Survey Article

Unsupervised Learning of Morphology

Harald Hammarström*
Radboud Universiteit and Max Planck
Institute for Evolutionary Anthropology

Lars Borin**
University of Gothenburg

This article surveys work on Unsupervised Learning of Morphology. We define Unsupervised Learning of Morphology as the problem of inducing a description (of some kind, even if only morpheme segmentation) of how orthographic words are built up given only raw text data of a language. We briefly go through the history and motivation of this problem. Next, over 200 items of work are listed with a brief characterization, and the most important ideas in the field are critically discussed. We summarize the achievements so far and give pointers for future developments.

1. Introduction

Morphology is understood here in its usual sense in linguistics, namely, as referring to (the linguistic study and description of) the internal structure of words. More specifically, we understand morphology following Haspelmath (2002, page 2) as "the study of systematic covariation in the form and meaning of words." For our purposes, we assume that we have a way of identifying the text words of a language, ignoring the fact that the term **word** has eluded exhaustive cross-linguistic definition. Similarly, we assume a number of commonly made distinctions in linguistic morphology, whose basic import is indisputable, but where there is an ongoing discussion on exactly where to draw the boundaries with respect to particular phenomena in individual languages.

Generally, a distinction is made between **inflectional morphology** and **word formation**. Inflectional morphology deals with the various realizations of the "same" lexical word, depending on the particular syntactic context in which the word appears. Typical examples of inflection are verbs agreeing with one or more of their arguments in the clause, or nouns inflected in particular case forms in order to show their syntactic relation to other words in the phrase or clause, for example, showing which verb argument they express. Word formation deals with the creation of new lexical words from existing

* Centre for Language Studies, Radboud Universiteit, Postbus 9103, 6500 HD Nijmegen, The Netherlands/ Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. E-mail: h.hammarstrom@let.ru.nl.

** Språkbanken, Department of Swedish Language, University of Gothenburg, Box 200, SE-405 30 Göteborg, Sweden. E-mail: lars.borin@svenska.gu.se.

Submission received: 2 March 2010; revised submission received: 19 August 2010; accepted for publication: 4 October 2010.

© 2011 Association for Computational Linguistics

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

ones, for example, agent nouns from verbs. If the same kinds of mechanisms are used as in inflectional morphology (i.e., the resulting word is derived out of only one existing word), linguists talk about **derivational morphology**. If two or more existing lexical words are combined in order to make up a new word, the terms **compounding** or **incorporation** are used, depending on the categories of the words involved.

There is a fairly wide array of formal means available cross-linguistically for expressing inflectional and derivational categories in languages. Most commonly, however, some form of affixation is involved—that is, some phonological material is added to the end of the word (suffixation), to the beginning of the word (prefixation), or (much more rarely) inside the stem of the word (infixation). Suffixes and prefixes (but rarely infixes) can form long chains, where the different positions, or "slots," express different kinds of inflectional or derivational categories. If a language has suffixing and/or prefixing—sometimes called concatenative morphology—it obviously follows that text words in that language can be segmented into a sequence of morphological elements: a stem and a number of suffixes after the stem and/or prefixes before the stem.[1]

Morphology is one of the oldest linguistic subdisciplines, and this brief presentation by necessity omits many intricacies and greatly simplifies a vast scholarship. (For standard, in-depth, introductions to this fascinating field, see, e.g., Nida [1949], Jensen [1990], Spencer and Zwicky [1998], or Haspelmath [2002].)

In language technology applications, a morphological component forms a bridge between texts and structured information about the vocabulary of a language. Some kind of morphological analysis and/or generation thus forms a basic component in many natural language processing applications. Many languages have quite complex morphological systems, with the number of potential inflected forms of a single lexical word running into the thousands, requiring a substantial amount of work if the linguistic knowledge of the morphological component is to be defined manually. For this reason, researchers often turn to machine learning approaches. This survey article is concerned with unsupervised approaches to morphology learning.

For the purposes of the present survey, we use the following definition of **Unsupervised Learning of Morphology** (ULM).

Input: Raw (unannotated, non-selective[2]) natural language text data
Output: A description of the morphological structure (there are various levels to be distinguished; see subsequent discussion) of the language of the input text
With: As little supervision (parameters, thresholds, human intervention, model selection during development, etc.) as possible

Some approaches have explicit or implicit biases towards certain kinds of languages; they are nevertheless considered to be ULM for this survey. Morphology may be narrowly taken as to include only derivational and inflectional affixation, where the number of affixes a root may take is finite[3] and the order of the affixes may not be

1 The picture is less simple in reality, because affixation is often accompanied by so-called morphophonological changes—changes in the shape of the stem or affix involved, or both—which often have the effect of blurring the boundaries between the elements.
2 With the term non-selective we intend to exclude text data that requires manual selection (e.g., curated singular–plural pairs).
3 The number of inflectional affixes is finite by definition. The derivational affixes—especially in heavily agglutinating languages—may be recursive, but are in practice finite.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

permuted.[4] This survey also subsumes attempts that take a broader view including clitics[5] and compounding (and there seems to be no reason in principle to exclude incorporation and lexical affixes: see Mithun [1999], pages 37–67, for some examples). Many, but not all, approaches focus on concatenative morphology/compounding only.

All works considered in this survey are designed to function on orthographic words, that is, raw text data in an orthography that provides a ready-made segmentation of text into words. Crucially, this excludes the rather large body of work that only targets word segmentation, that is, segmenting a sentence or a full utterance into words (cf. Goldsmith [2010] who also overviews word segmentation). However, works that explicitly aim to treat both word segmentation and morpheme segmentation in one algorithm are included. Hence, subsequent uses of the term segmentation in the present survey are to be understood as morpheme segmentation rather than word segmentation. We prefer the term segmentation to analysis because, in general in ULM, the algorithm does not attempt to label the segments.

There have been other approaches to machine learning of morphology than pure ULM as defined here, the most popular ones being:

- approaches that require selective input, such as "singular–plural pairs," or "all members of a paradigm" (Garvin 1967; Klein and Dennison 1976; Golding and Thompson 1985; Wothke 1985; McClelland and Rumelhart 1986; Brasington, Jones, and Biggs 1988; Tufis 1989; Zhang and Kim 1990; Borin 1991; Theron and Cloete 1997; Oflazer, McShane, and Nirenburg 2001, for example)
- approaches where some (small) amount of annotated data, some (small) amount of existing rule sets, or resources such as a machine-readable dictionary or a parallel corpus, are mandatory (Yarowsky and Wicentowski 2000; Yarowsky, Ngai, and Wicentowski 2001; Cucerzan and Yarowsky 2002; Neuvel and Fulop 2002; Johnson and Martin 2003; Rogati, McCarley, and Yang 2003, for example)

Such approaches are excluded from the present survey, unless the required data (e.g., paradigm members) are extracted from raw text in an unsupervised manner as well. We also exclude the special case of the second approach where morphology learning means not "learning the morphological system of a language," but rather "learning the inflectional classes of out-of-vocabulary words," namely, approaches where an existing morphological analysis component is used as the basis for guessing in which existing paradigm an unknown text word should belong (e.g., Antworth 1990; Mikheev 1997; Bharati et al. 2001; Forsberg, Hammarström, and Ranta 2006; Lindén 2008; Lindén 2009).

One of the matters that varies the most between different authors is the desired outcome. It is useful to set up the implicational hierarchy shown in Table 1 (which

4 There are, however, rare cases of languages which allow the permutation of specific pairs of prefixes, such as Kagulu (Petzell 2007), Yimas and Karawari (Foley 1991, pages 31–32) as well as Chintang, Bantawa, and possibly other Kiranti languages where prefix ordering in general is very free (Rai 1984; Bickel et al. 2007).

5 Clitics are affix-like elements that attach to words in particular syntactic positions, rather than to words of particular categories as proper affixes do. The English genitive *-'s* is sometimes classified as a clitic, because you can say things like *The girl I met yesterday's purse* (the *-'s* attaches to the end of the noun phrase, regardless of the part of speech of the last word, an adverb in this case). This could not happen with an inflectional suffix like the plural *-s*: **The girl I met yesterdays* cannot mean *The girls I met yesterday*.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Table 1
Levels of power of morphological analysis.

	Form	Meaning
Affix list	A list of the affixes	
⇑		
Same-stem decision	Given two words, decide if they are affixations of the same stem	Given two words, decide if they are affixations of the same lexeme
⇑		
Segmentation	Given a word, segment it into stem and affix(es)	
Morphological analysis		A functional labeling for the affixes in the segmentation
⇑		
Inflection tables	A list of the affixation possibilities for all stems	
Paradigm list		A list of the paradigms for all stem types, complete with functional labels for paradigm slots
⇑		
Lexicon+Paradigm	A list of the paradigms and a list of all stems with information of which paradigm each stem belongs to	
⇑		
Justification	A linguistically and methodologically informed motivation for the morphological description of a language	

need of course not correspond to steps taken in an actual algorithm). The division is implicational in the sense that if one can do the morphological analysis of a lower level in the table, one can also easily produce the analysis of any of the levels above it. Reflecting a fundamental assumption underlying most ULM work, form and meaning (semantics) are kept separate in the table (see Section 2). For example, if one can perform segmentation into stem and affixes, one can decide if two words are of the same stem (if meaning is disregarded) or the same lexeme (if meaning is taken into account). The converse need not hold; it is perfectly possible to answer the question of whether two words are of the same stem with high accuracy, without having to commit to what the actual stem should be.

Many recent articles fail to deal properly with previous and related work, some reinvent heuristics that have been proposed earlier, and there is little modularization taking place. Previous surveys and overviews for a general audience are Borin (1991), Batchelder (1997, pages 66–68), Powers (1998), Clark (2001, pages 80–82), Xanthos (2007, pages 95–107), Goldsmith (2001), Daelemans (2004, page 1898), Roark and Sproat (2007, pages 116–136), Hammarström (2007b, pages 10–15), Chan (2008, pages 48–60), Hammarström (2009b, pages 14–21), Borin (2009), Goldsmith (2010), and, to a more limited extent, the related-work sections of individual research papers. Kurimo, Creutz, and Turunen (2007), Kurimo, Creutz, and Varjokallio (2008a, 2008b), Kurimo and Turunen (2008), Kurimo and Varjokallio (2008), and McNamee (2008) are overviews of systems in the MorphoChallenge of the respective year. However, we will try to be more comprehensive than previous surveys and discuss the ideas in the field critically.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

We will not attempt a comparison in terms of accuracy figures as this is wholly impossible, not only because of the great variation in goals but also because most descriptions do not specify their algorithm(s) in enough detail. Fortunately, this aspect is better handled in controlled competitions, such as the *Unsupervised Morpheme Analysis—MorphoChallenge*[6] which offers tasks of segmentation of Finnish, English, German, Arabic, and Turkish.

2. History and Motivation of ULM

Usually and justifiedly, the work of Harris (1955, 1967) is given as the starting point of ULM. From another perspective, however, the same work by Harris can be said to equally represent the culmination of an endeavor in the linguistic school of thought known as **American structuralism**, to formalize the process of linguistic description into so-called **linguistic discovery procedures**.

The variety of American structuralism which concerned itself most with the formalization of linguistic discovery procedures is often connected with the name of Leonard Bloomfield, and its core tenet may be succinctly summed up in Bloomfield's oft-quoted dictum: "The only useful generalizations about language are inductive generalizations" (Bloomfield 1933, page 20). The so-called "extremist Post-Bloomfieldians" took this program a step further: "From Bloomfield's justified insistence on formal, rather than semantic, features as the *starting-point* for linguistic analysis, this group (especially Harris) set up as a theoretical aim the description of linguistic structure exclusively in terms of distribution" (Hall 1987, page 156).

The earliest reason for interest in ULM was thus—at least in part—*methodological* and arguably even *ideological*, but not (unlike at least some of the later ULM work) motivated by, for example, a desire to simulate language acquisition in humans.

More or less simultaneously with but independently of Harris, the Russian linguist Andreev launched a program much like that of Harris.[7] Andreev's work is much less known than that of Harris's, and for this reason we will describe it in some detail here. In a series of publications (Andreev 1959, 1963, 1965b, 1967), he develops an "algorithm for statistical-combinatory modeling of languages." This is part of a research program which, just like that of Harris, aims at eliminating semantics and considerations of meaning completely from the process of "discovery" of language structure.

Thus, Andreev claims to be able to go from unsegmented transcribed speech all the way up to syntax, using basically one and the same approach grounded in text (corpus) statistics. Given our focus on ULM, here we will be concerned only with his approach as applied to morphological segmentation.

Andreev's approach is much more explicitly based in text statistics—and to some extent in language typology—than Harris's work. The algorithm for morphological segmentation is described in some detail in the works of Andreev and his colleagues. It relies on statistics of letter frequencies in a text corpus, and of average word length in characters and average sentence length in words. From these statistics he calculates a number of heuristic thresholds which are used to iteratively grow affix candidates from characters at given positions in text words, and paradigm candidates from the resulting segmentations. Instead of looking at successor/predecessor counts or transition probabilities, Andreev looks at character positions in relation to word edges, from the first and

6 Web site `http://www.cis.hut.fi/morphochallenge2009/` accessed 10 September 2009.
7 To our knowledge, Andreev never refers to Harris's work.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

the last character inwards no further than the average word length. At each position, the amount of **overrepresentation** is calculated for each character found in this position in some word. The overrepresentation ("correlative function" in Andreev's terminology) is defined as the relative frequency of the character in this word position divided by its relative frequency in the corpus. The character–position combinations are used in order of decreasing overrepresentation in an iterative see-saw procedure, where affix and stem candidates are collected in alternating iterations of the algorithm. Andreev's approach reflects the same intuition as that of Harris; we would expect word-edge sequences of highly overrepresented characters to be flanked by marked differences in predecessor or successor counts calculated according to Harris's method.

A concrete example of how Andreev's method works (with the finer details omitted) is the following, originally presented by Andreeva (1963), but the presentation here is partly based on that in Andreev (1967).

In a 900,000-word corpus of electronics texts in Russian, the most overrepresented letter was <j> (Russian <й>) in the last position of the word, where it is eight times as frequent as in the corpus as a whole (Andreeva 1963, page 49). For the words ending in <j>, its most overrepresented predecessor was <o>, and using some thresholds derived from corpus statistics, the first affix candidate found was *-oj* (Russian -<ой>).Removing this ending from all words in which it appears and matching the remainders of the words (i.e., putative stems) against the other words of the corpus, yields a set of words from which additional suffix candidates emerge (including the null suffix). This set of words is then iteratively reduced, using the admissible suffix candidates (those below a certain length exceeding a heuristic threshold of overrepresentation) in each step, as long as at least two stem candidates remain. In other words: There must be at least two stems in the corpus appearing with all the suffix candidates. In the Russian experiment reported by Andreeva (1963), a complete adjective paradigm was induced, with 12 different suffixes. The initial suffix candidate, *-oj*, has a high functional load and consequently a high text frequency: It is the most ambiguous of the Russian adjective suffixes, appearing in four different slots in the adjective paradigm, and is also homonymous with a noun suffix.

In Andreev (1965b) the method is tested extensively on Russian, which is the subject of several papers in the volume, and a number of other languages: Albanian (Peršikov 1965), Armenian (Melkumjan 1965), Bulgarian (Fedulova 1965), Czech (Ožigova 1965), English (Malahovskij 1965), Estonian (Hol'm 1965), French (Kordi 1965), German (Fitialova 1965), Hausa (Fihman 1965a), Hungarian (Andreev 1965a), Latvian (Jakubajtis 1965), Serbo-Croatian (Panina 1965), Swahili (Fihman 1965b), Ukrainian (Eliseeva 1965), and Vietnamese (Jakuševa 1965). As an aside, we may note that only after the turn of the millennium are we again seeing this variety of languages in ULM work. Most of these studies are small-scale proof-of-concept experiments on corpora of varying sizes (from a few thousand words in many of the studies up to close to one million words for Russian). The outcomes are more often than not quite small "paradigmoid fragments," that is, incomplete and not always in correspondence with traditional segmentations. It is noteworthy, however, that the method could not produce a single instance of morphological segmentation for Vietnamese (Jakuševa 1965, page 228), which is as it should be, because Vietnamese is often held forth as a language without morphology.

The papers describing these experiments are short, and it is not always clear exactly what has been done. In fact, computers are not mentioned at all in most of the papers; on the contrary, it is quite clear that at least some of the experiments have been carried out manually. In principle, because Andreev and the other authors in Andreev (1965b) describe the procedure in great detail, it should be possible to replicate some of the

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

findings (cf. Altmann and Lehfeldt 1980, pages 195–198). To our knowledge, there has been one attempt to do this, by Cromm (1997), who reimplemented the method and tested it on the German Bible, experimenting with various parameter settings and also making some changes to the method itself. He notes that several parameters that Andreev provides mostly without motivation or comment in fact can be changed in a more accepting direction, leading to much increased recall without much loss in precision. Unfortunately, however, in his short paper, Cromm does not provide enough information about the algorithm or the changes that he made to it, so that the Russian original is still the only publicly available source for the details of Andreev's approach.

A very different, more practically oriented, motivation for ULM came in the 1980s, beginning with the supervised morphology learning ideas by Wothke (1985, 1986) and Klenk (1985a, 1985b) which later led to partly unsupervised methods (see the following). Because full natural language lexica, at the time, were too big to fit in working memory, these authors were looking for a way to analyze or stem running words in a "nicht-lexikalisches" manner, that is, without the storage and use of a large lexicon. This motivation is now obsolete.

The interest in purebred ULM was fairly low until about 1990, however, with only a few works appearing between the mid 1960s and 1990. Especially in the 1980s, the focus in computational morphology was on the development of finite-state approaches with hand-written rules, but in the course of the following decade, interest in ULM rose greatly, in the wake of a general increased attention during the 1990s to statistical and information-theoretically informed approaches in natural language processing. In speech processing, the problem of word segmentation is ever-present, and as the computational tools for taking on this problem became increasingly sophisticated and increasingly available not least as the result of a general development of computing hardware and software, researchers in linguistics and computational linguistics started taking a fresh look at the problems of word segmentation and ULM.

The work of Goldsmith (2000, 2001, 2006) represents a kind of focal point here. He pulls together a number of strands from earlier work, sets them against a theoretical background informed both by information theory (MDL) and linguistics, and uses them specifically to address the problem of ULM—in particular, unsupervised learning of inflectional morphology—and not, for instance, that of word segmentation or of stemming for information retrieval, and so forth.

Further, there has been the idea that ULM could contribute to various open questions in the field of first-language acquisition (see, e.g., Brent, Murthy, and Lundberg 1995; Batchelder 1997; Brent 1999; Clark 2001; Goldwater 2007). However, the connection is still rather vague and even if ULM has matured, it is not clear what implications, if any, this has for child language acquisition. Children have access to semantics and pragmatics, not just text strings, and it would be very surprising if such cues were not used at all in first language acquisition. Further, if some ULM technique was shown to be successful on some reasonably sized corpora, it does not automatically follow that children can (and do, if they can) use the same technique. Most current ULM techniques crucially involve long series of number crunching that seem implausible for the child-learning setting.

After the turn of the century, ULM has become something of a growth industry in language technology. There are several reasons for this. One obvious reason is a generally increased interest in machine learning, both theoretically (as a research area interesting in itself and as a possible tool for modeling human language acquisition and language learning) and for pragmatic reasons, as a way to reduce the manual work

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

involved in the construction of the lexical and grammatical knowledge bases needed for the realization of sophisticated language technology applications.[8]

Another reason has to do with the acceptance of the world as multilingual and the understanding that language communities are very unequally endowed with language technology resources. There are on the order of 7,000 languages spoken in the world today (Lewis 2009). Their size in number of first-language speakers is very unevenly distributed. The top 30 languages in the world account for more than 60% of its population. At the other end of the scale, we find that most languages are spoken by quite small communities:

> There are close to 7,000 languages in the world, and half of them have fewer than 7,000 speakers each, less than a village. What is more, 80% of the world's languages have fewer than 100,000 speakers, the size of a small town. (Ostler 2008, page 2)

On the whole, small language communities will tend to have correspondingly small financial and other resources that could be spent on the development of language technology, but the cost of, for example, constructing a lexicon or a parser for a language is more or less constant, and not proportional to the number of speakers of the language.

At the same time, it has been observed over and over again that the use or non-use of a language in a particular situation—where the language could in principle be used, but where there is a choice available between two or more languages—is intimately connected with the attitudes towards the language among the participants. This is perhaps the most reliable determiner of language use, and not factors such as effort, lack of vocabulary, and so on, which in many cases seem to be post hoc rationalizations motivating a choice made on attitudinal grounds. Another way of expressing this is that languages are more or less prestigious in the eyes of their speakers, and that linguistic inferiority complexes seem to be common in the world.

However, rather than taking status as an inherent and immutable characteristic of a language, we should see it for what it is, namely, a perceived characteristic, something that lies in the eye of the beholder. As such, it can be influenced by human action. Important for our purposes here is that it has been suggested that making available modern information and communication technologies for a language, including the creation of linguistic resources and language technology for it, may serve to raise its status (see, e.g., the papers in Saxena and Borin 2006).

This, then, is another reason for pursuing ULM: to be able to provide language technology to language communities lacking the requisite resources. However, ULM, at least as understood for the purposes of this survey, requires a written language, which would still exclude a substantial majority of the world's languages (Borin 2009). Note that the remainder—languages with a tradition of writing—are not on the whole small language communities; in the first instance, we are talking about the few hundred most spoken languages in the world, for example, the 313 languages with at least one million native speakers (accounting for about 80% of the world's population) surveyed by the Linguistic Data Consortium some years back in their *Low-density language survey* (Strassel, Maxwell, and Cieri 2003; Borin 2009).

8 Another pragmatic, less savory reason is a general downplaying of linguistic knowledge in the language technology research community (Reiter 2007).

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

The hope is often expressed in the literature that ULM and other unsupervised methods could be employed in order to rapidly and cheaply (in terms of human effort) bootstrap basic language technology resources for new languages.

It should be noted that, even for larger languages, because of the human effort needed to build computational morphological resources, many such implementations are not released to the public domain. Also, open domain texts will always contain a fair share of (inflected) previously unknown words that are not in the lexicon. There has to be strategy for such out-of-dictionary words—a ULM-solving algorithm is one possibility. The ULM problem as specified, therefore, still has a role to play for larger languages.

Finally, and closely related to the preceding reason, ULM and other kinds of machine learning of linguistic information are increasingly seen as providing potential tools in **language documentation**.[9]

It has been realized for some time that languages are disappearing at a rapid rate in the modern world (Krauss 1992, 2007). Many linguists see this loss of linguistic diversity as a disaster in the cultural and intellectual sphere on a par with the loss of the world's biodiversity in the ecological sphere, only on a grander scale; languages are going extinct more rapidly than species. Enter language documentation (Gippert, Himmelmann, and Mosel 2006), which is construed as going well beyond traditional descriptive linguistic fieldwork, aspiring as it does to capture all aspects—linguistic, cultural, and social—of a language community's day-to-day life, in video and audio recordings of a wide range of sociocultural activities, in still images, and in representative artifacts. Basic linguistic descriptions of lexicon and grammar made on the basis of transcribed recordings still form an important component of language documentation, however, and with the realization that languages are disappearing at a far faster rate than linguists can document them, it is natural to look for ways of making this process less labor-intensive.[10]

In summary, we have seen the following motivations for ULM (in chronological order):

- Linguistic theory
- Elimination of the lexicon
- Child language acquisition
- Morphological engine bootstrapping
- Language description and documentation bootstrapping

As noted, the motivation of eliminating the lexicon is now obsolete, whereas the others are active to various degrees. By far the most popular motivation has been, and still is, that of inducing a morphological analyzer/segmentation from raw text data (with little human intervention) in a well-described language. However, as we have argued herein, the timing is right for the momentum to carry over also to under-described languages.

9 To our knowledge, Eguchi (1987, page 168) is the first author to suggest ULM as one of several computational aids to the language documentation fieldworker.

10 For example, in the instructions for the recent large-scale language documentation effort BOLD:PNG—*Basic Oral Language Documentation: Papua New Guinea*—we read: "Try not to spend more than an hour transcribing a minute of text." and "As before, try not to spend more than an hour translating a minute of text." `www.boldpng.info/bold/stage3`.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

3. Trends and Techniques in ULM

3.1 Roadmap and Synopsis of Earlier Studies

A chronological listing of earlier work (with very short characterizations) is given in Table 2. Several papers are co-indexed if they represent essentially the same line of work by essentially the same author(s).

Given the number of algorithms proposed, it is impossible to go through the techniques and ideas individually. However, we will attempt to cover the main trends and look at some key questions in more detail.

The problem has been approached in four fundamentally different ways, which we may summarize in the following way.

(a) **Border and Frequency:** In this family of methods, if a substring occurs with a variety of substrings immediately adjacent to it, this is interpreted as evidence for a segmentation border. In addition, frequent or somehow overrepresented substrings are given a direct interpretation as candidates for segmentation. A typical implementation is to subject the data to a compression formula of some kind, where frequent long substrings with clear borders offer the optimal compression gain. The outcome of such a compression scheme gives the segmentation. In addition, for those approaches which also target paradigms, stem–suffix co-occurrence statistics are gathered given the segmentation produced, rather than all possible segmentations.

(b) **Group and Abstract:** In this family of methods, morphologically related words are first grouped (clustered into sets, paired, shortlisted, etc.) according to some metric, which is typically string edit distance, but may include semantic features (Schone 2001), distributional similarity (Freitag 2005), or frequency signatures (Wicentowski 2002). The next step is to abstract some morphological pattern that recurs among the groups. Such emergent patterns provide enough clues for segmentation and can sometimes be formulated as rules or morphological paradigms.

(c) **Features and Classes:** In this family of methods, a word is seen as made up of a set of features—*n*-grams in Mayfield and McNamee (2003) and McNamee and Mayfield (2007), and initial/terminal/mid-substring in De Pauw and Wagacha (2007). Features which occur on many words have little selective power across the words, whereas features which occur seldom pinpoint a specific word or stem. To formalize this intuition, Mayfield and McNamee and McNamee and Mayfield use TF-IDF, and De Pauw and Wagacha use entropy. Classifying an unseen word reduces to using its features to select which word(s) it may be morphologically related to. This decides whether the unseen word is a morphological variant of some other word, and allows extracting the "variation" by which they are related, such as an affix.

(d) **Phonological Categories and Separation:** In this family of methods, the phonemes (approximated by graphemes) are first classed into categories, foremostly, vowel versus consonant. Thereafter, each word is separated into its vowel skeleton and its consonant skeleton, after which various

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Table 2
Very brief roadmap of earlier studies.

	Model	Superv.	Experimentation	Learns what?
Harris 1955, 1968, 1970	C	T	English	Segmentation
Andreev 1965a, Andreev 1967, Chapter 2, Peršikov 1965; Melkumjan 1965; Fedulova 1965; Ožigova 1965; Malahovskij 1965; Hol'm 1965; Kordi 1965; Fitialova 1965; Fihman 1965a; Andreev 1965a; Jakubajtis 1965; Panina 1965; Fihman 1965b; Eliseeva 1965; Jakuševa 1965	C	T	Vietnamese to Hungarian (I)	Segmentation
Gammon 1969	C	T	English	Segmentation
Lehmann 1973, pages 71–93	C	T	French (I)	Segmentation
de Kock and Bossaert 1969, 1974, 1978	C	T	French/Spanish	Lexicon+ Paradigms
Faulk and Gustavson 1990	C	T	English (I)	Segmentation
Hafer and Weiss 1974	C	T	English (IR)	Segmentation
Klenk and Langer 1989	C	T+SP	German	Segmentation
Langer 1991	C	T+SP	German	Segmentation
Redlich 1993	C	T	English (I)	Segmentation
Klenk 1992, 1991	C	T+SP	Spanish	Segmentation
Flenner 1992, 1994, 1995	C	T+SP	Spanish	Segmentation
Janßen 1992	C	T+SP	French	Segmentation
Juola, Hall, and Boggs 1994	C	T	English	Segmentation
Brent 1993, 1999; Brent, Murthy, and Lundberg 1995; Snover 2002; Snover, Jarosz, and Brent 2002; Snover and Brent 2001, 2003	C	T	English/Child-English/Polish/French	Segmentation
Deligne and Bimbot 1997; Deligne 1996	C	T	English/French (I)	Segmentation
Yvon 1996	C	T	French (I)	Segmentation
Kazakov 1997; Kazakov and Manandhar 1998, 2001	C	T	French/English	Segmentation
Jacquemin 1997	C	T	English	Segmentation
Cromm 1997	C	T	German	Segmentation
Gaussier 1999	C	T	French/English (I)	Lexicon+ Paradigms
Déjean 1998a, 1998b	C	T	Turkish/English/Korean/French/Swahili/Vietnamese (I)	Affix Lists
Medina Urrea 2000, 2003, 2006b	C	T	Spanish	Affix List
Schone and Jurafsky 2000, 2001a; Schone 2001	C	T	English	Segmentation
Goldsmith 2000, 2001, 2006; Belkin and Goldsmith 2002; Goldsmith, Higgins, and Soglasnova 2001; Hu et al. 2005b; Xanthos, Hu, and Goldsmith 2006	C	T	English (I)	Lexicon+ Paradigms
Baroni 2000, 2003	C	T	Child-English/English	Affix List
Cho and Han 2002	C	T	Korean	Segmentation
Sharma, Kalita, and Das 2002, 2003; Sharma and Das 2002	C	T	Assamese	Lexicon+ Paradigms
Baroni, Matiasek, and Trost 2002	C/NC	T	English/German (I)	Related word pairs
Bati 2002	C/NC	T	Amharic	Lexicon+ Paradigms
Creutz 2003, 2006; Creutz and Lagus 2002, 2004, 2005a, 2005b, 2005c, 2007; Creutz, Lagus, and Virpioja 2005; Hirsimäki et al. 2003; Creutz et al. 2005	C	T	Finnish/Turkish/English	Segmentation

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Table 2
(*continued*)

	Model	Superv.	Experimentation	Learns what?
Kontorovich, Don, and Singer 2003	C	T	English	Segmentation
Medina Urrea and Díaz 2003; Medina-Urrea 2006a, 2008	C	T	Chuj/Ralámuri/Czech	Affix List
Mayfield and McNamee 2003; McNamee and Mayfield 2007	-	-	8 West European languages (IR)	Same-stem
Zweigenbaum, Hadouche, and Grabar 2003; Hadouche 2002	C	T	Medical French	Segmentation
Pirrelli et al. 2004; Pirrelli and Herreros 2007; Calderone 2008	C	T	Italian/English/Arabic	Unclear
Johnson and Martin 2003b	C	T	Inuktitut	Unclear
Katrenko 2004	C	T	Ukrainian	Lexicon+ Paradigms
Ćavar et al. 2004a, 2004b; Ćavar, Rodrigues, and Schrementi 2006; Ćavar et al. 2006	C	T	Child-English	Unclear
Rodrigues and Ćavar 2005, 2007	NC	T	Arabic	Segmentation
Monson 2004, 2009; Monson et al. 2004, 2007a, 2007b, 2008, 2008a, 2008b	C	T	English/Spanish/ Mapudungun (I)	Segmentation
Yarowsky and Wicentowski 2000; Wicentowski 2002, 2004	C/NC	T	30-ish mostly European type languages	Segmentation + Rewrite Rules
Gelbukh, Alexandrov, and Han 2004	C	-	English	Segmentation
Argamon et al. 2004	C	T	English	Segmentation
Goldsmith et al. 2005; Hu et al. 2005a	C/NC	T	Unclear	Unclear
Bacchin, Ferro, and Melucci 2005, 2002a, 2002b; Nunzio et al. 2004	C	T	Italian/English	Segmentation
Oliver 2004, Chapter 4–5	C	T	Catalan	Paradigms
Bordag 2005a, 2005b, 2007, 2008	C	T	English/German	Segmentation
Hammarström 2005, 2006a, 2006b, 2007b, 2009a, 2009b	C	-	Maori to Warlpiri	Same-stem
Bernhard 2005a, 2005b, 2006, 2007, 2008	C	T	Finnish/Turkish/English	Segmentation+ Related sets of words
Keshava and Pitler 2005	C	T	Finnish/Turkish/English	Segmentation
Johnsen 2005	C	T	Finnish/Turkish/English	Segmentation
Atwell and Roberts 2005	C	T	Finnish/Turkish/English	Segmentation
Dang and Choudri 2005	C	T	Finnish/Turkish/English	Segmentation
ur Rehman and Hussain 2005	C	T	Finnish/Turkish/English	Segmentation
Jordan, Healy, and Keselj 2005, 2006	C	T	Finnish/Turkish/English	Segmentation
Goldwater, Griffiths, and Johnson 2005; Goldwater 2007; Naradowsky and Goldwater 2009	C	T	English/Child-English	Segmentation
Freitag 2005	C	T	English	Segmentation
Golcher 2006	C	-	English/German	Lexicon+ Paradigms
Arabsorkhi and Shamsfard 2006	C	T	Persian	Segmentation
Chan 2006, Chan 2008, pages 101–139	C	T	English	Paradigms
Demberg 2007	C/NC	T	English/German/ Finnish/Turkish	Segmentation
Dasgupta and Ng 2006, 2007a; 2007b; Dasgupta 2007	C	T	Bengali	Segmentation
De Pauw and Wagacha 2007	C/NC	T	Gikuyu	Segmentation
Tepper 2007; Tepper and Xia 2008	C/NC	T+RR	English/Turkish	Analysis
Xanthos 2007	NC	T	Arabic	Lexicon+ Paradigms
Majumder et al. 2007; Majumder, Mitra, and Pal 2008	C	T	French/Bengali/French/ Bulgar- ian/Hungarian	Analysis
Zeman 2008, 2009	C	-	Czech/English/German/ Finnish	Segmentation+ Paradigms
Kohonen, Virpioja, and Klami 2008	C	T	Finnish/Turkish/English	Segmentation

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Table 2
(*continued*)

	Model	Superv.	Experimentation	Learns what?
Goodman 2008	C	T	Finnish/Turkish/English	Segmentation
Golénia 2008	C	T	Turkish/Russian	Segmentation
Pandey and Siddiqui 2008	C	T	Hindi	Segmentation+ Paradigms
Johnson 2008	C	T	Sesotho	Segmentation
Snyder and Barzilay 2008	C/NC	T	Hebrew/Arabic/Aramaic/ English	Segmentation
Spiegler et al. 2008	C	T	Zulu	Segmentation
Moon, Erk, and Baldridge 2009	C	T	English/Uspanteko	Segmentation
Poon, Cherry, and Toutanova 2009	C	T	Arabic/Hebrew	Segmentation

Abbreviations: C = Concatenative; I = Impressionistic evaluation; IR = Evaluation only in terms of Information Retrieval Performance; NC = Non-concatenative; RR = Hand-written rewrite rules; SP = Some manually curated segmentation points; T = Thresholds and Parameters to be set by a human.

> frequency techniques reminiscent of those of the (a) approaches can be applied. This strategy is targeted towards the special kind of non-concatenative morphology called **intercalated morphology**[11] with the observation that, empirically, in those (relatively few) languages which have intercalated morphology, it does seem to depend on vowel/consonant considerations. In Xanthos (2007), the phonological categories are inferred in an unsupervised manner (cf. Goldsmith and Xanthos 2009) whereas in Bati (2002) and Rodrigues and Ćavar (2005, 2007) they are seen as given by the writing system.

The first two, (a) and (b), enjoy a fair amount of popularity in the reviewed collection of work, though (a) is much more common and was the only kind used up to about 1997. The last two, (c) and (d), have been utilized only by the sets of authors cited therein.

Let us now look at some salient questions in more detail. The following notation will be used in formal statements:

- $w, s, b, x, y, \ldots \in \Sigma^*$: lowercase-letter variables range over strings of some alphabet Σ and are variously called words, segments, strings, and so forth.
- $W, S, \ldots \subseteq \Sigma^*$: capital-letter variables range over sets of words/strings/segments.
- $\mathcal{C}, \ldots$: capital-letter caligraphic variables range over multisets of words/strings/segments.
- $|\cdot|$: is overloaded to denote both the length of a string and the cardinality of a set.
- $w[i]$: denotes the character at position i in the string w. For example, if $w = hello$ then $w[1] = h$.
- $w[i:j]$: denotes the segment from position i to j (inclusive) of the string w. For example, if $w = hello$ then $w[1:|w|] = hello$.

11 Also known as templatic morphology or root-and-pattern morphology.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

- $W_{\mathcal{C}}$ is used to denote the set of words in a corpus $\mathcal{C}$.
- $f^{p}_{W}(x) = |\{z|xz \in W\}|$: the (prefix) frequency of x, that is, the number of words in W with initial segment x.
- $f^{s}_{W}(x) = |\{z|zx \in W\}|$: the (suffix) frequency of x, that is, the number of words in W with final segment x.

Subscript letters are dropped when understood from the context.

3.2 Border and Frequency Methods

3.2.1 Letter Successor Varieties. Most (if not all) authors trace the inspiration for their border heuristics back to Harris (1955). In fact, Harris defines a family of heuristics, all based on letter successor/predecessor varieties. They were originally presented as applying to utterances made up of *phoneme* sequences (Harris 1955), but they apply just the same to words, namely, grapheme sequences (Harris 1970). The basic counting strategy, labelled letter successor varieties (LSV) by Hafer and Weiss (1974), is as follows.

> Given a set of words W, the letter successor variety of a string x of length i is defined as the number of distinct letters that occupy the $i + 1$st position in words that begin with x in W:
>
> $$LSV(x) = |\{z[|x| + 1]|z = xy \in W\}|$$

Table 3 shows an example of a letter successor count on a tiny contrived wordlist.

We may define the letter predecessor variety (LPV) analogously. For a given suffix x, the $LPV(x)$ is the number of distinct letters that occupy the position immediately preceding x in the words of W that end in x. LSV/LPV counts for an example word are shown in Table 4.

It should be noted that Harris (1955, page 192, footnote 4) explicitly targets the variety in letter successors *types* (i.e., is only interested in which letters *ever* occur in the successor position, as opposed to being interested in their frequencies). For example, if there are two different letters occurring in successor position, one occurring a thousand times and the other once, Harris's letter successor variety is still two—the same as if the two letters occurred once each. Subsequent authors have suggested that the full frequency distribution of the *token* letter successors carries a better signal of morpheme boundary. After all, if there is a significant token frequency skewing, this suggests that we are in the middle of coherent morpheme. Moreover, mere type counts may be influenced by phonotactic constraints (consonant after vowel, etc.), which come out less significant in token frequency counts (Goldsmith 2006, page 6). Already the earliest

Table 3
Example of LSV-counts for some example prefixes (bottom) based on a small example word list (top).

$W = \{$abide, able, abode, and, art, at, bat$\}$				
x	a	ab	abe	...
$\{z\|z = xy \in W\}$	{abide, able, abode, and, art, at}	{abide, able, abode}	$\emptyset$	...
$LSV(x)$	4 (b,n,r,t)	3 (i,l,o)	0	...

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Table 4
LSV counts for *d-*, *di-*, *dis-*, . . . , *disturbance-* and LPV counts for *-e*, *-ce*, *-nce*, . . . , *-disturbance*. All figures are computed on the Brown Corpus of English (Francis and Kucera 1964), using the 27 letter alphabet [$a - z$] plus the apostrophe. There are $|W| = 42{,}353$ word types in lowercase.

LSV		13	20	21	6	1	1	3	1	1	1	1
	d	i	s	t	u	r	b	a	n	c	e	
LPV	0	1	1	1	1	1	1	19	6	12	25	

follow-ups to Harris (Gammon 1969; Hafer and Weiss 1974; Juola, Hall, and Boggs 1994) experiment with replacing the raw LSV/LPV counts with the entropy of the character *token* distribution. The character token distribution after a given segment can be seen as a probability distribution whose events are the characters of the alphabet. The entropy of this probability distribution then measures how unpredictable the next character is after a given segment. In general, for a discrete random variable X with possible values $x_1, \ldots, x_n$, the expression for entropy takes the following form:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

Thus, with alphabet Σ, the letter successor entropy (LSE) for a prefix x is defined as

$$LSE(x) = -\sum_{c \in \Sigma} \frac{f^p(xc)}{f^p(x)} \log_2 \frac{f^p(xc)}{f^p(x)}$$

At least two authors (Golcher 2006; Hammarström 2009b) have questioned entropy as the appropriate measure for highlighting a morpheme boundary. Entropy measures how skewed the distribution is as a whole, that is, how deviant the most deviant member is, in addition to the second member, the third, and so on. If there is no morpheme boundary, the morpheme continues with (at least) one character. So *one* deviant, highly predictable, character is necessary and sufficient to signal a non-break, and it is arguably irrelevant if there are second- and third-place, and so forth, highly predictable characters that also signal the absence of a morpheme boundary. For example, the character token distribution before *-ng* is shown in Table 5. Obviously, the fact that of the 3,352 occurrences of *-ng*, 3,258 of them are preceded by *-i-*, says that the absence of a morpheme boundary is highly likely. Now, does it matter that also another 35 are *-o-* versus only 4 for *-e-*? Entropy would also take into account the skewedness of *-o-* versus *-e-*, whereas for Hammarström (2009b) and Golcher (2006) only the skewedness of the most skewed character (i.e., the character that potentially constitutes the morpheme continuation) is interesting, in this example *-i-*. Therefore, these approaches only use the maximally skewed character to predict the presence/absence of a morpheme boundary. The letter successor max-drop (LSM) for a prefix x is defined as the fraction not occupied by its maximally skewed one-character continuation:

$$LSM(x) = 1 - max_{c \in \Sigma} \frac{f^p(xc)}{f^p(x)}$$

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Table 5
The character token distributon for the character immediately preceding *-ng*, computed on the Brown Corpus of English (Francis and Kucera 1964).

-ng 3,352														
-n-	1	*-l-*	1	*-h-*	1	*-e-*	4	*-u-*	26	*-a-*	26	*-o-*	35	*-i-*	3,258

Table 6
Normalized LPV/LPE/LPM-scores for *-e*, *-ce*, *-nce*, . . . , *-disturbance*. All figures are computed on the Brown Corpus of English (Francis and Kucera 1964), using the 27-letter alphabet $[a-z]$ plus the apostrophe. There are $|W| = 42,353$ word types in lowercase.

	d	i	s	t	u	r	b	a	n	c	e
$\overline{LPV}$	0.03	0.03	0.03	0.03	0.03	0.03	0.70	0.22	0.44	0.92	
$\overline{LPE}$	0.0	0.0	0.0	0.0	0.0	0.0	0.74	0.28	0.38	0.81	
$\overline{LPM}$	0.0	0.0	0.0	0.0	0.0	0.0	0.83	0.53	0.37	0.85	

Which one of LSV/LSE/LSM is the "correct" one? The answer, of course, depends on one's theory of affixation, for which the field has no single answer (see Section 3.6, subsequently).

Empirically, however, the three measures are highly correlated. To compare the three, we normalize them to their maxima in order to get a "border" score ≤ 1. The maximum achievable LSV is the alphabet size, so the normalized $\overline{LSV(x)} = \frac{LSV(x)}{|\Sigma|}$. The maximum achievable LSE is a uniform distribution across the alphabet, so the normalized $\overline{LSE(x)} = \frac{LSE(x)}{-|\Sigma| \cdot (\frac{1}{|\Sigma|} \log_2 \frac{1}{|\Sigma|})}$. The maximum achievable LSM is a uniform distribution across the alphabet, so the normalized $\overline{LSM(x)} = \frac{LSM(x)}{1-\frac{1}{|\Sigma|}}$. The predecessor analogues $\overline{LPV}, \overline{LPE}, \overline{LPM}$ are obvious. Table 6 shows an example word and its normalized predecessor scores of the three kinds.

As in the example, the three different measures have nearly the same story to tell in general, at least for English. For the three measures, Table 7 shows the Pearson product-moment correlation coefficient between the LPH/LPE/LPM-values of all terminal segments, as well as the Pearson product-moment correlation coefficient between the LPH/LPE/LPM-ranks of all terminal segments. Most usages in the literature of the letter successor counts have been relative to other counts on the same language. In such cases, the rank correlations show that all three measures can be expected to have near identical effects.

Table 7
The Pearson product-moment correlation coefficient between LPH/LPE/LPM-values (r) and the Pearson product-moment correlation coefficient between LPH/LPE/LPM-ranks (r-rank). All values are computed on the Brown Corpus of English (Francis and Kucera 1964), using the 27-letter alphabet $[a-z]$ plus the apostrophe. There are $|W| = 42,353$ word types in lowercase.

	LPH&LPE	LPE&LPM	LPM&LPH
r	0.872	0.957	0.729
r-rank	0.999	0.998	0.996

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

A number of concrete ways to use LSV/LPVs for segmentation are suggested by Harris (1955) and Hafer and Weiss (1974); for instance:

(a) **Cutoff:** By far the easiest way to segment a test word is first to pick some cutoff threshold k and then break the word wherever its successor (or predecessor or both) variety reaches or exceeds k.

(b) **Peak and plateau:** In the peak and plateau strategy, a cut in a word w is made after a prefix x if and only if $LSV(w[1:|x|-1]) \leq LSV(x) \geq LSV(w[1:|x|+1])$; that is, if the successor count for x forms a local "peak" or sits on a "plateau" of the LSV-sequence along the word.

(c) **Complete word:** A break is made after a word prefix (or before a word suffix) if that prefix (or suffix) is found to be a complete word in the corpus word list W.

These and similar strategies have been discussed and evaluated in various settings in the literature, and it is unlikely that any strategy based on LSV/LSE/LSM-counts alone will produce high-precision results. The example in Table 6 showing morpheme border heuristics on a specific word illustrates the matter at heart. Any intuitively plausible theory of affixation should allow abundant combination of morphemes without respect to their phonological form, which predicts that high LSV/LSE/LSM values should emerge at morpheme boundaries. However, there appears to be no reason why the converse should hold—high LSV/LSE/LSM values could emerge in other places of the word as well. Indeed, any frequent character at the end or beginning of a word may also be expected to show high LSV/LSE/LSM around it, such as the *-e* at the end of *disturbance* which has higher values than, for example, *-ance*. Therefore, simply inferring that high LSV/LSE/LSM values indicate a morpheme border is not a sound principle in general.

A different (but less successful, even when supervised) way to use character sequence counts is that associated with Ursula Klenk and various colleagues (Klenk and Langer 1989; Klenk 1991, 1992; Langer 1991; Flenner 1992, 1994, 1995; Janßen 1992). For each character bigram c_1c_2, they record, with some supervision in the form of manual curation, at what percentage there is a morpheme boundary before $|c_1c_2$, between $c_1|c_2$, after $c_1c_2|$, or none. A new word can then be segmented by sliding a bigram window and taking the split which satisfies the corresponding bigrams the best. For example, given a word *singing*, if the window happens to be positioned at *-gi-* in the middle, the bigram splits *ng|*, *g|i*, and *|in* are relevant to deciding whether sin*g*|*i*ng is a good segmentation. Exactly how to do the split by sliding the window and combining such bigram split statistics is subject to a fair amount of discussion. It became apparent, however, that the appropriateness of a bigram split is dependent on, for example, the position in a word—*-ed* is likely at the end of a word, but hardly in any other position—and exception lists and cover-up rules had to be introduced, before the approach was abandoned altogether.

3.2.2 Frequency Heuristics. For reasons just explained, most (if not all) recent authors in the border-and-frequency tradition have incorporated another measure, complementary to a morpheme border heuristic. This measure is nearly always directly or indirectly related to frequency, that is, frequent segments of some kind are singled out. Frequency has been used in many different ways. The simplest way is to look at the raw frequency of segments of any length, but, inevitably, this will sweep in any short

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

segment. Indeed, better candidates for morphemic segmentation are segments which are somehow overrepresented, that is, more frequent than random. There are various ways to define this property as well, including the following.

Overrepresentation as more-frequent-than-its-length: For a segment x of $|x|$ characters, it is overrepresented to the degree that it is more common than expected from a segment of its length. This applies to a segment in any position.

$$\frac{f(x)}{|\Sigma|^{|x|}}$$

Overrepresentation as more-frequent-than-its-parts: For a segment $x = c_1c_2 \ldots c_n$ of n characters, it is overrepresented to the degree that it is more common than expected from a co-occurrence of its parts. This applies to a segment in any position.

$$\frac{f(c_1c_2 \ldots c_n)}{f(c_1)f(c_2) \ldots (c_n)}$$

Overrepresentation as more-frequent-as-suffix: For a segment x, it is overrepresented to the degree that its probability as a suffix is higher than in any other (non-final) position. This applies to a segment in terminal position (but with obvious analogues for other positions).

With such measures, many authors have singled out affixes above a certain over-representation-value threshold or overrepresentation-rank threshold.

Threshold values are unsatisfactory because typically there is no theory in which to interpret them. Although they may be set ad hoc with some success, such settings do not automatically generalize. Such considerations have led many authors to devise compression-inspired models for exploiting skewed frequencies. In particular, several different sets of authors have invoked Minimum Description Length (MDL) as the motivation for a given formula to compress input data into a morphologically analyzed representation.[12]

The MDL principle is a general-purpose method of statistical inference. It views the learning/inference process as data compression: For a given set of hypotheses H and data set D, we should try to find the hypothesis in H that compresses D most (Grünwald 2007, pages 3–40). Concretely, such a calculation can take the the following form. If $L(H)$ is the length, in bits, of the description of the hypothesis; and $L(D|H)$ is the length, in bits, of the description of the data when encoded with the help of the hypothesis, then MDL aims to minimize $L(H) + L(D|H)$.

In principle, all of the works that have invoked MDL in their ULM method act as follows. A particular way Q of describing morphological regularities is conceived that has two components which we may call patterns P and data D. A coding scheme is devised to describe any P and to describe any collection of actual words with some specific P and D. A greedy search is done for a local minimum of the sum $L(P) + L(D|P)$ to describe the set of words W (in some approaches) or the bag of word tokens $\mathcal{C}$ (in

12 To our knowledge, Brent (1993) is the first author to do so for *morphological* segmentation.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

other approaches) of the input text data.[13] To take one concrete example, Goldsmith's (2006) particular way Q of describing morphological regularities is to allow for a list of stems, a list of affixes, a list of signatures (structures indicating which stems may appear with which affixes, i.e., a list of pointers to stems, and a list of pointers to suffixes). The search is then among different lists of stems, affixes, and signatures to see which is the shortest to account for the words of the corpus. Further details of such coding schemes need not concern us here, but for a range of options see, for example, Goldsmith (2001, 2006), Xanthos, Hu, and Goldsmith (2006), Creutz and Lagus (2007), Argamon et al. (2004), Arabsorkhi and Shamsfard (2006), Ćavar et al. (2004b), Baroni (2003), or Brent, Murthy, and Lundberg (1995).

It should be noted that the label MDL, in at least the terminology of Grünwald (2007, pages 37–38), is infelicitous for such cases where the P, D-search is not among different description *languages*, but among varations within a fixed language Q. For example, in the stem-affixes-signatures way of description (a specific Q), the search does not include other (possibly more parsimonious?) ways of description that do not use stems, affixes, or signatures at all. For the MDL-label to apply with its full philosophical underpinnings, the scope must include any possible compression algorithm, namely, any Turing machine. In this respect it is important to note that, compared to the schemes devised so far, Lempel-Ziv compression, another description language, should yield a superior compression (as, in fact, conceded by Baroni 2000, pages 146–147). MDL-inspired optimization schemes have achieved very competitive results in practice, however, and must be considered the leading paradigm to exploit skewed frequencies for morphological analysis.

3.2.3 Paradigm Induction. The next step after segmentation is to induce systematic alternation patterns, or (inflectional) paradigms,[14] and this is usually done as an extension of a border-and-frequency approach. For purposes of ULM, a paradigm is typically defined as a maximally large set of affixes whose members systematically occur on an open class of stems. For a number of reasons, finding paradigms is a major challenge. The number of theoretically possible paradigms is exponential in the number of affixes (as paradigms are *sets* of affixes). Paradigms do not need to be disjoint; in real languages they are typically not. Rather, words in the same part of speech tend to share affixes across paradigms (Carstairs 1983). In addition, without any language-specific knowledge, basically the only evidence at hand is co-occurrence of stems and affixes (i.e., when a word occurs in the corpus it evidences the co-occurrence of a [hypothetical] stem and suffix making up that word). Paradigm induction would be an easy problem if all affixes that *could* legally appear on a word *did* appear on each such word in a raw text corpus. This is, as is well known, far from the case. A typical corpus distribution is that a few lexemes appear very frequently but by far most lexemes appear once or only a few times (Baayen 2001). What this means for morphology is that most lexemes will appear with only one or a minority of their possible affixes, even in languages with relatively little morphology. Of course, there is also the risk that some rare affix, for example, the Classical Greek alternative medial 3p. pl. aorist imperative ending -σθων

13 As most approaches define their task as capturing the set of legal morphological forms, their goal should be to compress W, but see Goldwater (2007, pages 53–59) for arguments for compressing $\mathcal{C}$.

14 Note also that paradigm information can be fed back into the segmentation process in that some affixes which do poorly according to some paradigm-related measure (e.g., affixes that do not take part in many paradigms) can be weeded out.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

(Blomqvist and Jastrup 1998), may not appear at all even in a very large Classical Greek corpus.

More formally, consider a morphological paradigm (set of suffixes) P that is a true paradigm according to linguistic analysis. If k lexemes that are inflected according to P occur in a corpus, each of the k lexemes will occur in $1 \leq i \leq |P|$ forms. The number of forms i that a lexeme occurs in is likely not to be normally distributed. Most lexemes will occur in only one form, and only very few, if any, lexemes will occur in all $|P|$ forms. It appears that for most languages and most paradigms, the number of lexemes that occur in i forms tends to decrease logarithmically in i (Chan 2008, pages 75–84). As an example, consider the three most common paradigms in Swedish and the frequency of forms in Table 8.

Works which have attempted nevertheless to tackle the matter of paradigms, at least for languages with one-slot morphology, include Zeman 2008, 2009, Hammarström (2009b), and Monson (2009). They explicitly or implicitly make use of the following two heuristics to narrow down the search space:

- Languages tend to have a small number of paradigms (where "small" means fewer than 100 paradigms with at least 100 member stems each).
- Languages tend to have only small paradigms (where "small" means fewer than 50), that is, the number of affixes in each paradigm is small. Agglutinative languages, which have several layers of affixes, can be said to obey this generalization in the sense that each layer has few members, whereas conversely, the full paradigm achieves considerable size combinatorially.

Although we know of no empirical evulation of them, in the impression of the present authors, the two heuristics appear to be cross-linguistically valid.

Chan (2006) is an exceptionally clean study of inducing paradigms, assuming that the segmentation is already given. The problem then takes the form of a matrix with

Table 8
The three most common paradigms in Swedish according to the SALDO lexicon and morphological resources (Borin, Forsberg, and Lönngren 2008), as computed on the SUC 1.0 corpus (Ejerhed and Källgren 1997) of 55,000 word types.

Adjective 1st decl (e.g., *gul* 'yellow')		Noun 3rd decl (e.g., *tid* 'time')		Verb 1st conj (e.g., *lag-* 'fix')	
-a	2022	-"	1619	-a	1001
-"	1821	-en	1141	-ade	948
-t	1572	-er	1072	-ar	883
-e	221	-erna	583	-at	579
-are	208	-s	310	-as	482
-s	114	-ens	259	-ande	423
-aste	90	-ernas	136	-ad	387
-ast	46	-ers	40	-ades	273
-as	39			-ats	207
-es	13			-andes	5
-ts	4			-ads	3
-ares	1				

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

stems on one axis and suffixes on the other axis. Chan then makes use of known techniques from linear algebra, in particular Latent Dirichlet Allocation, to break the full matrix into smaller dense submatrices, which, when multiplied together, resemble the full matrix. There is only one humanly tuned threshold, namely, when to stop breaking into smaller parts.

3.3 Group and Abstract

In contrast to the methods that use a heuristic for finding morpheme boundaries, the grouping methods are much less sensitive to continuous segments. String edit distance is the most straightforward metric for which to find pairs or sets of morphologically related words (see, e.g., Gaussier 1999; Yarowsky and Wicentowski 2000; Schone and Jurafsky 2001a; Baroni, Matiasek, and Trost 2002; Hu et al. 2005a; Bernhard 2006, pages 101–117; Bernhard 2007; Majumder et al. 2007; Majumder, Mitra, and Pal 2008). In addition, as unsupervised methods for semantic clustering (e.g., Latent Semantic Analysis) and distributional clustering became more mature, these could be included as well (Schone and Jurafsky 2000, 2001a; Schone 2001; Baroni, Matiasek, and Trost 2002; Freitag 2005). More remarkable, however, is that Yarowsky and Wicentowski (2000) and Wicentowski (2002, 2004) have shown that frequency signatures can also be used to (heuristically) find morphologically related words. The example they use is *sang* versus *sing*, whose relative frequency distribution in a corpus is 1,427/1,204 (or 1.19/1), whereas *singed*[15] versus *sing* is 9/1,204 (Yarowsky and Wicentowski 2000, pages 209–210). This way, *sing* can be heuristically said to be parallel to *sang* rather than *singed*, and indeed the distribution for *singed* versus *singe* (its true relative) is 9/2, that is, much closer to 1.

Suppose now that groups of morphologically related words are somehow heuristically extracted. For example, one group might be {*play*, *player*, *played*, *playing*} and another might be {*bark*, *barks*, *barked*, *barking*}. The next step would be to find what is common among several groups, not just one. Abstracting morphological alternations given a family of groups is a thorny issue. For instance, Baroni, Matiasek, and Trost (2002) leave the matter largely in the exploration phase. Wicentowski (2004) presents a finished theory based on constraining the abstraction to find patterns in terms of prefix, suffix, and stem alternations.

The outstanding question for the group-and-abstract approaches, related not only to grouping but also to abstracting, is how to find one and the same morphological process (umlauting, adding a suffix, etc.) that operates over a maximal number of groups. The search space is huge, considering not only the group space but also the large number of potential morphological processes itself.

The group-and-abstract approaches are also characterized by the ubiquitous use of ad hoc thresholds. However, there are clear advantages in that they are in principle capable of handling non-concatenative morphology and in that issues of semantics (of stems) are addressed from the beginning.

The work by de Kock and Bossaert (1969, 1974, 1978), Yvon (1996), Medina Urrea (2003) and partly Moon, Erk, and Baldridge (2009) can favorably be seen as a mid-way between the border-and-frequency and group-and-abstract approaches as they rely on

15 That is, the past tense of the verb *singe*.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Table 9
Example feature values for the words *ngĩthiĩ* (I went) and *tũgĩthiĩ* (we went) adapted from De Pauw and Wagacha (2007, page 1518). B=-features describe a subset at the start of the word form, E=-features indicate patterns at the end of the word, and I=-features describe patterns inside the word form.

class	features
ngĩthiĩ	B=n B=ng B=ngĩ B=ngĩt B=ngĩth B=ngĩthi I=g I=gĩ I=gĩt I=gĩth I=gĩthi E=gĩthiĩ I=ĩ I=ĩt I=ĩth I=ĩthi E=ĩthiĩ I=t I=th I=thi E=thiĩ I=h I=hi E=hiĩ I=i E=iĩ
tũgĩthiĩ	B=t B=tũ B=tũg B=tũgĩ B=tũgĩt B=tũgĩth B=tũgĩthi I=ũ I=ũg I=ũgĩ I=ũgĩt I=ũgĩth I=ũgĩthi E=ũgĩthiĩ I=g I=gĩ I=gĩt I=gĩth I=gĩthi E=gĩthiĩ I=ĩ I=ĩt I=ĩth I=ĩthi E=ĩthiĩ I=t I=th I=thi E=thiĩ I=h I=hi E=hiĩ I=i E=iĩ

sets of four members with a particular affixation arrangement ("squares"),[16] whose existence is governed much by the frequency of the affixes in question.

3.4 Features and Classes

The features-and-classes methods share with the group-and-abstract methods the virtue of not being tied to segmental morpheme choices. As mentioned earlier, in this family of methods a word is seen as made up of a set of features which have no internal order—*n*-grams in Mayfield and McNamee (2003) and McNamee and Mayfield (2007), and beginning/terminal/internal segments in De Pauw and Wagacha (2007).

For example, Table 9 shows two words and their features in Gĩkũyũ, a tonal Bantu language of Kenya. As designed by De Pauw and Wagacha (2007), initial (B=), middle (I=), or final (E=) segments of a given word constitute its features. A majority of features enumerated this way will not be morphologically relevant, whereas a minority is. For example, in this case, I=h is just an arbitrary character without morpheme status, whereas I=*ngĩthi* happens to be equal to a stem. The idea is that arbitrary features such as I=h will be too common in the training data to provide a useful constraint, whereas a more specialized feature like I=*ngĩthi* might indeed trigger useful morphological generalization properties.

The input word list W thus transforms into a training set of word–feature pairs, which can be fed into a standard maximum entropy classifier. The next step is, for each word, to ask the classifier for the k closest classes, namely, words (which will include the word itself and $k-1$ others with significant feature overlap). Clearly, such clusters may capture relations that string-edit-distance clustering does not. De Pauw and Wagacha

16 Based on the famous **Greenberg square**, which is a concrete means of illustrating the minimal requirement for postulating a paradigm: We need a minimum of two attested stems and two attested suffixes (or in the general case arbitrary morphological processes), where both stems must occur with both suffixes:

$stem_A{+}sx_1$ $stem_B{+}sx_1$
$stem_A{+}sx_2$ $stem_B{+}sx_2$

As it is used in linguistics, both the stems and the suffixes in the square must represent attested form–meaning combinations. This requirement is normally given up in the work reviewed here, where only the form of postulated stems and affixes is available, but not the meaning. The Greenberg square goes back to the age-old linguistic notion of proportional analogy (Anttila 1977).

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

(2007, pages 1517–1518) further suggest how specific morphological information, such as prefixes, tonal changes, etc., may be abstracted from such clusters.

Clearly, feature-based methods provide an interesting new avenue for non-segmental and long-distance phenomena, but are so far largely unexplored and not free from thresholds and parameters.

3.5 Phonological Categories and Separation

These approaches specifically target the special kind of non-concatenative morphology called intercalated morphology (or templatic morphology or root-and-pattern morphology) famous mainly from Semitic languages, such as Arabic. They start out by assuming that graphemes can be subdivided into those that take part in the root, and those that take part in the pattern. For the languages so far targeted, Arabic (Rodrigues and Ćavar 2005, 2007; Xanthos 2007) and Amharic (Bati 2002), this is largely true, or a transcription is used where it is largely true. Rodrigues and Ćavar (2005, 2007) and Bati (2002) hard-code the transition from the graphemic representation of a word to its (potential) root and pattern parts. This can be said to constitute a strong language specific bias, tantamount to supervision. Xanthos (2007), on the other hand, starts out only by assuming that there *exists* a distinction between root and pattern graphemes and subsequently learns which graphemes are which. See Goldsmith and Xanthos (2009) for an excellent survey on how to do this (something which falls under learning phonological categories rather than morphology learning). Basically, it is possible only because there are systematic combination constraints between different phonemes (approximated by graphemes); for example, vowels and consonants alternate in a very non-random manner.

Once each word is divided into its potential root and pattern, the morphology learning problem is similar to morphology learning given roots and suffixes, that is, the typical model for learning concatenative morphology, where the task is to weed out noise, to decide where patterns ("suffixes") start and end, which patterns are spurious, and so on. All these authors who have addressed intercalated morphology use a variant of MDL (see the border-and-frequency techniques in Section 3.2). The accuracy of ULM on languages with intercalated morphology appears to be similar to the accuracy on other languages (cf. Section 4.3).

3.6 General Strengths and Weaknesses

A perhaps worrying tendency is that, despite extensive cross-citation, there is little transfer between different groups of authors and there is a fair amount of duplication of work. The lack of a broadly accepted theoretical understanding is possibly related to this fact. Few approaches have an abstract model of how words are formed, and thus cannot explain why (or why not) the heuristics employed fail, what kind of errors are to be expected, and how the heuristics can be improved. Nevertheless, a model for the simplest kind of concatenative morphology is emerging, namely, that two sets of random strings, B and S, combine in some way to form a set of words W. For Gelbukh, Alexandrov, and Han (2004), the segmentation task is to find minimal size $|X| + |Y|$ such that $W \subset \{xy | x \in X, y \in Y\}$. For example, if $W = \{ad, ae, bd, be, cd, ce\}$, then the minimal size $|X| + |Y| = 5$ with $X = \{a, b, c\}$ and $Y = \{d, e\}$. For Bacchin, Ferro, and Melucci (2005) as well as in the word-segmentation version of Deligne (1996), the segmentation task is to find a configuration of splits s_i for each $w_i = x_i y_i \in W$ such that each x_i and

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

y_i occur in as many splits as possible. More precisely, the product, over all words, of the number of splits for the parts x and y should be maximized. Formally, let $x_i y_i$ be the parts of w_i induced by splits s_i and let $p(x) = |\{i|x = x_i\}| = |\{w_i|xy_i = w_i\}|$ be the number of words in which x equals the first part of the split and similarly let $p'(y) = |\{i|y = y_i\}| = |\{w_i|x_i y = w_i\}|$ be the number of words in which y equals the last part of the split. Then the task is to find splits that maximize the following expression:

$$arg \max_{[s_1,\ldots,s_{|W|}]} \prod_{w_i \in W} p(x_i) \cdot p'(y_i)$$

For example, if $W = \{ad, ae, bd, be, cd, ce, ggg\}$, then the configuration of splits $a|d$, $a|e$, $b|d$, $b|e$, $c|d$, $c|e$, $g|gg$ yields the product $(2 \cdot 3)^6 \cdot (1 \cdot 1)$.

Brent (1999) devises a precise, but more elaborate, way of constructing W from B and S, but at the cost of a large search space, and whose global maximum is hard to characterize intuitively. The same holds for the extension by Snover (2002). Kontorovich, Don, and Singer (2003), Snyder and Barzilay (2008), Goldwater (2007), Johnson (2008), and Poon, Cherry, and Toutanova (2009) should also be noted for containing generative models.

Most approaches, of any of the kinds (a)–(d) described in Section 3.1, explicitly or implicitly target languages which have (close to) one-slot morphology, that is, a word (or stem) typically takes not more than one prefix and not more than one suffix. Many (indeed most; Dryer 2005) languages deviate more or less from this model. At first, it may seem that multi-slot morphology can be handled by the same algorithms as one-slot morphology, by iterating the process used for one-slot morphology. A decade of ULM has shown that the matter is not so simple, because heuristics for one slot languages do not necessarily generalize to the outermost slot of a multi-slot language.

The (c) and (d) approaches do not combine easily with the others but it is conceivable that the (a) and (b) type of approaches may be mutually enhancing. Results from the (a) methods may serve to cut down the search space for the (b) methods, and the (b) methods may provide a way to circumvent thresholds for the (a) methods. There is also the possibility of serial combination where, for example, the (a) methods target concatenative morphology and the (b)—or (c)—methods attempt the remaining cases. Presumably because most methods so far do not produce a clean, well-defined result, various forms of hybridization of techniques by different authors have yet to be systematically explored.

Lastly, there are scattered attempts to address morphophonological changes in a principled way, though so far these have been developed in close connection with a particular segmentation method and target language (Schone 2001; Schone and Jurafsky 2001a; Wicentowski 2002, 2004; Tepper 2007; Kohonen, Virpioja, and Klami 2008; Tepper and Xia 2008).

4. Discussion

4.1 Language Dependence of ULM

As we mentioned in Section 3.6, most approaches have an explicit bias towards certain kinds of morphological systems, those for which we introduced the label "one-slot morphology." This is of course not a problem, if the purpose is to bootstrap a morphology for some languages which happen to belong to the right type. If the purpose is to say something about human language acquisition or language learning, or if the aim is to

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

devise a method that should work with any language, such a bias naturally becomes problematic.

The two human learning analogues which have most frequently been proposed in the literature on ULM and other machine learning of morphology are those of **language acquisition** and of **linguistic analysis** (e.g., as carried out as part of linguistic fieldwork). Depending on which of the two we choose, the kinds of biases that we may or may not allow become different. Language acquisition in humans is oral (or sign, but for practical reasons, we are leaving sign languages completely out of the discussion here), so expecting written input with word delimiters would then be an inadmissible bias. ULM as delimited in Section 1 is definitely closer to linguistic analysis than to language acquisition.

It may be instructive at this point to see what kinds of knowledge are supposed to be required in order to carry out the discovery procedures mentioned in Section 2:

> An analyst approaches a language which either he already knows in some practical way or with which he sets about to familiarize himself—preferably in a language learning situation. The analyst's background is the sum total of his practical knowledge of other languages, his previous analytical experience, and what he has learned from the linguistic research of other people. With this knowledge of the language to be analyzed and with this background knowledge, he makes certain guesses about the grammatical structure of the language. He then submits these guesses to a series of systematic checks in which he confirms, disproves, or modifies his original guesses—and makes a few better guesses en route. This systematic evaluation is based on a theory of the structure of language, and the theory itself (while containing elements of creative thinking) is based on empirical study. (Longacre 1964, page 12)

Mutatis mutandis, the procedure described in this quote, contains most elements of ULM and related methods proposed in the literature. Note that the quote just given stresses the importance of the knowledge that the linguist brings to the analysis and which informs the whole analytical process. This suggests that there may be a level of general knowledge about language (in general or a useful subset of languages), or about linguistic analysis, or both, which would be useful to ULM in general, something like the "knowledge" that white space is a word delimiter in written text, but on a higher level. One component of a research program on ULM would then be to formulate this kind of general knowledge in a way which makes sense given that the object of study is language, to test it, and to share it with the community of linguistic scholars. A concrete illustration could be the way that the old notion of proportional analogy (Anttila 1977) is refined and formalized in various ways and used to test segmentation hypotheses in works on ULM from the earliest times onwards (e.g., the "squares" mentioned in Section 3.3).

4.2 ULM and Semantics

As traditionally conceived, an inflectional paradigm links a set of word forms to structural descriptions expressed in terms of a stem carrying a lexical meaning and some formal expression of one or more morphosyntactic categories (or grammatical meanings) taken from a closed, small set of such categories. This bears emphasizing, because ULM work generally has been concerned only with the formal expression side of morphology; that is, instead of the traditional

table-s 'table N PL'

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

table-s 'table V 3SG'

it will give us simply

table-s

table-s

although it may tell us that the stem *table* appears in two paradigms.

As far as we know, there have been no attempts to induce functional labels using ULM, although it is conceivable that the same kind of techniques used, for example, in order to cluster words semantically (e.g., Latent Semantic Indexing/Analysis or Random Indexing), could be used also to classify the resulting morphs from a ULM segmentation (cf. Schone and Jurafsky [2001b] for a study of inducing part-of-speech class labels in a setting similar to that of ULM). The labelling problem can easily be considered independent of ULM by using a hand-segmented (or segmented by a hand-built morphological parser) input corpus.

4.3 Is ULM of Any Use?

As we said in Section 2, there is an explicit expectation frequently encountered in the more recent literature that ULM and other unsupervised methods could be employed in order to rapidly and cheaply (in terms of human effort) bootstrap basic language technology resources for new languages. However, looking at the literature, it seems that—at least in the area of inflectional morphology—the only approaches that have so far produced substantial results are the old-fashioned, hand-coded grammar-based ones, such as the work described by Trosterud (2004), where finite-state morphological processors and constraint grammar-based disambiguation components are developed for a number of related languages. The fact that the languages are related is of great help when dealing with successive languages after the first one. The morphological component for the first language, North Sámi, required approximately 2.5 person-years of highly qualified linguistic expert work to reach the prototype stage, whereas the analogous module for the closely related Lule Sámi was completed in an additional six months (Trosterud 2006).[17] This and other work in the same vein reported in the literature (e.g., by Artola-Zubillaga 2004 and Maxwell and David 2008) is characterized by deep and long-lasting involvement by linguistic expertise and further often by the creative use of digitized versions of conventional printed linguistic resources, especially dictionaries. The following observation is perhaps trivial, but bears stressing, because it is in fact often not heeded in practice: For this kind of approach to work, it is necessary that tools for providing systems with linguistic knowledge use a conceptual apparatus and notation familiar to the linguists who are supposed to be working with them. Relevant to our purposes here, the same holds for any attempt to kickstart the development of a morphological analyzer by using ULM: If the expectation is that the output of ULM should be manually "post-edited," this output must of course be intelligible to the linguist doing the post-editing.

17 As pointed out by one anonymous reviewer, this suggests that with the right organization of information flow among machine-learning components, ULM, too, could benefit from working with several closely related languages simultaneously.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Most ULM approaches reported in the literature are small proof-of-concept experiments, which generally founder on the lack of evaluation data. The MorphoChallenge[18] series does provide adequate gold-standard evaluation data for Finnish, English, German, Arabic, and Turkish as well as task-based Information Retrieval (IR) evaluation data for English, German, and Finnish. It can be seen that ULM systems are mature enough to enhance IR, but so far, ULM systems are not close to full accuracy on the gold standard and outside commentators have generally been unimpressed with these results (e.g., Mahlow and Piotrowski 2009, page vi). However, many (most?) of the strong-looking systems reported in the literature have not, for one reason or another, taken part in the MorphoChallenge. Taking MorphoChallenge results and proof-of-concept reports together, it seems that high accuracy by ULM systems is presently only achievable if the language has small amounts of one-slot concatenative morphology, whereas for morphologically more complex languages, parameter tuning and/or lower accuracy is to be expected.

We are not yet in a position to assess whether there are other tasks than IR which, in general, benefit significantly from (noisy) ULM, such as Speech Recognition (Hirsimäki et al. 2003, 2005, 2006; Kurimo et al. 2006) or Machine Translation (Sereewattana 2003; Virpioja et al. 2007; Bojar, Straňák, and Zeman 2008; Kirik and Fishel 2008; De Gispert et al. 2009; Fishel and Kirik 2010) because almost only the Morfessor system has been tested, and results are, if positive, not completely unambiguous. One usage of noisy ULM, at least, is for smoothing language identification models (Hammarström 2007a; Ceylan and Kim 2009).

Further, ULM approaches are data-hungry, which precludes their use with many low-density languages. There is much ongoing work addressing these issues, however, so we can probably expect some progress in this area (Bird 2009).

4.4 Future Directions

In practice, the near future should define a high-accuracy threshold-minimal system for one-slot morphology languages, using refinements of ideas already extant.

A major challenge, and the reason for duplication of work in the past, is to find a theory that explains why (or why not) a given algorithm works. Further study of theoretical properties of (stochastic) combining of string sets/bags are likely to hold the key to the culmination of the border-and-frequency methods—not further experimentation with ad hoc heuristics. The recent increased interest in Bayesian generative models in general in NLP may possibly serve as a catalyst.

In the group-and-abstract paradigm, working with feature sets of a word, as in De Pauw and Wagacha (2007), is an ingenious generalization that holds numerous advantages over string edit distances. Feature set comparisons are naturally defined over arbitrary collections, whereas string edit distances work on *pairs* of strings. Many morphologically related words differ in several characters and are therefore not particularly close in edit distance. Features instead of edit distances provide a neat framework, based on global properties of the feature distribution, of capturing the fact that some character mismatches do not really matter, whereas some character matches (although not necessarily long) are very significant.

For paradigm induction, it is clear that the ULM field has not made use of the large literature on clustering in other fields. Chan (2006) is a step in this direction, but further

18 Web site `www.cis.hut.fi/morphochallenge2009/` accessed 10 September 2009.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

steps are lacking; in particular, spectral clustering (of some kind) has not been explored for paradigm induction in ULM. Also here, given the typical skewed stem distributions and skewed suffix distributions (exemplified in Section 3.2), some theoretical work is needed to determine its implications for clustering.

Finally, we see ample opportunity for empirical investigations into lesser-known languages for which data has become available only recently (Abney and Bird 2010). This would clarify the potential of ULM usefulness for underdescribed and under-resourced languages.

5. Conclusion

After more than half a century of research, the field of ULM has made good progress (as have many other areas of computational linguistics), but there is still a long way to go before it will become practically useful or even theoretically interesting to linguists. In the terms of Table 1 in Section 1, the state of the art of ULM is somewhere in the region of "Segmentation" and "Inflection tables," if we are talking about linguistic form, but there has been next to no progress at all when it comes to linguistic meaning (e.g., functional labeling of affixes).

In the early days of ULM, the expectation was that it should constitute—when eventually achieved sometime in the future—a formalized version of a linguistic discovery procedure, that is, a knowledge-heavy enterprise. Instead, recent successes in the area have been largely contingent upon the rapid development in computational linguistics of statistical and information-theoretic knowledge-light (but robust) methodologies.

We believe (like Wintner 2009 for computational linguistics in general), however, that if ULM is to become a serious alternative to—or, equally likely, a natural component of—manually built computational morphology systems for a wide and diverse range of languages, and especially if we are to make headway in the area of semantics, we need to see more interaction between the present approaches to ULM with the computational techniques and mathematical modeling tools they can bring to bear on the problem on the one hand, and typologically informed linguistic research on morphology founded on a vast store of knowledge and methodology refined over two millennia on the other.

Acknowledgments

The authors wish to thank three anonymous referees for helpful comments and suggestions.

References

Abney, Steven and Steven Bird. 2010. The human language project: Building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 88–97, Uppsala.

Altmann, Gabriel and Werner Lehfeldt. 1980. *Einführung in die quantitative Phonologie*, volume 7 of *Quantitative Linguistics*. Bochum: Brockmeyer.

Andreev, Nikolaj Dmitrievič. 1959. Modelirovanije jazyka na base ego statističeskoj i teoretiko-množestvennoj struktury. In *Tezisy soveščanija po matematičeskoj lingvistike, 14–21 Aprelja 1959 goda*. Ministerstvo vysšego obrazovanija SSSR, Leningrad, pages 15–22.

Andreev, Nikolaj Dmitrievič. 1963. Algoritmy statistiko-kombinatornogo modelirovanija morfologii, sintaksisa, slovoobrazovanija i semantiki. In *Materialy po matematičeskoj lingvistike i mašinomu perevodu: Sbornik II*. Izdatel'stvo Leningradskogo universiteta, Leningrad, pages 3–44.

Andreev, Nikolaj Dmitrievič. 1965a. Opyt statistiko-kombinatornogo vydelenija pervogo morfologičeskogo tipa v vengerskom jazyke. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 205–211.

Andreev, Nikolaj Dmitrievič, editor. 1965b. *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Andreev, Nikolaj Dmitrievič. 1967. *Statistiko-kombinatornye metody v teoretičeskom i prikladnom jazykovedenii*. Nauka, Leningrad.

Andreeva, L. D. 1963. Statistiko-kombinatornoe vydelenie paradigmy pervogo mofologičeskogo tipa v russkom jazyke. In *Materialy po matematičeskoj lingvistike i mašinnomu pervodu: Sbornik II*. Izdatel'stvo Leningradskogo universiteta, Leningrad, pages 45–60.

Anttila, Raimo. 1977. *Analogy*. Mouton, The Hague.

Antworth, Evan L. 1990. *PC-KIMMO: A two-level processor for morphological analysis. Occasional Publications in Academic Computing 16*. Summer Institute of Linguistics, Dallas.

Arabsorkhi, Mohsen and Mehrnoush Shamsfard. 2006. Unsupervised discovery of Persian morphemes. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2006*, pages 175–178, Trento.

Argamon, Shlomo, Navot Akiva, Amihood Amir, and Oren Kapah. 2004. Efficient unsupervised recursive word segmentation using minimum description length. In *Proceedings of COLING 2004*, pages 1058–1064, Geneva.

Artola-Zubillaga, Xabier. 2004. Laying lexical foundations for NLP: The case of Basque at the *ixa* research group. In *SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages*, pages 9–18, Lisbon.

Atwell, Eric and Andrew Roberts. 2005. Combinatory hybrid elementary analysis of text. In *Proceedings of MorphoChallenge 2005*, pages 37–41, Helsinki University of Technology, Helsinki.

Baayen, Harald R. 2001. *Word Frequency Distributions*, volume 18 of *Text, Speech, and Language Technology*. Kluwer, Dordrecht.

Bacchin, Michela, Nicola Ferro, and Massimo Melucci. 2002a. The effectiveness of a graph-based algorithm for stemming. In *ICADL '02: Proceedings of the 5th International Conference on Asian Digital Libraries*, volume 2555 of *Lecture Notes in Computer Science*, pages 117–128, Singapore.

Bacchin, Michela, Nicola Ferro, and Massimo Melucci. 2002b. University of Padua at CLEF 2002: Experiments to evaluate a statistical stemming algorithm. In *Working Notes for CLEF 2002: Cross-Language Evaluation Forum Workshop*, pages 161–168, Rome.

Bacchin, Michela, Nicola Ferro, and Massimo Melucci. 2005. A probabilistic model for stemmer generation. *Information Processing and Management*, 41(1):121–137.

Baroni, Marco. 2000. *Distributional Cues in Morpheme Discovery: A Computational Model and Empirical Evidence*. Ph.D. thesis, University of California, Los Angeles.

Baroni, Marco. 2003. Distribution-driven morpheme discovery: A computational/ experimental study. *Yearbook of Morphology*, 213–248.

Baroni, Marco, Johannes Matiasek, and Harald Trost. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL/SIGPHON-2002*, pages 48–57, Philadelphia.

Batchelder, E. O. 1997. *Computational Evidence for the Use of Frequency Information in Discovery of the Infant's First Lexicon*. Ph.D. thesis, City University of New York.

Bati, Tesfaye Bayu. 2002. Automatic morphological analyser: An experiment using unsupervised and autosegmental approach. Master's thesis, Addis Ababa University, Ethiopia.

Belkin, Mikhail and John Goldsmith. 2002. Using eigenvectors of the bigram graph to infer morpheme identity. In *Morphological and Phonological Learning: Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 41–47, Philadelphia, PA.

Bernhard, Delphine. 2005a. Segmentation morphologique à partir de corpus. *Actes de TALN & RÉCITAL 2005*, volume 1, pages 555–564, Dourdan.

Bernhard, Delphine. 2005b. Unsupervised morphological segmentation based on segment predictability and word segments alignment. In Mikko Kurimo, Mathias Creutz, and Krista Lagus, editors, *Unsupervised segmentation of words into morphemes – Challenge 2005*, pages 18–22, Helsinki University of Technology, Helsinki.

Bernhard, Delphine. 2006. *Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales*. Ph.D. thesis, Université Joseph Fourier – Grenoble I.

Bernhard, Delphine. 2007. Apprentissage non supervisé de familles morphologiques par classification ascendante hiérarchique. In *Actes de la 14e conférence sur le Traitement*

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Automatique des Langues Naturelles, TALN 2007, volume 1, pages 367–376, Toulouse.

Bernhard, Delphine. 2008. Simple morpheme labelling in unsupervised morpheme analysis. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval*, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19–21, 2007, Revised Selected Papers, pages 873–880. Springer-Verlag, Berlin.

Bharati, Akshar, S. M. Bendre Rajeev Sangal, Pavan Kumar, and Aishwarya. 2001. Unsupervised improvement of morphological analyzer for inflectionally rich languages. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS-2001)*, pages 685–692, Tokyo.

Bickel, Balthasar, Goma Banjade, Martin Gaenszle, Elena Lieven, Netra Paudyal, Ichchha Rai, Manoj Rai, Novel Kishor Rai, and Sabine Stoll3. 2007. Free prefix ordering in Chintang. *Language*, 83(1):43–73.

Bird, Steven. 2009. Last words: Natural language processing and linguistic fieldwork. *Computational Linguistics*, 35(3):469–474.

Blomqvist, Jerker and Poul Ole Jastrup. 1998. *Grekisk grammatik: Graesk grammatik*, 2 edition. Akademisk Forlag, København.

Bloomfield, Leonard. 1933. *Language*. Henry Holt & Co, New York.

Bojar, Ondřej, Pavel Straňák, and Daniel Zeman. 2008. English–Hindi translation in 21 days. In *Proceedings of the ICON-2008 NLP Tools Contest*, pages 4–7, Pune.

Bordag, S. 2005a. Unsupervised knowledge-free morpheme boundary detection. Paper presented at the *Proceedings of Recent Advances in Natural Language Processing 2005 (RANLP '05)*, Borovets.

Bordag, Stefan. 2005b. Two-step approach to unsupervised morpheme segmentation. In Mikko Kurimo, Mathias Creutz, and Krista Lagus, editors, *Unsupervised segmentation of words into morphemes – Challenge 2005*, pages 23–27, Helsinki University of Technology, Helsinki.

Bordag, Stefan. 2007. *Elements of Knowledge-Free and Unsupervised Lexical Acquisition*. Ph.D. thesis, University of Leipzig, Leipzig.

Bordag, Stefan. 2008. Unsupervised and knowledge-free morpheme segmentation and analysis. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval*, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19–21, 2007, Revised Selected Papers, pages 881–891. Springer-Verlag, Berlin.

Borin, Lars. 1991. *The Automatic Induction of Morphological Regularities*. Ph.D. thesis, Uppsala University.

Borin, Lars. 2009. One in the bush: Low-density language technology. Research Reports from the Department of Swedish, No. GU-ISS-09-1. University of Gothenburg.

Borin, Lars, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the BLARK - SALDO, a freely available lexical database for Swedish language technology. In Joakim Nivre, Mats Dahllöf, and Beáta Megyesi, editors, *Resourceful language technology: Festschrift in honor of Anna Sågvall Hein*, volume 7 of *Studia Linguistica Upsaliensia*. Acta Universitatis Upsaliensis, Uppsala, pages 21–32.

Brasington, Ron, Steve Jones, and Colin Biggs. 1988. The automatic induction of morphological rules. *Literary and Linguistic Computing*, 3(2):71–78.

Brent, Michael. 1993. Minimal generative explanations: A middle ground between neurons and triggers. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pages 28–36, Boulder, CO.

Brent, Michael R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105.

Brent, Michael R., S. Murthy, and A. Lundberg. 1995. Discovering morphemic suffixes: A case study in minimum description length induction. In *Fifth International Workshop on Artificial Intelligence and Statistics*, pages 482–490. Fort Lauderdale, FL.

Calderone, Basilio. 2008. *Unsupervised Learning of Linguistic Structures*. Ph.D. thesis, Scuola Normale Superiore, Pisa.

Carstairs, Andrew. 1983. Paradigm economy. *Journal of Linguistics*, 19:115–125.

Ćavar, Damir, Joshua Herring, Toshikazu Ikuta, Paul Rodrigues, and Giancarlo Schrementi. 2004a. On induction of morphology grammars and its role in bootstrapping. In *Proceedings of Formal*

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Grammar 2004, pages 47–62, ESSLLI, Nancy, France.

Ćavar, Damir, Joshua Herring, Toshikazu Ikuta, Paul Rodrigues, and Giancarlo Schrementi. 2004b. On statistical parameter setting. In *Proceedings of the First Workshop on Psycho-Computational Models of Human Language Acquisition*, pages 9–16, Geneva.

Ćavar, Damir, Jushua Herring, Toshikazu Ikuta, Paul Rodrigues, and Giancarlo Schrementi. 2006. On unsupervised grammar induction from untagged corpora. In P. Kaszubski, editor, *PSiCL: Poznan' Studies in Contemporary Linguistics*, volume 41. Poznan', Poland: Adam Mickiewicz University, pages 57–71.

Ćavar, Damir, Paul Rodrigues, and Giancarlo Schrementi. 2006. Unsupervised morphology induction for part-of-speech tagging. In Aviad Eilam, Tatjana Scheffler, and Joshua Tauberer, editors, *Proceedings of the 29th Annual Penn Linguistics Colloquium*, volume 12(1) of *U. Penn Working Papers in Linguistics*. University of Pennsylvania Press, Philadelphia, pages 29–41.

Ceylan, Hakan and Yookyung Kim. 2009. Language identification of search engine queries. In *ACL-IJCNLP '09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, pages 1066–1074, Morristown, NJ.

Chan, Erwin. 2006. Learning probabilistic paradigms for morphology in a latent class model. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL 2006*, pages 69–78, New York City, NY.

Chan, Erwin. 2008. *Structures and Distributions in Morphology Learning*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Cho, Sehyeong and Seung-Soo Han. 2002. Automatic stemming for indexing of an agglutinative language. In T. Yakhno, editor, *Advances in Information Systems*, volume 2457 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, pages 154–165.

Clark, Alexander. 2001. *Unsupervised Language Acquisition*. Ph.D. thesis, University of Sussex.

Creutz, Mathias. 2003. Unsupervised segmentation of words using prior distributions of morph length and frequency. In *Proceedings of the ACL 2003*, pages 280–287, Sapporo.

Creutz, Mathias. 2006. *Induction of the Morphology of Natural Language: Unsupervised Morpheme Segmentation with Application to Automatic Speech Recognition*. Ph.D. thesis, Helsinki University of Technology, Espoo, Finland.

Creutz, Mathias and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 21–30, Philadelphia, PA.

Creutz, Mathias and Krista Lagus. 2004. Induction of a simple morphology for highly-inflecting languages. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51, Barcelona.

Creutz, Mathias and Krista Lagus. 2005a. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR '05)*, pages 106–113, Espoo.

Creutz, Mathias and Krista Lagus. 2005b. Morfessor in the Morpho Challenge. In Mikko Kurimo, Mathias Creutz, and Krista Lagus, editors, *Unsupervised segmentation of words into morphemes – Challenge 2005*, pages 12–17, Helsinki University of Technology, Helsinki.

Creutz, Mathias and Krista Lagus. 2005c. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Technical report A81, Publications in Computer and Information Science, Helsinki University of Technology.

Creutz, Mathias and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1–3):1–33.

Creutz, Mathias, Krista Lagus, Krister Lindén, and Sami Virpioja. 2005. Morfessor and hutmegs: Unsupervised morpheme segmentation for highly-inflecting and compounding languages. In *Proceedings of the Second Baltic Conference on Human Language Technologies*, pages 107–112, Tallinn.

Creutz, Mathias, Krista Lagus, and Sami Virpioja. 2005. Unsupervised morphology induction using morfessor. In *Finite State*

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Methods in Natural Language Processing: 5th International Workshop, FSMNLP 2005, pages 300–301, Helsinki.

Cromm, Oliver. 1997. Affixerkennung in deutschen wortformen: Ein nicht-lexikalisches segmentierungsverfahren nach N. D. Andreev. *LDV-Forum*, 14(2):4–13.

Cucerzan, Silviu and David Yarowsky. 2002. Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings of CoNLL-2002*, pages 1–7, Taipei.

Daelemans, Walter. 2004. Computational linguistics. In Geert Booij, Christian Lehmann, Joachim Mugdan, and Stavros Skopetas, editors, *Morphologie/Morphology: Ein internationales Handbuch zur Flexion und Wortbildung [An International Handbook on Inflection and Word-Formation]*, volume 17.2 of *Handbücher zur Sprach- und Kommunikationswissenschaft*. Mouton de Gruyter, Berlin, pages 1893–1900.

Dang, Minh Thang and Saad Choudri. 2005. Simple unsupervised morphology analysis algorithm (SUMAA). In Mikko Kurimo, Mathias Creutz, and Krista Lagus, editors, *Proceedings of MorphoChallenge 2005*, pages 47–51, Helsinki University of Technology, Helsinki.

Dasgupta, Sajib. 2007. Toward language-independent morphological segmentation and part-of-speech induction. Master's thesis, The University of Texas at Dallas.

Dasgupta, Sajib and Vincent Ng. 2006. Unsupervised morphological parsing of bengali. *Language Resources and Evaluation*, 3–4:311–330.

Dasgupta, Sajib and Vincent Ng. 2007a. High-performance, language-independent morphological segmentation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistic*, pages 155–163, Rochester, NY, Association for Computational Linguistics.

Dasgupta, Sajib and Vincent Ng. 2007b. Unsupervised word segmentation for Bangla. In *Proceedings of the 5th International Conference on Natural Language Processing (ICON 2007)*, pages 15–24, Hyderabad.

De Gispert, Adrià, Sami Virpioja, Mikko Kurimo, and William Byrne. 2009. Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 73–76, Boulder, CO.

de Kock, Josse and Walter Bossaert. 1969. Towards an automatic morphological segmentation. In *International Conference on Computational Linguistics, COLING*, pages 10–11, Sånga-Säby.

de Kock, Josse and Walter Bossaert. 1974. *Introducción a la lingüística automática en las lenguas Románicas*, volume 202 of *Biblioteca románica hispánica 2: Estudios y ensayos*. Gredos, Madrid.

de Kock, Josse and Walter Bossaert. 1978. *The Morpheme: An Experiment in Quantitative and Computational Linguistics*. Van Gorcum, Amsterdam.

De Pauw, Guy and Peter W. Wagacha. 2007. Bootstrapping morphological analysis of Gĩkũyũ using maximum entropy learning. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH 2007)*, pages 1517–1520, Antwerp.

Déjean, Hervé. 1998a. *Concepts et algorithmes pour la découverte des structures formelles des langues*. Ph.D. thesis, Université de Caen Basse Normandie.

Déjean, Hervé. 1998b. Morphemes as a necessary concept for structures discovery from untagged corpora. In *NeMLaP3/CoNLL98 Workshop on Paradigms and Grounding in Language Learning*, pages 295–298, Philadephia, PA.

Deligne, Sabine. 1996. *Modèles de séquences de longueurs variables: application au traitement du langage écrit et de la parole*. Ph.D. thesis, École Nationale Supérieure des Télécommunications, Paris.

Deligne, Sabine and Frédéric Bimbot. 1997. Inference of variable-length linguistic and acoustic units by multigrams. *Speech Communication*, 23(3):223–241.

Demberg, Vera. 2007. A language-independent unsupervised model for morphological segmentation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 920–927, Prague.

Dryer, Matthew S. 2005. Prefixing versus suffixing in inflectional morphology. In Bernard Comrie, Matthew S. Dryer, David Gil, and Martin Haspelmath, editors, *World Atlas of Language Structures*. Oxford University Press, Oxford, pages 110–113.

Eguchi, Paul K. 1987. Fieldworker and computer: An end user's view of computer

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

ethnology. *Senri Ethnological Studies*, 20:165–174.
Ejerhed, Eva and Gunnel Källgren. 1997. Stockholm Umeå Corpus version 1.0, SUC 1.0. Technical report, Department of Linguistics, Umeå University.
Eliseeva, K. A. 1965. Statistiko-kombinatornoe modelirovanie pervogo tipa v ukrainskoj morfologii. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 85–88.
Faulk, R. D. and F. Goertzel Gustavson. 1990. Segmenting discrete data representing continuous speech input. *IBM Systems Journal*, 29(2):287–296.
Fedulova, N. I. 1965. Vydelenie pervogo morfologičeskogo tipa v bolgarskom jazyke. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 110–115.
Fihman, B. S. 1965a. Vydelenie pervogo morfologičeskogo tipa v jazyke hausa po algoritmu statistiko-kombinatornogo modelirovanija. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 189–195.
Fihman, B. S. 1965b. Vydelenie pervogo morfologičeskogo tipa v jazyke suahili po algoritmu statistiko-kombinatornogo modelirovanija. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 196–204.
Fishel, Mark and Harri Kirik. 2010. Linguistically motivated unsupervised segmentation for machine translation. In *Proceedings of the Seventh International Language Resources and Evaluation (LREC'10)*, pages 1741–1745, Valletta.
Fitialova, I. B. 1965. Statistiko-kombinatornoe vydelenie pervogo morfologičeskogo tipa v nemeckom jazyke. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 158–171.
Flenner, Gudrun. 1992. *Ein quantitatives Morphsegmentierungsverfahren für spanische Wortformen*. Ph.D. thesis, Georg-August-Universität Göttingen.
Flenner, Gudrun. 1994. Ein quantitatives Morphsegmentierungssystem für spanische Wortformen. In Ursula Klenk, editor, *Computatio Linguae II: Aufsätze zur algorithmischen und quantitativen Analyse der Sprache*, volume 83 of *Zeitschrift für Dialektologie und Linguistik: Beihefte*. Franz Steiner, Stuttgart, pages 31–62.
Flenner, Gudrun. 1995. Quantitative Morphsegmentierung im Spanischen auf phonologischer Basis. *Sprache und Datenverarbeitung*, 19(2):63–78.
Foley, William. 1991. *The Yimas Language of New Guinea*. Stanford University Press, Stanford, CA.
Forsberg, Markus, Harald Hammarström, and Aarne Ranta. 2006. Lexicon extraction from raw text data. In *Proceedings of the 5th International Conference, FinTAL*, pages 488–499, Turku.
Francis, Nelson W. and Henry Kucera. 1964. Brown corpus. Department of Linguistics, Brown University, Providence, RI.
Freitag, Dayne. 2005. Morphology induction from term clusters. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 128–135, Ann Arbor, MI.
Gammon, Edward. 1969. Quantitative approximations to the word. In *International Conference on Computational Linguistics, COLING*, pages 1–28, Sånga-Säby.
Garvin, Paul L. 1967. The automation of discovery procedure in linguistics. *Language*, 43(1):172–178.
Gaussier, Éric. 1999. Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing at the 37th Annual Meeting of the Association for Computational Linguistics (ACL-1999)*, pages 24–30, Philadephia, PA.
Gelbukh, Alexander F., Mikhail Alexandrov, and Sang-Yong Han. 2004. Detecting inflection patterns in natural language by minimization of morphological model. In Alberto Sanfeliu, José Francisco Martínez Trinidad, and Jesús Ariel Carrasco-Ochoa, editors, *Proceedings of Progress in Pattern Recognition, Image Analysis and Applications, 9th Iberoamerican Congress on Pattern Recognition, CIARP '04*, volume 3287 of *Lecture Notes in Computer Science*. Springer-Verlag, Berlin, pages 432–438.
Gippert, Jost, Nikolaus P. Himmelmann, and Ulrike Mosel, editors. 2006. *Essentials of Language Documentation*. Mouton de Gruyter, Berlin.
Golcher, Felix. 2006. Statistical text segmentation with partial structure analysis. In *Proceedings of KONVENS 2006*, pages 44–51, Konstanz.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Golding, Andrew and Henry S. Thompson. 1985. A morphology component for language programs. *Linguistics*, 23:263–284.
Goldsmith, John. 2000. Linguistica: An automatic morphological analyzer. In *Proceedings from the Main Session of the Chicago Linguistic Society's 36th Meeting*, pages 125–139, Chicago, IL.
Goldsmith, John. 2001. Unsupervised learning of the morphology of natural language. *Computational Linguistics*, 27(2):153–198.
Goldsmith, John, Derrick Higgins, and Svetlana Soglasnova. 2001. Automatic language-specific stemming in information retrieval. In Carol Peters, editor, *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2000 Workshop*, Lecture Notes in Computer Science. Springer-Verlag, Berlin, pages 273–283.
Goldsmith, John, Yu Hu, Irina Matveeva, and Colin Sprague. 2005. A heuristic for morpheme discovery based on string edit distance. Technical Report of Computer Science Department, University of Chicago, IL. TR-2005-4.
Goldsmith, John and Aris Xanthos. 2009. Learning phonological categories. *Language*, 85(1):4–38.
Goldsmith, John A. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12(4):353–371.
Goldsmith, John A. 2010. Segmentation and morphology. In Alexander Clark, Chris Fox, and Shalom Lappin, editors, *Handbook of Computational Linguistics and Natural Language Processing*, Blackwell Handbooks in Linguistics. Wiley-Blackwell, Oxford, pages 364–393.
Goldwater, Sharon. 2007. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.
Goldwater, Sharon, Tom Griffiths, and Mark Johnson. 2005. Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005]*, pages 459–466, Vancouver.
Golénia, Bruno. 2008. Learning rules in morphology of complex synthetic languages. Master's thesis, Université de Paris V.
Goodman, Sarah A. 2008. Morphological induction through linguistic productivity. In *Working Notes for the CLEF 2008 Workshop*, Aarhus.
Grünwald, Peter D. 2007. *The Minimum Description Length Principle: Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA.
Hadouche, Fadila. 2002. Détection de relations morphologiques en corpus basée sur les cooccurrences. Master's thesis, DESS, Centre de Recherche en Ingénierie Multilingue, CRIM, France.
Hafer, Margaret A. and Stephen F. Weiss. 1974. Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10:371–385.
Hall, Jr., Robert A. 1987. Bloomfield and semantics. In Robert A. Hall, Jr., editor, *Leonard Bloomfield: Essays on his Life and work*. John Benjamins, Amsterdam, pages 155–160.
Hammarström, Harald. 2005. A new algorithm for unsupervised induction of concatenative morphology. In *Finite State Methods in Natural Language Processing: 5th International Workshop, FSMNLP 2005*, pages 288–289, Helsinki.
Hammarström, Harald. 2006a. A naive theory of morphology and an algorithm for extraction. In *SIGPHON 2006: Eighth Meeting of the Proceedings of the ACL Special Interest Group on Computational Phonology*, pages 79–88, New York, NY.
Hammarström, Harald. 2006b. Poor man's stemming: Unsupervised recognition of same-stem words. In *Information Retrieval Technology: Proceedings of the Third Asia Information Retrieval Symposium, AIRS 2006*, pages 323–337, Singapore.
Hammarström, Harald. 2007a. A fine-grained model for language identification. In *Proceedings of iNEWS-07 Workshop at SIGIR 2007*, pages 14–20, Amsterdam.
Hammarström, Harald. 2007b. Unsupervised learning of morphology: Survey, model, algorithm and experiments. Thesis for the Degree of Licentiate of Engineering, Department of Computer Science and Engineering, Chalmers University.
Hammarström, Harald. 2009a. Poor man's word-segmentation: Unsupervised morphological analysis for Indonesian. In *Proceedings of the Third International Workshop on Malay and Indonesian Language Engineering (MALINDO)*, Singapore.
Hammarström, Harald. 2009b. *Unsupervised Learning of Morphology and the Languages of the World*. Ph.D. thesis, Chalmers University of Technology and University of Gothenburg.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Harris, Zellig. 1967. Morpheme boundaries within words: Report on a computer test. In *Transformations and Discourse Analysis Papers 73*. Department of Linguistics, University of Pennsylvania, Philadelphia. Reprinted in Harris 1970.

Harris, Zellig S. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Harris, Zellig S. 1968. Recurrent dependence process: Morphemes by phoneme neighbors. In *Mathematical Structures of Language*, volume 21 of *Interscience tracts in pure and applied mathematics*. Interscience, New York, pages 24–28.

Harris, Zellig S. 1970. Morpheme boundaries within words: Report on a computer test. In Zellig S. Harris, editor, *Papers in Structural and Transformational Linguistics*, volume 1 of *Formal Linguistics Series*. D. Reidel, Dordrecht, pages 68–77.

Haspelmath, Martin. 2002. *Understanding morphology*. Arnold, London.

Hirsimäki, Teemu, Mathias Creutz, Vesa Siivola, and Mikko Kurimo. 2003. Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. In *Proceedings of Eurospeech 2003, Geneva*, pages 2293–2996. Geneva.

Hirsimäki, Teemu, Mathias Creutz, Vesa Siivola, and Mikko Kurimo. 2005. Morphologically motivated language models in speech recognition. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR '05)*, pages 121–126, Espoo.

Hirsimäki, Teemu, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen. 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, 20(4):515–541.

Hol'm, H. A. 1965. Vydelenie pervogo morfologičeskogo tipa v ėstonskom jazyke na osnove statistiko-kombinatornogo modelirovanija. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 212–224.

Hu, Yu, Irina Matveeva, John Goldsmith, and Colin Sprague. 2005a. Refining the SED heuristic for morpheme discovery: Another look at Swahili. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, pages 28–35, Ann Arbor, MI.

Hu, Yu, Irina Matveeva, John Goldsmith, and Colin Sprague. 2005b. Using morphology and syntax together in unsupervised learning. In *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*, pages 20–27, Ann Arbor, MI.

Jacquemin, Christian. 1997. Guessing morphology from terms and corpora. In *Proceedings, 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97)*, pages 155–165, Philadelphia, PA.

Jakubajtis, T. A. 1965. Statistiko-kombinatornoe vydelenie pervogo morfologičeskogo tipa v latyšskom jazyke. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 116–122.

Jakuševa, D. A. 1965. Opyt primenenija algoritma statistiko-kombinatornogo modelirovanija k v'etnamskomu jazyku. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 225–228.

Janßen, Axel. 1992. Segmentierung französischer Wortformen ohne Lexikon. In Ursula Klenk, editor, *Computatio Linguae: Aufsätze zur algorithmischen und quantitativen Analyse der Sprache*, volume 73 of *Zeitschrift für Dialektologie und Linguistik: Beihefte*. Franz Steiner, Stuttgart, pages 74–95.

Jensen, John T. 1990. *Morphology. Word Structure in Generative Grammar*. John Benjamins, Amsterdam.

Johnsen, Lars G. 2005. Morphological learning as principled argument. In Mikko Kurimo, Mathias Creutz, and Krista Lagus, editors, *Proceedings of MorphoChallenge 2005*, pages 33–36, Helsinki University of Technology, Helsinki.

Johnson, Howard and Joel Martin. 2003. Unsupervised learning of morphology for English and Inuktitut. In *HLT-NAACL 2003, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 43–45, Edmonton.

Johnson, Mark. 2008. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, Columbus, OH.

Jordan, Chris, John Healy, and Vlado Keselj. 2005. Swordfish: Using n-grams in an

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

unsupervised approach to morphological analysis. In Mikko Kurimo, Mathias Creutz, and Krista Lagus, editors, *Proceedings of MorphoChallenge 2005*, pages 42–46, Helsinki University of Technology, Helsinki.

Jordan, Chris, John Healy, and Vlado Keselj. 2006. Swordfish: An unsupervised ngram based approach to morphological analysis. In *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 657–658, New York, NY.

Juola, Patrick, Chris Hall, and Adam Boggs. 1994. Corpus-based morphological segmentation by entropy changes. In *Third Conference on the Cognitive Science of Natural Language Processing*, Dublin.

Katrenko, Sophia. 2004. Towards unsupervised learning of morphology applied to Ukrainian. *Student Session: 16th European Summer School in Logic, Language and Information*, pages 138–148, Nancy.

Kazakov, Dimitar. 1997. Unsupervised learning of naïve morphology with genetic algorithms. In *ECML'97 – Workshop Notes on Empirical Learning of Natural Language Tasks*, pages 105–112, Prague.

Kazakov, Dimitar and Suresh Manandhar. 1998. A hybrid approach to word segmentation. In *Proceedings of the 8th International Workshop on Inductive Logic Programming (ILP-98)*, pages 125–134, Madison, WI.

Kazakov, Dimitar and Suresh Manandhar. 2001. Unsupervised learning of word segmentation rules with genetic algorithms and inductive logic programming. *Machine Learning*, 43:121–162.

Keshava, Samarth and Emily Pitler. 2005. A simpler, intuitive approach to morpheme induction. In Mikko Kurimo, Mathias Creutz, and Krista Lagus, editors, *Proceedings of MorphoChallenge 2005*, pages 28–32, Helsinki University of Technology, Helsinki.

Kirik, Harri and Mark Fishel. 2008. Modelling linguistic phenomena with unsupervised morphology for improving statistical machine translation. Paper presented at the Workshop on Unsupervised Methods in NLP, held in conjunction with SLTC'08, Royal Institute of Technology, Stockholm, Sweden.

Klein, Sheldon and Terry A. Dennison. 1976. An interactive program for learning the morphology of natural languages. In *Proceedings of the 3rd International Meeting on Computational Linguistics*, pages 343–353, Debrecen.

Klenk, Ursula. 1985a. Ein nicht-lexikalisches Verfahren zur Erkennung spanischer Wortstämme. In Ursula Klenk, editor, *Strukturen und Verfahren in der maschinellen Sprachverarbeitung*. AQ-Verlag, Dudweiler, pages 47–65.

Klenk, Ursula. 1985b. Recognition of Spanish inflectional endings based on the distribution of characters. In *Computers in Literary and Linguistic Computing: Proceedings of the Eleventh International Conference [L'ordinateur et les recherches littéraires et linguistiques: actes de la XIe Conférence internationale]*, pages 246–253, Louvain.

Klenk, Ursula. 1991. Verfahren der Segmentierung von Wörtern in Morphe: Mit einer Untersuchung zum Spanischen. In Jürgen Rolshoven und Dieter Seelbach, editor, *Romanistische Computerlinguistik: Theorien und Implementationen*, volume 266 of *Linguistische Arbeiten*. Niemeyer, Tübingen, pages 197–206.

Klenk, Ursula. 1992. Verfahren morphologischer Segmentierung und die Wortstruktur des Spanischen. In Ursula Klenk, editor, *Computatio Linguae: Aufsätze zur algorithmischen und quantitativen Analyse der Sprache*, volume 73 of *Zeitschrift für Dialektologie und Linguistik: Beihefte*. Franz Steiner, Stuttgart, pages 110–124.

Klenk, Ursula and Hagen Langer. 1989. Morphological segmentation without a lexicon. *Literary and Linguistic Computing*, 4(4):247–253.

Kohonen, Oskar, Sami Virpioja, and Mikaela Klami. 2008. Allomorfessor: Towards unsupervised morpheme analysis. In Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access*, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarthus, Denmark, September 17–19, 2008, Revised Selected Papers, pages 975–982, Springer-Verlag, Berlin.

Kontorovich, L., D. Don, and Y. Singer. 2003. A Markov model for the acquisition of morphological structure. Technical report CMU-CS-03-147, School of

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Computer Science, Carnegie Mellon University, June.
Kordi, E. E. 1965. Ishodnye dannye dlja statistiko-kombinatornogo modelirovanija morfologii sovremennogo francuzckogo jazyka i vydelenie pervogo morfologičeskogo tipa. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 172–180.
Krauss, Michael. 1992. The world's languages in crisis. *Language*, 68(1):4–10.
Krauss, Michael E. 2007. Mass language extinction and documentation: The race against time. In O. Miyaoka, O. Sakiyama, and M. Krauss, editors, *Vanishing Languages of the Pacific Rim*. Oxford University Press, Oxford, pages 3–24.
Kurimo, Mikko, Mathias Creutz, and Ville Turunen. 2007. Overview of Morpho Challenge in CLEF 2007. In *Working Notes for the CLEF 2007 Workshop*, Budapest.
Kurimo, Mikko, Mathias Creutz, and Matti Varjokallio. 2008a. Morpho challenge evaluation by information retrieval experiments. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, and Anselmo Peñas, editors, *Advances in Multilingual and Multimodal Information Retrieval*, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19–21, 2007, Revised Selected Papers, pages 991–998. Springer-Verlag, Berlin.
Kurimo, Mikko, Mathias Creutz, and Matti Varjokallio. 2008b. Morpho challenge evaluation using a linguistic gold standard. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, and Anselmo Peñas, editors, *Advances in Multilingual and Multimodal Information Retrieval*, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19–21, 2007, Revised Selected Papers, pages 864–872. Springer-Verlag, Berlin.
Kurimo, Mikko, Antti Puurula, Ebru Arisoy, Vesi Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumäe, and Murat Saraçlar. 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 487–494, New York.
Kurimo, Mikko and Ville Turunen. 2008. Unsupervised morpheme analysis evaluation by IR experiments—Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus.
Kurimo, Mikko and Matti Varjokallio. 2008. Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard—Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus.
Langer, Hagen. 1991. *Ein automatisches Morphsegmentierungsverfahren für deutsche Wortformen*. Ph.D. thesis, Georg-August-Universität zu Göttingen.
Lehmann, Hubert. 1973. *Linguistische Modellbildung und Methodologie*. Max Niemeyer Verlag, Tübingen.
Lewis, M. Paul, editor. 2009. *Ethnologue: Languages of the World*. Available at `www.ethnologue.com/`.
Lindén, Krister. 2008. A probabilistic model for guessing base forms of new words by analogy. In *Proceedings of CICLing-2008: 9th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 106–116, Springer, Berlin.
Lindén, Krister. 2009. Entry generation by analogy—encoding new words for morphological lexicons. *Northern European Journal of Language Technology*, 1(1):1–25.
Longacre, Robert E. 1964. *Grammar Discovery Procedures*. Mouton, The Hague.
Mahlow, Cerstin and Michael Piotrowski. 2009. Preface. In *State of the Art in Computational Morphology: Proceedings of the Workshop on Systems and Frameworks for Computational Morphology, SFCM 2009*, pages v–viii, Zurich.
Majumder, Prasenjit, Mandar Mitra, and Dipasree Pal. 2008. Bulgarian, Hungarian and Czech stemming using YASS. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, pages 49–56, Budapest.
Majumder, Prasenjit, Mandar Mitra, Swapan K. Parui, Gobinda Kole, Pabitra Mitra, and Kalyankumar Datta. 2007. YASS: Yet another suffix stripper. *ACM Transactions on Information Systems*, 25(4):18:1–20.
Malahovskij, L. V. 1965. Načal'nyj ėtap statistiko-kombinatornogo modelirovanija morfologii anglijskogo jazyka. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 137–149.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Maxwell, Michael and Anne David. 2008. Joint grammar development by linguists and computer scientists. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 27–34, Hyderabad.

Mayfield, James and Paul McNamee. 2003. Single n-gram stemming. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 415–416, New York, NY.

McClelland, James L. and David E. Rumelhart. 1986. On learning the past tenses of English verbs. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*. MIT Press, Cambridge, MA, pages 216–271.

McNamee, Paul. 2008. Retrieval experiments at Morpho Challenge 2008. In *Working Notes for the CLEF 2008 Workshop*, Aarhus.

McNamee, Paul and James Mayfield. 2007. N-gram morphemes for retrieval. In *Working Notes for the CLEF 2007 Workshop*, Budapest.

Medina Urrea, Alfonso. 2000. Automatic discovery of affixes by means of a corpus: A catalog of Spanish affixes. *Journal of Quantitative Linguistics*, 7(2):97–114.

Medina Urrea, Alfonso. 2003. *Investigación cuantitativa de afijos y clíticos del español de México: Glutinometría en el* Corpus del Español Mexicano Contemporáneo. Ph.D. thesis, El Colegio de México, México, D.F.

Medina-Urrea, Alfonso. 2006a. Affix discovery by means of corpora: Experiments for Spanish, Czech, Ralámuli and Chuj. In Alexander Mehler and Reinhard Köhler, editors, *Aspects of Automatic Text Analysis*, volume 209 of *Studies in Fuzziness and Soft Computing*. Springer, Berlin, pages 277–299.

Medina Urrea, Alfonso. 2006b. Towards the automatic lemmatization of 16th century Mexican Spanish: A stemming scheme for the CHEM. In *Computational Linguistics and Intelligent Text Processing, 7th International Conference, CICLing 2006*, pages 101–104, Mexico City.

Medina-Urrea, Alfonso. 2008. Affix discovery based on entropy and economy measurements. In Nicholas Gaylord, Alexis Palmer, and Elias Ponvert, editors, *Computational Linguistics for Less-Studied Languages*, volume X of *Texas Linguistics Society*. CSLI Publications, Stanford, CA, pages 99–112.

Medina Urrea, Alfonso and E. C. Buenrostro Díaz. 2003. Características cuantitativas de la flexión verbal del Chuj. *Estudios de Lingüística Aplicada*, 38:15–31.

Melkumjan, M. R. 1965. Ishodnye dannye i statistiko-kombinatornoe vydelenie paradigmy pervogo morfologičeskogo tipa v armjanskom jazyke. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 123–136.

Mikheev, Andrei. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23(3):405–423.

Mithun, Marianne. 1999. *The Languages of Native North America*. Cambridge Language Surveys. Cambridge University Press, Cambridge.

Monson, Christian. 2004. A framework for unsupervised natural language morphology induction. In *ACL 2004: Student Research Workshop*, pages 67–72, Barcelona.

Monson, Christian. 2009. *ParaMor: From Paradigm Structure to Natural Language Morphology Induction*. Ph.D. thesis, Carnegie Mellon University.

Monson, Christian, Jaime Carbonell, Alon Lavie, and Lori Levin. 2007a. ParaMor: Finding paradigms across morphology. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, and Anselmo Peñas, editors, *Advances in Multilingual and Multimodal Information Retrieval*, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19–21, 2007, Revised Selected Papers, pages 892–899. Springer-Verlag, Berlin.

Monson, Christian, Jaime Carbonell, Alon Lavie, and Lori Levin. 2007b. ParaMor: Minimally supervised induction of paradigm structure and morphological analysis. In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 117–125, Prague.

Monson, Christian, Jaime Carbonell, Alon Lavie, and Lori Levin. 2008. ParaMor and Morpho Challenge 2008. Using unsupervised paradigm acquisition for prefixes. In Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access*, 9th Workshop of the Cross-Language Evaluation Forum,

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

CLEF 2008, Aarthus, Denmark, September 17–19, 2008, Revised Selected Papers, pages 967–974. Springer-Verlag, Berlin.

Monson, Christian, Alon Lavie, Jaime Carbonell, and Lori Levin. 2004. Unsupervised induction of natural language morphology inflection classes. In *SIGPHON 2004: Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology*, pages 52–61, Barcelona.

Monson, Christian, Alon Lavie, Jaime Carbonell, and Lori Levin. 2008a. Evaluating an agglutinative segmentation model for ParaMor. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 49–58, Columbus, OH.

Monson, Christian, Ariadna Font Llitjós, Vamshi Ambati, Lori Levin, Alon Lavie, Alison Alvarez, Roberto Aranovich, Jaime Carbonell, Robert Frederking, Erik Peterson, and Katharina Probst. 2008b. Linguistic structure and bilingual informants help induce machine translation of lesser-resourced languages. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 2854–2859, Marrakech.

Moon, Taesun, Katrin Erk, and Jason Baldridge. 2009. Unsupervised morphological segmentation and clustering with document boundaries. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 668–677, Singapore.

Naradowsky, Jason and Sharon Goldwater. 2009. Improving morphology induction by learning spelling rules. In *UCAI 2009, Proceedings of the 21st, International Joint Conference on Artificial Intelligence*, Pasadena, California, USA, July 11–17, 2009, pages 1531–1537.

Neuvel, Sylvain and Sean A. Fulop. 2002. Unsupervised learning of morphology without morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, Philadelphia, pages 31–40.

Nida, Eugene A. 1949. *Morphology. The Descriptive Analysis of Words*, 2nd edition. The University of Michigan Press, Ann Arbor, MI.

Nunzio, G. M. Di, N. Ferro, M. Melucci, and N. Orio. 2004. Experiments to evaluate probabilistic models for automatic stemmer generation and query word translation. In *Proceedings of the Cross-Language Evaluation Forum (CLEF): Methodology and Metrics (CLEF 2003)*, pages 220–235.

Oflazer, Kemal, Marjorie McShane, and Sergei Nirenburg. 2001. Bootstrapping morphological analyzers by combining human elicitation and machine learning. *Computational Linguistics*, 27(1):59–85.

Oliver, A. 2004. *Adquisició d'informació lèxica i morfosintàctica a partir de corpus sense anotar: aplicació al rus i al croat*. Ph.D. thesis, Universitat de Barcelona.

Ostler, Nicholas. 2008. Is it globalization that endangers languages? In *UNESCO/UNU Conference: Globalization and Languages: Building our Rich Heritage*, pages 206–211, Paris.

Ožigova, G. I. 1965. Statistiko-kombinatornoe modelirovanie paradigmy pervogo morfologičeskogo tipa v češskom jazyke. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 89–103.

Pandey, Amaresh Kumar and Tanveer J. Siddiqui. 2008. An unsupervised Hindi stemmer with heuristic improvements. In *AND '08: Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*, pages 99–105, New York, NY.

Panina, N. A. 1965. Opyt statistiko-kombinatornogo vydelenija paradigmy pervogo morfologičeskogo tipa v fserbohorvatskom jazyke. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 104–109.

Peršikov, V. F. 1965. Iz opyta statistiko-kombinatornogo modelirovanija albanskoj morfologii. In Nikolaj Dmitrievič Andreev, editor, *Statistiko-kombinatornoe modelirovanie jazykov*. Nauka, Leningrad, pages 181–188.

Petzell, Malin. 2007. *A lingustic description of Kagulu*. Ph.D. thesis, Göteborgs Universitet.

Pirrelli, Vito, Basilio Calderone, Ivan Herreros, and Michele Virgilio. 2004. Non-locality all the way through: Emergent global constraints in the Italian morphological lexicon. In *SIGPHON 2004: Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology*, pages 52–61, Barcelona.

Pirrelli, Vito and Ivan Herreros. 2007. Learning morphology by itself. In *Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5)*, pages 269–290, Fréjus.

Poon, Hoifung, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

morphological segmentation with log-linear models. In *Proceedings of NAACL '09: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 209–217, Morristown, NJ.

Powers, David M. W. 1998. Reconciliation of unsupervised clustering, segmentation and cohesion. In *NeMLaP3/CoNLL '98 Workshop on Paradigms and Grounding in Language Learning*, Sydney, pages 307–310.

Rai, Novel Kishore. 1984. *A Descriptive Study of Bantawa*. Ph.D. thesis, Poona University.

Redlich, A. Norman. 1993. Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, 5(2):289–304.

Reiter, Ehud. 2007. Last words: The shrinking horizons of computational linguistics. *Computational Linguistics*, 33(2):283–287.

Roark, B. and Richard W. Sproat. 2007. Machine learning of morphology. In *Computational Approaches to Morphology and Syntax*, volume 4 of *Oxford Surveys in Syntax and Morphology*. Oxford University Press, Oxford, pages 116–136.

Rodrigues, Paul and Damir Ćavar. 2005. Learning Arabic morphology using information theory. In *The Panels 2005: Proceedings from the Annual Meeting of the Chicago Linguistic Society*, volume 41–2, pages 49–58, Chicago, IL.

Rodrigues, Paul and Damir Ćavar. 2007. Learning Arabic morphology using statistical constraint-satisfaction models. In Elabbas Benmamoun, editor, *Perspectives on Arabic Linguistics: Papers from the Annual Symposium on Arabic Linguistics*, pages 63–75, Urbana, IL.

Rogati, Monica, Scott McCarley, and Yiming Yang. 2003. Unsupervised learning of Arabic stemming using a parallel corpus. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 391–398, Sapporo.

Saxena, Anju and Lars Borin, editors. 2006. *Lesser-Known Languages of South Asia: Status and Policies, Case Studies and Applications of Information Technology*. Mouton de Gruyter, Berlin.

Schone, Patrick. 2001. *Toward Knowledge-Free Induction of Machine-Readable Dictionaries*. Ph.D. thesis, University of Colorado.

Schone, Patrick and Daniel Jurafsky. 2000. Knowledge-free induction of inflectional morphologies using latent semantic analysis. In *Conference on Natural Language Learning 2000 (CoNLL-2000)*, pages 67–72, Lisbon.

Schone, Patrick and Daniel Jurafsky. 2001a. Knowledge-free induction of inflectional morphologies. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 183–191, Pittsburgh, PA.

Schone, Patrick and Daniel Jurafsky. 2001b. Language-independent induction of part of speech class labels using only language universals. In *"Machine Learning: Beyond Supervision," Workshop at IJCAI-2001*, pages 53–60, Seattle, WA.

Sereewattana, Siriwan. 2003. Unsupervised segmentation for statistical machine translation. Master's thesis, University of Edinburgh.

Sharma, Utpal and Rajib Das. 2002. Classification of words based on affix evidence. In *International Conference on Natural Language Processing, ICON-2002*, pages 31–39, Mumbai.

Sharma, Utpal, Jugal Kalita, and Rajib Das. 2002. Unsupervised learning of morphology for building lexicon for a highly inflectional language. In *Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 1–10, Philadelphia.

Sharma, Utpal, Jugal Kalita, and Rajib Das. 2003. Root word stemming by multiple evidence from corpus. In *Proceedings of the 6th International Conference on Computational Intelligence and Natural Computation (CINC)*, pages 1593–1596, Cary, NC.

Snover, Matthew G. 2002. An unsupervised knowledge free algorithm for the learning of morphology in natural languages. Master's thesis, Department of Computer Science, Washington University.

Snover, Matthew G. and Michael R. Brent. 2001. A Bayesian model for morpheme and paradigm identification. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*, pages 482–490.

Snover, Matthew G. and Michael R. Brent. 2003. A probabilistic model for learning concatenative morphology. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*. MIT Press, Cambridge, MA, pages 1513–1520.

Snover, Matthew G., Gaja E. Jarosz, and Michael R. Brent. 2002. Unsupervised learning of morphology using a novel directed search algorithm: Taking the first step. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 11–20, Philadelphia.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Snyder, Benjamin and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT*, pages 737–745, Columbus, OH.

Spencer, Andrew and Arnold M. Zwicky, editors. 1998. *The Handbook of Morphology*. Blackwell, Oxford.

Spiegler, Sebastian, Bruno Golénia, Ksenia Shalonova, Peter Flach, and Roger Tucker. 2008. Learning the morphology of Zulu with different degrees of supervision. In *Spoken Language Technology Workshop, 2008 (SLT 2008)*, pages 9–12, Goa.

Strassel, Stephanie, Mike Maxwell, and Christopher Cieri. 2003. Linguistic resource creation for research and technology development: A recent experiment. *ACM Transactions on Asian Language Processing*, 2(2):101–117.

Tepper, Michael. 2007. Knowledge-lite induction of underlying morphology: A hybrid approach to learning morphemes using context-sensitive rewrite rules. Master's thesis, University of Washington.

Tepper, Michael and Fei Xia. 2008. A hybrid approach to the induction of underlying morphology. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 17–24, Hyderabad.

Theron, Pieter and Ian Cloete. 1997. Automatic acquisition of two-level morphological rules. In *Fifth Conference on Applied Natural Language Processing*, pages 103–110, Washington, DC.

Trosterud, Trond. 2004. Porting morphological analysis and disambiguation to new languages. In *SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages*, pages 90–92, Lisbon.

Trosterud, Trond. 2006. Grammatically based language technology for minority languages. In Anju Saxena and Lars Borin, editors, *Lesser-Known Languages of South Asia: Status and Policies, Case Studies and Applications of Information Technology*. Mouton de Gruyter, Berlin, pages 293–315.

Tufis, Dan. 1989. It would be much easier if WENT were GOED. In *Proceedings of the Fourth Conference of the European Chapter of the ACL*, pages 145–152, Manchester.

ur Rehman, Khalid and Iftikhar Hussain. 2005. Unsupervised morphemes segmentation. In Mikko Kurimo, Mathias Creutz, and Krista Lagus, editors, *Proceedings of MorphoChallenge 2005*, pages 52–56, Helsinki University of Technology, Helsinki.

Virpioja, Sami, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of Machine Translation Summit XI*, pages 391–498, Copenhagen.

Wicentowski, Richard. 2002. *Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework*. Ph.D. thesis, Johns Hopkins University, Baltimore, MD.

Wicentowski, Richard. 2004. Multilingual noise-robust supervised morphological analysis using the wordframe model. In *Proceedings of the ACL Special Interest Group on Computational Phonology (SIGPHON)*, pages 70–77, Barcelona.

Wintner, Shuly. 2009. Last words: What science underlies natural language engineering? *Computational Linguistics*, 35(4):641–644.

Wothke, Klaus. 1986. Machine learning of morphological rules by generalization and analogy. In *Proceedings of the 11th Conference on Computational Linguistics*, pages 289–293, Morristown, NJ.

Wothke, Klaus Christian. 1985. *Maschinelle Erlernung und Simulation morphologischer Ableitungsregeln*. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität zu Bonn.

Xanthos, Aris. 2007. *Apprentissage automatique de la morphologie: Le cas des structures racine-schème*. Ph.D. thesis, Université de Lausanne.

Xanthos, Aris, Yu Hu, and John Goldsmith. 2006. Exploring variant definitions of pointer length in MDL. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL 2006*, pages 32–40.

Yarowsky, David, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, Stroudsburg, PA, pages 1–8.

Yarowsky, David and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pages 207–216, Hong Kong.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024

Yvon, François. 1996. *Prononcer par analogie: motivation, formalisation et evaluation*. Ph.D. thesis, École Nationale Supérieure des Télécommunications, Paris.

Zeman, Daniel. 2008. Unsupervised acquiring of morphological paradigms from tokenized text. In *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, pages 892–899, Budapest.

Zeman, Daniel. 2009. Using unsupervised paradigm acquisition for prefixes. In Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access*, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarthus, Denmark, September 17–19, 2008, Revised Selected Papers, pages 983–990. Springer-Verlag, Berlin.

Zhang, Byoung-Tak and Yung-Taek Kim. 1990. Morphological analysis and synthesis by automated discovery and acquisition of linguistic rules. In *Papers Presented to the 13th International Conference on Computational Linguistics (COLING 1990)*, volume 2, pages 431–436, Helsinki.

Zweigenbaum, P., F. Hadouche, and N. Grabar. 2003. Apprentissage de relations morphologiques en corpus. In *Actes de TALN 2003*, pages 285–294, Batz-sur-mer.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00050 by guest on 29 March 2024