# Resource-light Approaches to Computational Morphology
# Part 1: Monolingual Approaches

Jirka Hana and Anna Feldman

**Abstract**

This article surveys *resource-light monolingual* approaches to morphological analysis and tagging. While supervised analyzers and taggers are very accurate, they are extremely expensive to create. Therefore, most of the world languages and dialects have no realistic prospect for morphological tools created in this way. The weakly-supervised approaches aim to minimize time, expertise and/or financial cost needed for their development. We discuss the algorithms and their performance considering issues such as accuracy, portability, development time and granularity of the output.

## 1 Introduction

*Morphological analysis* (MA) is the process of labeling a word with tags encoding its morphological properties. For example, the word *rose* in English could be analyzed as a noun or as a verb. Morphological analysis considers words out of context; *morphological tagging*, on the other hand, assigns each word a single tag based on the context the word is in. Therefore, *rose* would be tagged as a noun in *The rose smells nicely.* and as a verb in *He rose from his seat.* The granularity of tagsets varies. Some tagsets encode part-of-speech only, while some add additional grammatical information such as case, number, gender, tense, etc. Depending on the language and captured distinctions, a tagset usually contains between several tens to several thousands tags (e.g., about 40 in the English Penn Treebank tagset (Marcus et al., 1993), about 4,000 in the Czech Positional tagset (Hajič, 2004)). Morphological analysis and tagging may be accompanied by *lemmatization*, a procedure assigning each word its lemma.[1] For example, *rose* would be assigned the lemma *rose* (for the noun) or *rise* (for the verb).

[1] By *lemma* we mean a form distinguished from a set of all forms related by inflection. Lemmas are chosen by convention (e.g., nominative singular for nouns, infinitive for verbs). We consider the terms *base*, *canonical* or *citation form* to be synonymous. Some NLP articles (e.g., Yarowsky and Wicentowski, 2000), use the word *root* or *stem* to mean the same thing. We use these terms with their traditional linguistic meaning: *root* refers to a morpheme which is not an affix, while *stem* is a word without its inflectional affixes. For many English words, all these terms refer to the same string.

Morphological analysis, tagging and lemmatization are essential for many Natural Language Processing (NLP) applications, of both practical and theoretical nature. They are commonly used in syntactic parsers, grammar checkers, speech recognition systems, web searches, machine translation, text-to-speech synthesis, etc.

Modern taggers and analyzers are very accurate. However, the standard way to create them for a particular language requires a substantial amount of expertise, time and money. For example, the Czech analyzer developed by Hajič (2004) uses a lexicon with 300,000+ manually entered entries.[2] The creation of manually annotated corpora used for tagger training is also an extremely expensive undertaking. As a result, most of the world languages and dialects have no realistic prospect for morphological taggers or analyzers created this way.

Various techniques have been suggested to overcome this problem, including unsupervised methods. While completely unsupervised systems are scientifically interesting, shedding light on areas such as child language acquisition or general learnability, for many practical applications their precision is still too low. They also completely ignore linguistic knowledge accumulated over several millennia, often rediscovering rules that can be found in basic grammar books.

Another strand of research to overcome the lack of morphologically annotated data uses lightly supervised methods, such as bootstrapping from a small lexicon or labeled corpus, from manually encoded phonological or morphological rules or paradigms, or even from resources from a different (but perhaps related) language. We call these methods *resource-light*. Interestingly, Zipf's law (1935; 1949) provides a convincing argument for resource-light systems against both fully-unsupervised and fully-supervised systems. Not all manually provided resources have the same impact on the accuracy of the system. A small amount of high-impact resources can lead to an acceptable accuracy. At the same time, no manually compiled lexicon can cover an arbitrary text. Feldman and Hana (2010) document this on a statistic of Czech nouns obtained from the PDT corpus (Hajič et al., 2000). On one hand, 10% of the most frequent noun lemmas cover 74% of the noun tokens, and on the other, 50% of the less frequent lemmas cover only 4% of tokens; moreover they are very text specific – 70% of them do not occur in another portion of the corpus.

This survey focuses on approaches that require some kind of light supervision, such as seeding (e.g., Yarowsky and Wicentowski, 2000; Kohonen et al., 2010), manual correction (e.g., Bosch et al., 2008; Oflazer et al., 2001), and manual encoding of basic linguistic facts (e.g., Cucerzan and Yarowsky, 2002; Feldman and Hana, 2010; Tepper and Xia, 2010). Learning from a different language (e.g., Cucerzan and Yarowsky, 2002; Feldman and Hana, 2010; Bosch

[2] In this paper, we use the word lexicon to mean a list of stems or lemmas each with information about its paradigm.

et al., 2008), another resource-light strategy, will be discussed in our forthcoming survey (Feldman and Hana, 2012).

Resource-lightness falls on a continuum from a relatively high level of supervision to entirely unsupervised. Cucerzan and Yarowsky (2002), for instance, observe that one useful measure of minimal supervision is the additional cost of obtaining a desired functionality from existing commonly available knowledge sources. They note that for a remarkably wide range of languages, there exist plenty of reference grammar books and dictionaries which are invaluable linguistic resources. Clearly, different types of resources require a different amount of time and/or expertise. Cucerzan and Yarowsky (2002) need one day plus a reference grammar,[3] while Hana et al. (2011) need one week, a reference grammar and a resource for a related language. In turn, many so-called unsupervised taggers (e.g., Merialdo, 1994; Cutting et al., 1992; Brill, 1995; Banko and Moore, 2004; Wang and Schuurmans, 2005) need no annotation but require a list of all possible tags for each form which is generally equivalent to requiring a morphological analyzer.

In the following, we discuss various problems and methods of resource-light morphological induction in more detail. Table 4 summarizes the main points.

# 2 Morphological analysis, morphemes and inflections

In this section we discuss resource-light systems developed for processing morphology in various languages. In addition to resource-light systems, we briefly mention several prominent resource-intensive systems on one hand, and some unsupervised approaches on the other. These approaches are *not* the focus of this survey. The reason why we outline them here in a nutshell is that resource-light systems often incorporate ideas from both unsupervised approaches and resource intensive systems.

## 2.1 Approaches to morphological analysis

Simplifying somewhat, there are two major approaches to morphological analysis. Both approaches append strings to model the concatenative aspect of morphology, but they differ in how they handle phonological changes[4]:

1. two level morphology (2LM; Koskenniemi, 1983; Karttunen and Beesley, 1992; Beesley and Karttunen, 2003). In mainstream linguistics, phonology is handled via ordered rewrite rules. These rules can theoretically produce

[3] By reference grammars we mean grammars that usually focus on the fundamental grammar structures normally taught in basic or introductory courses. As the approaches rely on simple paradigm/ending tables (except Tepper and Xia (2008, 2010)), the quality and depth of the grammar does not make much difference.

[4] We use the term *phonological changes* to refer to all phonological, graphemic and allomorphic changes. See also section 2.3.

an unbounded number of intermediate forms before transforming an abstract sequence of morphemes into a corresponding surface form. While 2LM employs similar rules, they relate only two levels (lexical and surface level) and are all applied in parallel. A network of lexicons specifies the lexical forms of morphemes and the basics of their morphotactics (i.e., legal orderings and combinations). In its current version, 2LM is usually realized as a finite state transducer. The formalism is powerful enough to model the majority of morphologies, even though modeling some phenomena is less than straightforward.

2. so-called engineering approach. Such systems do not have a phonological component at all, or the component is rudimentary. Instead, phonological changes and irregularities are factored into endings and a higher number of paradigms. This implies that the terms *stem* and *ending* have slightly different meanings from the ones they traditionally have. A stem is the part of the word that does not change within its paradigm, and the ending is the part of the word that follows such a stem. Examples of such an approach are Mikheev and Liubushkina (1995) for Russian and Hajič (2004) for Czech. The advantages of such a system are its high speed, simple implementation and straightforward morphology specification. The problems are a very high number of paradigms (several hundreds in the case of Czech) and the impossibility to capture even the simplest and most regular phonological changes and so its limited ability to predict the behavior of new lexemes. For example, the English noun paradigm (*0 – s*) would be captured as multiple paradigms including, *0 – s*, *0 – es*, *y – ies*, *f – ves*.

In addition, for many languages, morphology can be modeled and implemented as a simple look-up table associating inflected forms with their analyses.

## 2.2 Learning of morphemes and/or paradigms

As stated above, traditional morphological analyzers rely on large lexicons specifying inflections of individual lemmas. These lexicons take years to develop and have to be constantly updated. Therefore, significant research has been done into ways of obtaining this information automatically. Trivially, unsupervised systems are resource-light. However, we focus on approaches that use at least some human supervision and therefore, here we mention only the most prominent unsupervised and weakly supervised approaches that are relevant in the subsequent discussion. See Hammarström and Borin (2010) for a detailed survey of unsupervised approaches.

*Linguistica* (Goldsmith, 2001, 2009) is one of the most cited systems for unsupervised morphological acquisition. From a plain text corpus, it learns derivational and inflectional paradigm approximations (called signatures) together with a lexicon. It uses several heuristics to find candidate segmentations of words into morphemes and then uses minimum description length (MDL Rissanen, 1989; Kazakov, 1997; de Marcken, 1995) to choose between them. When

Table 1: Results of (Kohonen et al., 2010) with various size of training data

| | Kohonen et al. (2010) | | | | | | Morfessor | soa |
|---|---|---|---|---|---|---|---|---|
| labeled data size | 500 | 600 | 800 | 1.5K | 3.5K | 10.5K | 0 | 0 |
| English | 61.1 | 65.2 | 65.6 | 68.3 | 69.1 | 72.9 | 59.8 | 66.2 |
| Finnish | 49.1 | 52.7 | 54.9 | 56.4 | 58.2 | 60.3 | 44.6 | 52.5 |

All numbers are F-measures obtained on the Morpho Challenge 2009 test data (Kurimo et al., 2010). The size of heldout data (500 words) is included in the size of training data.
For comparison, the Morfessor column reports the results of the original unsupervised Morfessor system in the best configuration for the language and soa reports the results of the best unsupervised tool at the Challenge (Bernhard, 2008).

comparing two grammars describing a corpus, MDL chooses the one which compresses the corpus the most (the size of the grammar is included in the comparison). The system has been successfully applied to a range of languages.

Unlike Linguistica, *Morfessor* (Creutz and Lagus, 2002, 2004, 2005) splits words into morphemes in a hierarchical fashion. This makes it more suitable for agglutinative languages, such as Finnish or Turkish, with a large number of morphemes per word. They use a Hidden Markov Model (HMM) to add a simple morphotactic model.

Kohonen et al. (2010) modify Morfessor to allow semi-supervised learning. In addition to a plain word list, the algorithm is given a set of 0-10,000 correctly segmented words. Also, a heldout of 500 correctly segmented words is used to optimize separate weights for unlabeled and labeled data, to prevent the small amount of labeled words (i.e., segmented) being overwhelmed by the large amount of unlabeled data. The results, summarized in Table 1, show that to surpass the unsupervised state of the art (Bernhard, 2008), one needs a relatively small list of segmented words (1000+500 for English and 100+500 for Finnish).

*Paramor* (Monson, 2009) is a system for unsupervised acquisition of paradigms from a list of words. It learns paradigms and a lexicon in several steps. It first considers all possible segmentations of words into candidate stems and endings. Then it creates schemes (partial paradigms with the associated stems) by joining endings that share a large number of associated stems. In the next step, similar schemes (as measured by cosine similarity) are merged. Finally, schemes proposing frequent morpheme boundaries not consistent with boundaries proposed by the character entropy measure are discarded.

All of these models (Linguistica, Morfessor, Paramor) are strictly concatenative and they are not suitable for discovering paradigms employing other morphological processes (interfixes, templates, metathesis, deletion, etc.).

## 2.3 Allomorphy and irregularity

Many morphemes have several contextually dependent realizations, the so-called allomorphs, due to phonological/graphemic changes or irregularities.[5] There are

[5] Here, we exclude morpheme variance due to different paradigmatic classes. Therefore, we consider *leaf*/*leav* (as in *leaves*), *happy*/*happi* (as in *happier*), and plural *-s*/*-es* to be

various approaches to allomorphy in unsupervised and semi-supervised methods. Actually, probably the most common approach, is to simply ignore it completely, as, for example, do all the systems discussed in the previous section.

There are at least two reasons to handle allomorphy. First, linguistically, it makes more sense to analyze *winning* as *win+ing* than as *winn+ing* or *win+ning*, and *happier* as *happy+er* than as *happ+ier*. For many applications, such as information retrieval, it is helpful to know that two morphs are variants of the same morpheme. Second, ignoring allomorphy makes the data appear more complicated and noisier than they actually are. Thus, the process of learning morpheme boundaries or paradigms is harder and less successful.

Since many allomorphs of the same morpheme have a similar form, some approaches (e.g., Yarowsky and Wicentowski, 2000; Oflazer et al., 2001; Cucerzan and Yarowsky, 2002) use Levenshtein edit distance (Levenshtein, 1966) to link them. Yarowsky and Wicentowski (2000) use edit distance to account for allomorphy of stems. Rather than treating all string edits as equal, they use weighted edit distance, i.e., their costs might be different for each pair of characters. Moreover, consonant clusters and vowel clusters might be treated as a single unit. The initial costs is set to prefer mutation of vowels over mutations of consonants[6] or costs from a similar language might be used. The values are iteratively re-estimated.

Cucerzan and Yarowsky (2002) compile concatenative aspects of inflectional paradigms (i.e., they list endings with their tags) on the basis of a grammar book, ignoring all allomorphic variations. In a sense, this is similar to using two-level morphology without any two-level rules. Obviously, many forms hypothesized on the basis of such information are inaccurate. They are later adjusted by matching them against the actual forms found in a corpus using weighted edit distance.

Such approaches treat allomorphic changes in all places of words as equally likely. However, most changes occur at morpheme boundaries. Wicentowski's (2002; 2004) WordFrame model uses a template to restrict the possible places of changes to the point of affixation[7] and the root vowel(s) (e.g., *foot* – *feet*, *Vater* 'father' – *Väter* 'fathers' in German). Cheng and See (2006) extend the model to handle infixation.

The methods above provide the information about default suffixes manually and the algorithm learns the allomorphic variance. Tepper and Xia (2008, 2010)

allomorphs. However, we exclude *-est* and *most* even though they both mark superlative, and Czech *-o*, *-í*, *-a*, *-ě*, *0*, ... even though they all mark nominative (*jmén-o* 'name', *staven-í* 'building', *žen-a* 'woman', *jeskyn-ě* 'cave', *pán* 'Mister' ).

[6] The authors claim that this is motivated by the fact that in morphological systems worldwide, vowels and vowel clusters tend to change more often during inflection than consonants. However, they do not provide any reference for this claim and we were not able to find any linguistic support for it.

[7] Affixation may involve allomorphic changes (often phonologically conditioned) at the location where the affix is added, so-called points-of-affixation changes, e.g., palatalization in Czech (*matk-a* 'mother' – *matč-in* 'mother's'). Sometimes these changes are realized only in spelling, e.g., gemination or elision in English such as *stir/stirred* and *close/closing*, respectively.

use the opposite approach. They specify the phonological/graphemic rules manually to improve unsupervised morpheme segmentation by Morfessor, an unsupervised system (see Section 2.2). They have tested their system on English and Turkish. They got significant improvements over the results of Morfessor: F-measure rose from 47% to 60% for English and from 37% to 55% for Turkish.[8] The authors needed a day to write rules for English (6 rules in total) and about a week for Turkish (10 in total). The rules are based on comprehensive grammars of English and Turkish: Huddleston and Pullum (2001) and Göksel and Kerslake (2005), respectively. While Tepper and Xia (2008, 2010)'s approach avoids the need of extensive manually encoded resources, they rely on detailed, sophisticated grammars for each language. It would be interesting to see how their system would perform using imperfect rules derived in a fraction of time. We hypothesize that these rules would be still useful, but unfortunately, the authors did not test this possibility.

## 2.4 Combination of evidence

All the systems above rely on word forms and word frequencies only, ignoring syntax and semantics. In fact, they work on frequency annotated word types, not on word tokens with an actual context. It has been shown by many researchers (e.g., Yarowsky and Wicentowski, 2000; Schone and Jurafsky, 2000; Baroni et al., 2002) that combining multiple sources of information gives better results than any single of them.

For example, Yarowsky and Wicentowski (2000) present an algorithm for a resource-light induction of present-past verb pairs (with suffixal and irregular morphology) from a large unannotated corpus by iterative combination of four alignment measures:

1. Alignment by frequency similarity assumes that two forms are forms of the same verb when their relative frequency fits the expected distribution. The distribution of irregular forms is approximated by the distribution of regular forms. Despite large lemma frequency differences between regular and irregular English verbs, the distributions of relative tense ratios for both past-tense-form/present-form and gerund/present-form are similar, e.g., the average past/present ratio for regular verbs is 0.847 and for irregular verbs is 0.842. Thus from the point of view of this metric, it is more likely that the past tense of *sing* is *sang* than *singed*, because the relative frequency of *sang/sing* is 1.19 while the relative frequency of *singed/sing* is 0.0007.

2. Alignment by context similarity relies on the assumption that forms of the same verb occur in the same context. This is true because subcategorization and selection restrictions usually do not change for most verbal

[8] Unfortunatelly, the system was tested on the Morpho Challenge 2007 data (`http://research.ics.tkk.fi/events/morphochallenge2007/`) while Kohonen et al. (2010) (see Section 2.2) tested their system on the Morpho Challenge 2009 data, so the results might not be directly comparable.

inflections (possibly with the exception of aspect and gerunds in some languages). Thus, for example, the English verb forms *pull*, *pulled* or *pulling* are all transitive and require a noun phrase object, such as *her hair*. To minimize the required human supervision, the authors identify the positions of subjects and objects using a set of simple regular expressions. As a result, many legitimate contexts are not matched. Nevertheless, the partial coverage is tolerable because these expressions are applied to a large corpus.

3. Alignment by weighted Levenshtein distance accounts for allomorphy of stems (see §2.3).

4. Alignment by a probabilistic function mapping lemmas to inflections, which is trained on the previous iteration of this algorithm.

Of the four measures, no single model is sufficiently effective on its own. Therefore, traditional classifier combination techniques are applied to merge scores of the four models.

They assume the following input: 1) a list of inflectional categories, each with canonical suffixes; 2) a large unannotated text corpus; 3) a list of candidate noun, verb, and adjective base forms (typically obtainable from a dictionary); 4) a rough mechanism for identifying the candidate parts of speech of the remaining vocabulary, not based on morphological analysis; 5) a list of consonants and vowels; 6) optionally, a list of common function words; 7) optionally, parameters of the model generated on previously studied languages to be used as seed information, especially if these languages are closely related.

There are some problems though. The algorithm is successfully tested on induction of English present-past verb pairs, but the paper uses them just as an example claiming it can be used to induce general morphological analyzers. In that case, however, two of the alignment measures would have to be significantly modified or probably replaced.

- First, the frequency alignment measure works well for verbal tense, but it would have to be modified to handle categories where one can expect multimodal distribution. For example, consider the number of nouns: the distribution is different for count nouns, mass nouns, plurale-tantum nouns, etc.

- Second, the context similarity measure relies on the assumption that inflected forms occur in a similar context. This is true for verbs, because subcategorization requirements usually do not change for most verbal inflections (possibly with the exception of aspect and gerunds in some languages). However, it is definitely not true for nominal categories: on one hand, they have very weak selectional requirements, and on the other hand, they are usually surrounded by agreeing attributes changing their inflection in sync with them.

## 2.5 Providing paradigms, learning lexicons

Using a basic reference grammar, it is relatively easy to provide information about inflectional endings, possibly organized into paradigms. In some languages, an analyzer built on such information would have an acceptable accuracy (e.g., in English most words ending in *ed* are past/passive verbs, and most word ending in *est* are superlative adjectives). However, in many languages, the number of homonymous endings is simply too high for such system to be useful. For example, the ending *a* has about 19 different meanings in Czech (Feldman and Hana, 2010) as is illustrated in Table 2.

Table 2: Homonymy of the *a* ending in Czech (from Feldman and Hana (2010))

| form | lemma | gloss | category |
|---|---|---|---|
| měst-a | město | town | noun neuter sg gen |
| | | | noun neuter pl nom (voc) |
| | | | noun neuter pl acc |
| tém-a | téma | theme | noun neuter sg nom (voc) |
| | | | noun neuter sg acc |
| žen-a | žena | woman | noun feminine sg nom |
| pán-a | pán | man | noun masculine anim sg gen |
| | | | noun masculine anim sg acc |
| ostrov-a | ostrov | island | noun masculine inanim sg gen |
| předsed-a | předseda | president | noun masculine anim sg nom |
| vidě-l-a | vidět | see | verb past feminine sg |
| | | | verb past neuter pl |
| vidě-n-a | | | verb passive feminine sg |
| | | | verb passive neuter pl |
| vid-a | | | verb transgressive masculine sg |
| dv-a | dva | two | numeral masculine sg nom |
| | | | numeral masculine sg acc |

Therefore, several researchers have used plain text corpora to automatically acquire a lexicon. For example, Hana et al. (2004); Hana (2008); Feldman and Hana (2010) build a system which relies on the inclusion of a limited amount of high-impact and low-cost manual resources while other resources are acquired automatically. They provide manual analyses of the most frequent words[9] and endings organized into paradigms including information about simple point-of-affixation stem changes. They call these resources “high-impact and low-cost”. The lexicon is automatically acquired. For each form, all hypothetical lexical entries consistent with the information about the paradigms are created. Then competing entries are compared and only those supported by the highest number of forms-tokens and/or form-types are retained. Most of the remaining entries are still non-existent; however, in the majority of cases they licence the same inflections as the correct entries, differing only in rare inflections. Table 3

[9] They report experiments that use 0K, 5K and 10K most frequent forms.

Table 3: Results of (Feldman and Hana, 2010)

| Language | Czech (nouns) | | | | Russian | Portuguese | Catalan |
|---|---|---|---|---|---|---|---|
| | | | | soa | | | |
| manual lexicon | 0 | 0 | 0 | 300K | 0 | 0 | 0 |
| manual word list | 0 | 5K | 10K | 0 | 1K | 0 | 1K |
| manual paradigms | + | + | + | + | + | + | + |
| manual derivations | 20 | 20 | 20 | ? | 0 | 0 | 0 |
| tagset size | | | | | 1063 | 259 | 289 |
| recall | 94.2 | 96.1 | 96.6 | 98.7 | 93.4 | 98.0 | 95.8 |
| ambiguity (tag/word) | 11.7 | 8.5 | 4.0 | 3.8 | 2.8 | 3.4 | 2.6 |

A lexicon entry specifies all inflectional forms of a lemma (by referring to its paradigm).
In case of Czech, the system was evaluated on PDT 1.0 (`http://ufal.mff.cuni.cz/pdt/`) and compared with the supervised state-of-the-art (soa; Hajič, 2004). Words of all POS categories were included in the input, but the tools were evaluated on nouns only. The authors restrict the evaluation on nouns as they are the most open class and the hardest to cover completely by supervised tools.

To be considered correct, the Czech analysis must contain a noun tag specifying the correct gender (4 usual values), number (2), case (7) and negation (2). In case of the other languages, the system was evaluated on all POS's.

summarizes the results of their system.

Similarly, Gasser (2010) uses a web-corpus and a MA guesser to expand a seed lexicon based on an online dictionary with 598 verb roots.

## 2.6 Computer-assisted creation of morphological analyzer

Out of all the approaches, the one by Oflazer et al. (2001) (similar to Mikheev and Liubushkina (1995)) is probably the closest to resource-intensive methods. They bootstrap a morphological analyzer relying on direct human supervision to produce two-level rules which are then compiled into a finite state transducer. The basic ideas can be summarized as follows:

1. Paradigm definitions are provided manually. Each paradigm is specified via an example – by listing all forms of a single word together with their tags. Additional examples can be added to provide information about regular and irregular allomorphy. Finally, the specification can contain a list of lemmas inflecting according to the paradigm.

2. The examples are automatically segmented into stems and endings. Stems are strings minimizing the sum of edit distances to all forms.

3. Transformation-based learning (Brill, 1995) is used to learn phonological changes accompanying concatenation of a stem with an ending.

4. A morphological analyzer is created by compiling the whole specification (stems with endings and transformations) into a finite state transducer.

5. The analyzer is tested on a corpus and all unanalyzed but "nearly analyzed" forms are reported. A human adjusts the specification and the process is repeated.

The obvious disadvantage of the method is the amount of supervision needed, which is much higher than that of any other system. However, as the system derives both morpheme boundaries and phonological rules automatically, the complexity of information and the level of expertise needed is significantly lower than that of fully supervised systems described in §2.1, while the results are probably comparable (we are not aware of any back-to-back comparison). Note also that the adjustment in the last step is done only on the basis of unanalyzed forms while some of the analyses provided by the system are not necessarily correct either.

# 3 Tagging

Supervised taggers are usually trained on large corpora (100,000+ words) that have been annotated by hand. In addition or instead, some taggers employ hand-written disambiguation rules (e.g., Hajič et al., 2001). For languages with many forms for each word, taggers often operate on the result of a morphological analyzer to alleviate data sparsity problem (Hajič, 2000). Since both annotated corpora and disambiguation rules are costly and tedious to produce, various less supervised alternatives have been explored.

Note, however, that many of the taggers commonly called unsupervised (e.g., Merialdo, 1994; Cutting et al., 1992; Brill, 1995; Banko and Moore, 2004; Wang and Schuurmans, 2005) are not entirely knowledge-free because they require a list of possible tags for each form. Moreover, as Banko and Moore (2004) pointed out these lists are usually obtained by collecting analyses from an annotated corpus. This means they do not contain improbable analyses which makes the task easier than disambiguating the output of a regular analyzer.

Smith and Eisner (2005) present a more resource light approach – they use a tag dictionary only for frequent words, while infrequent words can be assigned any tag. Similarly, Haghighi and Klein (2006) provide only several examples for each tag. Goldwater and Griffiths (2007) present a series of experiments with an unsupervised Bayesian tagger changing the size of the tag dictionary from a full dictionary compiled from a corpus to no dictionary at all (reducing the task to word clustering).

# 4 Conclusion

We have reviewed and compared resource-light methods for processing morphology, including segmentation, analysis, tagging, lexicon induction etc. Table 4 summarizes selected resource-light approaches discussed in this survey. The methods differ in the amount of supervision required, their accuracy, development time, and their portability. Naturally, unsupervised systems are the most portable and are the fastest to develop. However, they often lack precision and granularity of linguistic analysis to be useful for practical applications. At the same time, the development of systems that rely on manually encoded linguis-

| paper/system | languages | what | resources (+ raw corpus in all cases) | tagset | devel. time | comment |
|---|---|---|---|---|---|---|
| Kohonen et al. (2010) | Finnish, English | segmentation | 100+/1000+ (fi/en) segmented words | – | not specified | |
| Tepper and Xia (2008, 2010) | English, Turkish | segmentation | manual allomorphy rules; | – | en: 1 day ;<br>tr: 1 week | requires nontrivial linguistic knowledge |
| Oflazer et al. (2001) | Polish | segmentation and allomorphic rules | 1) manual paradigm definitions; 2) manually segmented examples | – | not specified | the highest amount of supervision required |
| Yarowsky and Wicentowski (2000) | English, Spanish | clustering of forms by inflection | 1) a list of POS and their associated canonical suffixes; 2) list of common nouns, verbs, adj's; 3) some rough mechanism (symbolic rules) for identifying POSs of the remaining vocabulary; 4) a list of consonants and vowels for each language<br>optional: 1) parameters from other languages; 2) common function words | – | not specified, but reasonably fast | not clear how to generalize to other POS |
| Hana et al. (2004); Hana (2008); Feldman and Hana (2010) | Russian, Czech, Catalan, Portuguese, Old Czech | morphological analysis | basic reference grammar;<br>optional: analyses of frequent forms (0-10k); list of frequent derivational suffixes | 250-2000 tags | 20-40 h/lg | |
| Cucerzan and Yarowsky (2003) | French, Romanian, Spanish; Slovene; Swedish | gender acquisition | dictionary to create a seed of <50 gender-annotated words | – | minimal | seed can be also projected from another language |
| Gasser (2010) | Tigrinya | lexicon acquisition | 1) reference grammar; 2) basic MA with a guesser; 3) list of 600 roots | – | not specified | |
| Cucerzan and Yarowsky (2002) | Kurdish, Romanian, Spanish | POS tagger | 1) small bilingual dictionary; 2) reference grammar | 250 tags | 4 h/lg | |

Table 4: Comparison of select resource-light monolingual approaches

tic data, be it phonological or morphological rules/paradigms or just seed lists, is clearly more time consuming. Such systems are less portable because for each new language, a certain amount of linguistic knowledge is required to be encoded by hand. Nevertheless, many systems that we reviewed here do not assume the time consuming linguistic encoding, but rely only on basic facts about a particular language. Note that Table 4 does not provide performance scores. The reason is that these systems are not directly comparable: some deal with morphological segmentation, others with detailed morphological analysis or lexicon acquisition, while others with POS tagging. In addition, these approaches were tested on different languages and various sizes of corpora. Thus, the performance numbers are meaningless out of context. Basic discussion of performance can be found in previous sections; for full details, we refer the interested reader to the original publications.

Our forthcoming survey (Feldman and Hana, 2012) deals with another strand of research on resource-light morphology, namely, with cross-lingual approaches. We will discuss various strategies to overcome the lack of training data by using resources available for another language or dialect, be it an annotated corpus, a grammar or a dictionary. Some approaches we discuss rely on parallel aligned or non-aligned corpora, other transfer necessary information from the source language to the target language only assuming access to bilingual non-parallel corpora.

# Acknowledgments

This research was supported by the Grant Agency of the Czech Republic (project ID: P406/10/P328) and by the U.S. NSF grants #0916280, #1033275. We would also like to thank the anonymous reviewers for their comments.

# References

Banko, M. and R. C. Moore (2004). Part-of-speech tagging in context. In *Proceedings of Coling 2004*, Geneva, Switzerland, pp. 556–561. COLING.

Baroni, M., J. Matiasek, and H. 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*, pp. 48–57.

Beesley, K. R. and L. Karttunen (2003). *Finite State Morphology.* CSLI Studies in Computational Linguistics. CSLI.

Bernhard, D. (2008). *Simple Morpheme Labelling in Unsupervised Morpheme Analysis*, pp. 873–880. Berlin, Heidelberg: Springer-Verlag.

Bosch, S., L. Pretorius, K. Podile, and A. Fleisch (2008). Experimental fast-tracking of morphological analysers for nguni languages. In *Proceedings of* *the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics 21*(4), 543–565.

Cheng, C. K. and S. L. See (2006). The revised wordframe model for the filipino language. *Journal of Research in Science, Computing and Engineering 3*(2), 17–23.

Creutz, M. and K. Lagus (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning – Volume 6*, MPL '02, Stroudsburg, PA, USA, pp. 21–30. Association for Computational Linguistics.

Creutz, M. and K. Lagus (2004). Induction of a simple morphology for highly-inflecting languages. In *Proceedings from the 7th regional meeting of the CLA special interest group in computational phonology (SIGPHON)*, pp. 43–51.

Creutz, M. and K. Lagus (2005). 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)*, pp. 106–113. Finland: Espoo.

Cucerzan, S. and D. Yarowsky (2002). Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings of the 6th conference on Natural language learning – Volume 20*, COLING-02, Stroudsburg, PA, USA, pp. 1–7. Association for Computational Linguistics.

Cucerzan, S. and D. Yarowsky (2003). Minimally supervised induction of grammatical gender. In *Proceedings of HLT-NAACL 2003: Main Conference*, pp. 40–47.

Cutting, D., J. Kupiec, J. Pedersen, and P. Sibun (1992). A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pp. 133–140.

de Marcken, C. (1995). *Unsupervised Language Acquisition*. Ph. D. thesis, MIT, Cambridge, MA.

Feldman, A. and J. Hana (2010). *A resource-light approach to morpho-syntactic tagging*. Amsterdam/New York, NY: Rodopi.

Feldman, A. and J. Hana (2012). Resource-light approaches to computational morphology, part 2: Cross-lingual approaches. *Compass*. Forthcoming.

Gasser, M. (2010). Expanding the lexicon for a resource-poor language using a morphological analyzer and a web crawler. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias

(Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA): European Language Resources Association (ELRA).

Göksel, A. and C. Kerslake (2005). *Turkish: A Comprehensive Grammar*. Routlege:London.

Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics 27*(2), 153–198.

Goldsmith, J. (2009). Morphological analogy: Only a beginning. In J. P. Blevins and J. Blevins (Eds.), *Analogy in Grammar — Form and Acquisition*. Oxford University Press.

Goldwater, S. and T. Griffiths (2007, June). A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, pp. 744–751. Association for Computational Linguistics.

Haghighi, A. and D. Klein (2006). Prototype-driven learning for sequence models. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, Stroudsburg, PA, USA, pp. 320–327. Association for Computational Linguistics.

Hajič, J., A. Böhmová, E. Hajičová, and B. Vidová-Hladká (2000). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In A. Abeillé (Ed.), *Treebanks: Building and Using Parsed Corpora*, pp. 103–127. Amsterdam:Kluwer.

Hajič, J. (2000). Morphological Tagging: Data vs. Dictionaries. In *Proceedings of ANLP-NAACL Conference*, Seattle, Washington, USA, pp. 94–101.

Hajič, J. (2004). *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Praha: Karolinum, Charles University Press.

Hajič, J., P. Krbec, P. Kveton, K. Oliva, and V. Petkevic (2001). Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In *Proceedings of ACL Conference*, Toulouse, France.

Hammarström, H. and L. Borin (2010). Unsupervised learning of morphology. *Computational Linguistics 37*(2), 309–350.

Hana, J. (2008). Knowledge- and labor-light morphological analysis. *OSUWPL 58*, 52–84.

Hana, J., A. Feldman, and K. Aharodnik (2011). A low-budget tagger for old czech. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR, USA, pp. 10–18. Association for Computational Linguistics.

Hana, J., A. Feldman, and C. Brew (2004, July). A resource-light approach to Russian morphology: Tagging Russian using Czech resources. In D. Lin and D. Wu (Eds.), *Proceedings of EMNLP 2004*, Barcelona, Spain, pp. 222–229. Association for Computational Linguistics.

Huddleston, R. and G. K. Pullum (2001). *The Cambridge Grammar of the English Language*. Cambridge University Press.

Karttunen, L. and K. R. Beesley (1992). Two-level rule compiler. technical report istl-92-2. Technical report, Xerox Palo Alto Research Center, Palo Alto, CA.

Kazakov, D. (1997). Unsupervised learning of naive morphology with genetic algorithms. In W. Daelemans, A. van den Bosch, and A. Weijtera (Eds.), *Workshop Notes of the ECML/Mlnet Workshop on Empirical Learning of Natural Language Processing Tasks*.

Kohonen, O., S. Virpioja, and K. Lagus (2010). Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, SIGMORPHON '10, Stroudsburg, PA, USA, pp. 78–86. Association for Computational Linguistics.

Koskenniemi, K. (1983). Two-level model for morphological analysis. In *IJCAI-83*, Karlsruhe, Germany, pp. 683–685.

Kurimo, M., S. Virpioja, V. T. Turunen, G. W. Blackwood, and W. Byrne (2010). Overview and results of morpho challenge 2009. In C. Peters, G. M. D. Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peas, and G. Roda (Eds.), *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, Volume 6241 of *Lecture Notes in Computer Science*, pp. 578–597. Springer.

Levenshtein (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory 10*(8), 707–710.

Marcus, M., B. Santorini, and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics 19*(2), 313–330.

Merialdo, B. (1994). Tagging english text with a probabilistic model. *Computational Linguistics 20*, 155–171.

Mikheev, A. and L. Liubushkina (1995). Russian Morphology: An Engineering Approach. *Natural Language Engineering 3*(1), 235–260.

Monson, C. (2009). *ParaMor: From Paradigm Structure to Natural Language Morphology Induction*. Ph. D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University.

Oflazer, K., S. Nirenburg, and M. McShane (2001). Bootstrapping morphological analyzers by combining human elicitation and machine learning. *Computational Linguistics 27*(1), 59–85.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* Singapore: World Scientific Publishing Co.

Schone, P. and D. Jurafsky (2000). Knowlege-free induction of morphology using latent semantic analysis. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2000).*

Smith, N. A. and J. Eisner (2005). Contrastive estimation: training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, Stroudsburg, PA, USA, pp. 354–362. Association for Computational Linguistics.

Tepper, M. and F. 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), Hyderabad, India, Jan 7-12*, pp. 17–24.

Tepper, M. and F. Xia (2010). Inducing morphemes using light knowledge. *ACM Transactions on Asian Language Information Processing 9*(1), 3:1–3:38.

Wang, Q. I. and D. Schuurmans (2005). Improved estimation for unsupervised part-of-speech tagging. In *Proceedings of the 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'2005)*, pp. 219–224.

Wicentowski, R. (2002). *Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework.* Ph. D. thesis, The John Hopkins University, Baltimore, MD.

Wicentowski, R. (2004). Multilingual noise-robust supervised morphological analysis using the wordframe model. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology*, SIGMorPhon '04, Stroudsburg, PA, USA, pp. 70–77. Association for Computational Linguistics.

Yarowsky, D. and R. Wicentowski (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Meeting of the Association for Computational Linguistics*, pp. 207–216.

Zipf, G. K. (1935). *The Psychobiology of Language.* Houghton-Mifflin.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least-Effort.* Addison-Wesley.