# Combining Finite State and Corpus-based Techniques for Unknown Word Prediction

Kostadin Cholakov and Gertjan van Noord
University of Groningen
PO Box 716
9700 AS Groningen, The Netherlands
*k.cholakov@rug.nl*, *g.j.m.van.noord@rug.nl*

## Abstract

Many NLP systems make use of various lexicons and dictionaries. However, unknown words are a major problem for such resources when applied to real-life data. We propose a method that combines finite state techniques and web queries to deliver possible analyses for a given unknown word and to generate its paradigm. We ensure the general applicability of our approach by applying it to a test set of Dutch words.

## 1 Introduction

Unknown words are a major hindrance for the performance of NLP tools that make use of lexicons and dictionaries. To overcome this problem, most applications try to extract (partially) the necessary morphological knowledge by implementing various heuristics and unknown word guessers.

In this paper, we present a two-phase method which delivers an accurate morphological analysis for a given unknown word and generates its paradigm. It deals with *open-class* words– nouns, adjectives and verbs based on the assumption that the other word classes are already covered by most lexicons. We test its efficiency and accuracy by applying it to real-life Dutch data. Dutch is a language with a productive inflectional morphology that exhibits quite a few interesting phenomena and thus poses a challenge for morphological processing.

In the first phase we use finite state techniques which are one of the most common approaches for morphological processing (Beesley and Karttunen, 2003; Petitpierre and Russel, 1995) since they can conveniently be used for both analysis and generation. We employ a small set of *non-deterministic* 'unweighted' finite state transducers (FSTs) whose manually encoded rules cover *regular* morphological phenomena.

Since our method deals with unknown words, it does not have access to any additional information apart from the limited knowledge provided by the word structure. Restricted only by that limited knowledge the FSTs, in *analysis mode*, identify all *possible* forms and roots allowed by the word structure. For example, the word *schnabbel* (a gig) is analysed as a singular noun, a base adjective and a first person singular present verb. As a consequence, three possible roots are produced. Since there is no way to know which of them is the correct one, in *generation mode*, all possible paradigms for each of these roots are generated.

The problem of *disambiguating* the output of the FSTs is dealt with in the second phase. We use Yahoo to search the web for each root and generated paradigm form and, based on the number of occurrences found, we try to identify the correct root and paradigm of the unknown word. Commercial search engines have already been successfully used for various NLP tasks (Keller and Lapata, 2003) and it is our claim that they are sufficient for ours as well. Since the *whole paradigm* of the unknown word is generated, it would be very difficult to find a large number of occurrences for each form in a wrong paradigm. For example, the generated adjective and verb forms for *schnabbel* have no or very few occurrences on the web and they can be safely rejected.

A similar approach, described in (Adolphs, 2008), applies finite state techniques to generate possible inflectional classes for unknown German words. However, disambiguation is done by using metrics based on frequency counts obtained from a corpus. Thus disambiguation depends heavily on the size and the gender of the corpus which is a drawback in comparison with the virtually unlimited data in the web our method has access to. If a word is, for instance, both a noun and a verb, it is possible that it would occur only as a noun in a given corpus and the method would fail to deal with the morphological ambiguity. (Nakov et al., 2003) use a rule-based approach to guess the morphological classes of unknown German nouns where each induced rule is ranked in the manner of (Mikheev, 1997). However, it is not clear if the method can scale to other word classes. (van den Bosch and Daelemans, 1999) apply memory-based learning to provide a detailed morphological analysis of Dutch. The method is tested on frequent dictionary words and only an *estimate* is provided about its expected performance on real-world data.

We should mention that the work described here is part of an algorithm for the automated acquisition of lexical types for words unknown to the Alpino grammar and parser (van Noord, 2006). The information provided by the generated paradigms is used as features in a statistical classifier which predicts lexical types for each unknown word. We also take into account occurrences of the unknown word in different contexts to extract additional features, including the type(s) assigned by the Alpino POS tagger (Prins and van Noord, 2001) and types which Alpino allows as plausible in the particular context. Therefore, *both* the morphology of the unknown word and its context are considered in the prediction process. For more details, see (Cholakov, 2009).

The remainder of the paper is organised as follows. Section 2 presents the morphological phenomena which are relevant for our experiments. Section 3 describes the FSTs and investigates their coverage and degree of non-determinism. Section 4 describes the web heuristics used to disambiguate the output of the FSTs and presents the experiments with the test data. Section 5 concludes the paper.

## 2 Morphological Phenomena

In this section we present the morphological phenomena which the FSTs account for. We begin by presenting some rules that have effect on all the three POS classes we consider.

In Dutch, the vowels **a, e, o, u** and **i** are either long or short. A vowel is long if it is doubled (e.g., *m**aa**n*– moon), if it is in a vowel combination (*l**ie**f*– sweet, dear) or when it is at the end of a syllable (*m**a**-ken*– to make). The general rule is that the type of a vowel, short or long, is preserved in all word forms. In particular contexts, a vowel is kept long by doubling it:

(1) *ma-ken–m**aa**k*
(to make, INF–1st PER.SG.PRES)

After removing the *-en* suffix from the infinitive, we get the form **mak* and the vowel turns into a short one. To prevent this, *a* is doubled. When *en* is added to form plural in (2), *u* would become a long vowel– **st**u**-ken*. To prevent this, the following consonant is doubled– *stuk-ken*.

(2) *stuk–stu**k**ken* (piece–pieces)

However, there are some exceptions to these rules. Depending on whether the syllable containing the vowel in question is stressed or not, the type of the vowel can change for words with stems ending in *-el, -er* and *-ig*. If the syllable is stressed, then the vowel preserves its type in all word forms as shown in (3-a). If the respective syllable is not stressed, then the vowel is not doubled and it turns into a short one as in (3-b).

(3) a. *de-len–deel* (parts–part)
b. *re-ge-len–re-gel* (rules–rule)

The same also applies to the doubling of consonants. In (4-a) *r* is doubled to keep the vowel short. However, this is not the case in (4-b) because the stress falls on another syllable and thus, *e* turns into a long vowel.

(4) a. *sper–sper-ren*
(to bar, 1st PER.SG.PRES–INF)
b. *coun-ter–coun-te-ren* (to strike back, 1st PER.SG.PRES–INF)

One last important rule is that a morpheme cannot end in *-v* or *-z* and they are replace by *f* and *s*, respectively: *rei-zen–reis* (to travel, INF–1st PER.SG.PRES) and *le-ven–leef* (to live, INF–1st PER.SG.PRES).

***Noun Inflection.*** Most Dutch nouns form plural by adding *-en* to the singular form, as shown in (2). However, some nouns take *-s* to form their plural: *jongen-jongens* (boy-boys), *tram-trams*.

***Adjective Inflection.*** Most adjectives in Dutch have base, comparative and superlative forms. Comparative is normally formed by adding *-er* to the base form: *snel-sneller* (fast-faster). Superlative is formed by adding the suffix *-st* to the base form: *snel-snelst*. However, base forms that end in *-s* take only *-t* to form superlative:

(5) *geeps-geepst* (pale–the most pale)

Additionally, when adjectives are used attributively, they get an *-e* suffix. The only exception is when they precede a neutral noun which is not used with a definite article or a pronoun. Some adjectives, for example the ones ending in *-en* which are mostly adjectives denoting material, do not exhibit that kind of inflection: *de gouden ring* (the gold ring).

***Verb Inflection.*** The starting point for verb inflection is the verb stem. The only thing which needs to be explained in connection with the results of our experiments is the formation of the past participle (psp). It is formed by adding the prefix *ge-* to the stem and, depending on the final consonant of the stem, either a *t* or *d* suffix is attached. However, if the stem already ends in *-t* or *-d*, no suffix is added.

An exception to these rules are verbs with *separable* particles which form psp by inserting *ge* between the separable particle and the verb stem–*op**ge**ruimd* (to clean). Verbs starting with *be-, er-, ge-, her-, mis-, ont-* and *ver-* form psp without *ge-*: *vertel-vertel**d*** (to tell). These particles are also known as *inseparable*. For a detailed discussion of Dutch morphology, see (de Haas and Trommelen, 1993).

## 3 Finite State Morphology

We employ a set of FSTs to cover the morphological phenomena presented in the previous section. For example, if the input word is *stukken* from (2), the transducer for plural nouns should produce *stuk* as a possible root form in analysis mode, and then, given this root form, it should output *stukken* in generation mode. In our experiments, the root of a given noun is its singular form, the root of an adjective– its base form, and the one of a verb is its stem.

We have used the Stuttgart Finite State Transducer (SFST) tools to implement and run a number of separate transducers which are shown in Table 1.

| POS | transducers |
|---|---|
| nouns | singular, plural |
| adj | base, comparative, superlative |
| verbs | sg1, sg2/3, pl/inf, past-sg, past-pl, psp |

**Table 1:** *Transducers used*

We use words from the CELEX morphological database (CELEX, 1995) as a development set in order to: **i)** investigate if all target phenomena are covered by the FSTs and **ii)** to have some notion of their degree of non-determinism. CELEX contains about

380K word forms corresponding to nearly 125K headwords. It is a very suitable resource for our purposes since it covers a large number of different morphological phenomena.

Three word sets are randomly selected: 2000 plural nouns, 2000 superlative adjectives and 2000 first person singular verbs and they are processed with the FSTs. We take superlative adjective forms and plural nouns in order to ensure that we deal with comparable adjectives and countable nouns. The paradigms of the selected words are extracted from CELEX, so we can check the paradigms generated by the FSTs against them.

The results for the analysis of the three extracted word sets are given in Table 2. First, the words of each set are analysed with the respective transducer and the output is a set of possible root forms for each word. The number of analyses is divided by the number of analysed words to get the *analyses per word* ratio. This indicates how deterministic the applied transducer is in analysis mode.

Next, we use the candidate roots to generate paradigms for each word and we check them against the paradigm we extracted from CELEX for the respective word. If the correct paradigm was found, the root which it was generated from is saved and thus, a list of correct roots for each of the three word sets is produced.

| | nouns | adj | verbs |
|---|---|---|---|
| analysed | 1995 | 1999 | 1963 |
| analyses/word | 1.18 | 2.93 | 1.11 |
| correct roots | 1946 | 1998 | 1567 |

**Table 2:** *Analysis mode results*

The words which failed to be analysed are very irregular forms which makes it impossible for an analysis to be produced, e.g. *vaklui-vakman* (experts-expert), since our method deals only with regular inflection.

The FSTs do not have access to information about the word stress and therefore, cases like (3-a) and (3-b) are ambiguous because it is not clear whether the vowel should be doubled or not. For example, the possible roots for *regelen* would be **regeel* and *regel*.

The number of analyses for adjectives is so high because it is not clear whether the word is an adjective that ends in *s* and takes only *-t* to form superlative as shown in (5). If so, there is also no way to know if the root form ends in *-s* or if *s* is a replacement for a *-z*. Thus, for almost each adjective three analyses will be delivered– *boost* (angry): **boo*, **boos* and *booz*. This is also the reason for the non-determinism of the first person singular verb transducer– the output of the analysis for *leef* is *leev* and **leef*.

There are also words for which no correct paradigms are generated. This is due to the fact that those words have at least one irregular form in their paradigms–*schip-schepen* (ship-ships). However, we do not see these irregularities as an obstacle for the performance of our method since they form a closed class and are supposed to be already included in most lexicons and dictionaries.

Next, each of the three lists with correct roots is processed by the transducers for the respective POS in generation mode to investigate the non-determinism of the generation. Table 3 shows the average number of *generated forms per root* for the transducers with non-deterministic generation.

| transducer | forms/root |
|---|---|
| plural | 1.21 |
| base | 1.21 |
| comparative | 1.38 |
| pl/inf | 1.36 |
| psp | 1.11 |

**Table 3:** *Generation: non-deterministic transducers*

Except for the psp transducer, the non-determinism is caused by cases like (4-a) and (4-b) where, depending on which syllable is stressed, the final *l, r* and *g* can be doubled or not. The non-determinism of the psp generator is due to the fact that some of the verb particles can be both separable and inseparable. For example, we have *omklemd* (to clasp) but there is also *omgekanteld* (to tip over). Therefore, both **omkanteld* and *omgekanteld* are generated.

# 4 Disambiguation Phase

## 4.1 Web Search Heuristics

Since we showed that the FSTs cover all target morphological phenomena, the only remaining issue is to disambiguate their non-deterministic output. We resolve this problem by using Yahoo to obtain the search hits for a given root and the word forms it generates. If a given form is found more times than a certain threshold (500 in our experiments), it is very likely that the form is a valid member of the paradigm. A paradigm is considered valid if all its forms have passed the threshold. Further, we limit the search only to pages considered by Yahoo to be in Dutch and which are from the Netherlands.

In order to evaluate the performance of our method, a test set that consists of nouns, adjectives and verbs which have between 40 and 100 occurrences in large Dutch newspaper corpora (∼530M words) has been created. This selection is based on the assumption that unknown words are typically *less frequent*. The corpora have already been parsed with the Alpino parser and many words have been added to the lexicon of the Alpino grammar. The lexical entries of the chosen words provide our gold standard.

Verbs are the easiest POS for processing with our method because in most cases, all 6 forms– *sg1, sg2/3, pl/inf, past-sg, past-pl* and *psp*– are different and the chance of finding enough counts for all forms in a wrong paradigm is minimal. We start by checking if there are enough search hits for the root, i.e. the sg1 form in this case. In order to be sure that the occurrences found are actually the ones of the root, the query sent to Yahoo consists of the first person singular personal pronoun and the root– *ik* + root.

However, it is very difficult to capture cases like *opruimen* with that query since the separable particle occurs at the end of the sentence. To deal with this, we automatically check whether the given candidate root starts with a separable particle and if so, we ignore that particle and search only for the verb stem– *ik ruim*.

Since the combination of *ik* and the root form can be very rare for some verbs, we set a threshold of only 50 occurrences. Even if a wrong root is able to get through, it would be always discarded, if there are not enough search hits for one of the forms it generates. The same threshold is also used for the past singular and the past plural forms since those are not that frequently used in Dutch.

For adjectives, we want to find the base, comparative and superlative forms and their inflected counterparts. However, not every adjective has all six forms. Non-comparable adjectives like *amorf* (amorphous) have only the base and the base inflected form. There are also some adjectives like *romantisch* (romantic) which form superlative by using the word *meest* (most) in front of the base form.

We start again by looking if there are enough search hits for the proposed root which is the base adjective form in this case. If all six forms are found, then the generated paradigm is taken to be the final result. However, in order to be able to deal with cases like *romantisch*, we also allow for paradigms where only the two base and the two comparative forms are found.

The adjectives for which no paradigm has been generated are processed again but this time we assume that these adjectives are non-comparable and thus, we search only for the two base forms. If their counts are above the threshold, the paradigm is considered to be valid. We do this in a separate round in order to avoid situations where a wrong root is able to produce two valid base forms due to chance.

Nouns are the most complicated POS to process because there are only 2 forms, singular and plural, which makes it more difficult to obtain reliable results. One example is *schel-schellen* (bell-bells). If the input word is *schellen*, there are two possible roots for it– *schel* and **schellen*. In the latter case, there are enough hits for the generated plural form **schellens* because *schellen* happens to be a family name and *schellens* is its genitive form– *Schellens methode* (Schellen's method). Thus a wrong paradigm will be generated.

To prevent this, we search for the combination of the indefinite article *een* and the root. However, not all nouns combine with *een* (e.g. mass nouns) and that is why, if there are not enough search hits, we also search for the combination of the root and *is* (to be, 3rd PER.SG.PRES). The queries for *een schellen* and *schellen is* return less than 100 hits and this root is correctly discarded.

Once a root form has passed the threshold, we can also determine the definite article of the noun. Depending on their gender and number, Dutch nouns are used either with the *de* definite article (for masculine and feminine, and also plural) or the *het* article (for neuter). We search for occurrences of *de* + root and *het* + root and return the article with the higher number of search hits.

The query for the generated noun plural form is *de+plural*. Since the generation is not deterministic, there might be more than one plural form which passes the threshold. In this case, the most frequent one is taken to be the final output. If there is more than one root that was able to generate a paradigm, the paradigm of the one with the most search hits is taken to be the final outcome.

However, there are many compounds whose root forms, combined with the indefinite article have very low number of occurrences: *Marshall-plan* (the Marshall plan), *zaak-Bouterse* (the Bouterse case), *zendingswerk* (missionary work). Normally, the compound head is a common word and the paradigm could be generated by processing the head instead of the whole compound. In most of the cases, the head is the rightmost part of the compound– *plan* in *Marshall-plan* and *werk* (work) in *zendingswerk*.

For words which are joined by a hyphen, splitting the compound is straightforward– the word on the right side is considered to be the head of the compound. However, if this word starts with a capital letter, like in *zaak-Bouterse*, we assume that the head of the compound is on the left side since the right part is most probably a name. For compounds without a hyphen, we take the chunk from the third letter to the end of the word. If there are at least 10,000 search hits for this chunk, it is considered to be the head of the compound. Otherwise, the chunk from the fourth letter is taken and so on, until we find a chunk with enough occurrences.

Let us put it all together. The query for the root is sent but the threshold set for it is 100 because nouns tend to occur less frequently together with the indefinite article. After a valid root is found, the definite article is determined and then the search hits for the generated plural form(s) are obtained. If no paradigm was generated for a given word, we assume that it might be a low-frequent compound and we try to split it. If there is a valid head, the same procedure is applied to it in an attempt to generate its paradigm which is also the paradigm of the whole compound.

## 4.2 Results and Error Analysis

The test set, described in Section 4.1, consists of 2593 unique words but the gold standard contains a total of 2781 entries– 1368 nouns, 729 adjectives and 684 verbs. This is due to the fact that some morphologically ambiguous words have more than one valid paradigm. For instance, many psp can also act as adjectives: *gebruind* is the psp of the verb *bruinen* (to turn/make brown) but it is also an adjective– 'tanned, sunburnt'. That is why this word is listed both as an adjective and a verb. There are 188 cases of morphological ambiguity.

However, our method is able do deal with them. When an input word comes in, the FSTs for each POS try to process it and if some of them manage to generate a paradigm, the web heuristics for the respective POS are applied. After the disambiguation phase, only the verb and the adjective paradigms for *gebruind* 'survived' and they are correctly taken to be the final outcome. However, in very rare cases there is morphological ambiguity within the same POS. For instance, the word *kussen* is the plural form of *kus* (a kiss) but it also means 'a cushion' whose plural form is *kussens*. In that case, only the paradigm whose root form has more search hits is returned.

Table 4 presents the overall results and the results for each POS. *Coverage* indicates the number of words with a paradigm generated and *accuracy* is the number of the generated correct paradigms.

| | overall | nouns | adjectives | verbs |
|---|---|---|---|---|
| total | 2781 | 1368 | 729 | 684 |
| coverage | 96.55% | 98% | 98.91% | 90.94% |
| accuracy | 99.63% | 99.33% | 100% | 99.84% |

**Table 4:** *Web experiment results*

No paradigm has been generated for the uncountable noun *smelt* (smelt). All other nouns without paradigms are compounds which form plural in an irregular way. The same is also valid for the verbs which have not been covered– all of them have irregular past or psp forms which are not handled by our technique. Most of the adjectives not covered are non-comparable adjective ending in *-en* and designating material, e.g. *satijnen* (satin). Such adjectives have only a base form and thus, no valid paradigm is found for them.

Next, we see in Table 4 that some of the generated noun and verb paradigms are wrong. Most of the wrong noun ones are compounds whose head is the word *kind* (child). For example, the compound *pleegkind* (foster child) is correctly split and we try to generate a paradigm for its head *kind*. However, it has an irregular plural form– *kinderen*. Nevertheless, the 'regular' plural form **kinden* has more than 1000 occurrences in the web and it is taken to be a valid one. A manual examination of the results of the web search showed that all the occurrences were actually a typo– the actual meaning was *kind en* (child and) but in many web forums, blogs, etc. they were written as one word.

Another source of error for the noun paradigms are words from French origin, e.g. *secretaris-generaal* (secretary-general). In this case the head of the compound is on the left of the hyphen and thus it is the part that is inflected. The right plural form is *secretaris**sen**-generaal*. However, our compound splitting heuristics takes the right part to be the head which, in this case, happens to be the noun *generaal* (general). Its paradigm is generated but this is not the correct paradigm of the compound.

The only verb with a wrong paradigm generated is the irregular verb *schijten* (to shit). The incorrect psp generated by the psp FST, *geschijt*, has enough search hits because it happens to be a less frequent case of nominalisation– namely, *ge + root*. This error, however, can be tolerated, since it is potentially possible in the rare case of irregular verbs which allow for such nominalisation and have roots that end in *-t* or *-d*.

The achieved results show clearly that simple but carefully designed web queries can be successfully used to disambiguate the output of a non-deterministic finite state morphology for Dutch. Our method is able to deal with all major and common morphological phenomena except for few cases that include irregular forms or very rare combinations of factors.

## 5 Conclusion

We proposed a combination of finite state techniques and specially designed web queries and heuristics to deliver morphological analyses for a given unknown word and to generate its paradigm. Our approach does not require access to sophisticated linguistic information but instead employs the web as a linguistic resource. The successful application of the method to real-life Dutch data proved its efficiency and high accuracy.

Naturally, languages with a large number of inflectional variants would be more problematic for our approach due to the much larger morphological ambiguity they exhibit. However, we also expect that the high number of forms in the paradigms would facilitate the disambiguation of the FSTs output. As we showed, it is much easier to validate a paradigm that consists of 6 forms since the web results are more reliable for it. We will investigate the scalability of our method in future research.

## References

Adolphs, P. (2008). Acquiring a poor man's inflectional lexicon for German. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology.* CSLI Publications, Palo Alto, CA, USA.

CELEX (1995). The CELEX lexical database-Dutch, English, German. CD-ROM.

Cholakov, K. (2009). Towards morphologically enhanced automated lexical acquisition. In *Proceedings of the 14th ESSLLI Student Session*, Bordeaux, France.

de Haas, W. and Trommelen, M. (1993). *Morfologisch Handboek van het Nederlands: Een overzicht van de woordvorming.* SDU, 's Gravenhage, The Netherlands.

Keller, F. and Lapata, M. (2003). Using the web the obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Mikheev, A. (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics*, 23.

Nakov, P., Bonev, Y., Angelova, G., Gius, E., and von Hahn, W. (2003). Guessing morphological classes of unknown German nouns. In *Proceedings of RANLP 2003*, Borovets, Bulgaria.

Petitpierre, D. and Russel, G. (1995). Mmorph– the multext morphology program. Technical report, Carouge, Switzerland.

Prins, R. and van Noord, G. (2001). Unsupervised POS-tagging improves parcing accuracy and parsing efficiency. In *Proceedings of IWPT*, Beijing, China.

van den Bosch, A. and Daelemans, W. (1999). Memory-based morphological analysis. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 285–292, San Francisco, CA.

van Noord, G. (2006). At last parsing is now operational. In *Proceedings of TALN*, Leuven, Belgium.