Using Rough Sets Theory to predict German Word Stress

Stefan Rapp Michael Jessen Grzegorz Dogil

14th February 1995

Institut für Maschinelle Sprachverarbeitung
– Experimentelle Phonetik –
Universität Stuttgart
Azenbergstr. 12
70174 Stuttgart
e-mail: {rapp,jessen,dogil}@phonetik.ims.uni-stuttgart.de

Abstract

Rough Set Theory [Paw82, Paw91] is a framework for reasonably dealing with imprecise or uncertain data. It can be used to implement application independent symbolic machine learning techniques. A special application of a Rough Set based machine learning algorithm is presented that can predict german word stress by extracting symbolic rules from sample data. A comparison is made between the predictive power of different codings of the sample data.

Keywords: Machine Learning, Phonology, Rough Set Theory, German Word Stress.

1 Introduction

In this paper, an application of a machine learning program in a linguistic domain is presented. The purpose of this paper is twofold: On the one hand, it addresses phonologists, because the results derived with a machine learning technique give interesting insights to the domain under consideration, which is stress assignment in German morphologically simple words. On the other hand, it addresses AI researchers since it shows the advantagous use of AI techniques for examining aspects of existing theories and formation of new theories.

Symbolic rules generated by machine learning algorithms can

- provide a basis for comparing different theories by trying to map the computationally derived rules to the rules of different theories stated by human researchers,
- discover some new rules with high predictive power which were up to now not considered by researchers,
- show structural disadvantages e. g. of different codings of a problem domain,
- suggest novel viewpoints to the application problem domain and thus give new impulses for future research in the field.

While the first and second of these points have been predominant during the first stage of research, this paper gives a stronger bias to the latter two points. The problem domain under consideration is a current issue in linguistic research: the prediction of the primary stress in German morphologically simple words with symbolic rules.

In this study symbolic rules have been induced with a machine learning approach that is based on the theory of Rough Sets. Since this interesting approach is until now not very well known among AI researchers, the fundamentals of Rough Set Theory will be presented in the next section of the paper.

Another point that will be summarised in this paper is the way the data were selected and computationally preprocessed to extract attributes for the machine learning program.

2 Rough Set Theory

Rough Set Theory was developed as a theoretical framework for reasoning about unprecise data. It can be seen as an extension of classic set theory to suit the needs of three-valued logic.

A universe U of objects x_i is partitioned by an *indiscernibility relation* R. Strictly speaking, R is a family of equivalence relations R_i.[1] The introduction of an indiscernibility relation is a very natural thing to do: the indiscernibility relation might correspond to our senses by which we perceive our environment, or to the possible discrete values we obtain from measuring some real–valued attributes, or, in a machine learning context, to the finite number of combinations of attribute values that describe example instances.

In Rough Set Theory, the indiscernibility relation is said to be the *knowledge* that we have about the universe U. Objects belonging to the same equivalence class cannot be distinguished. This clearly corresponds to every day experience that we cannot distinguish two objects unless they differ in at least one aspect (we thus *know* a difference).

The basic concept in Rough Set Theory (like in classic set theory) is the set. The problem under consideration is the description of an arbitrary set with the knowledge that is expressed by the indiscernibility relation. How can we describe a set X that contains an element, say x_j, that seems to be equal to another one, say x_k, which is not in the set? We cannot. But we can approximate the set with two other sets that are an upper and lower limit to the set. These two sets are built from elementary sets as follows:

$$\begin{aligned} \underline{R}X &= \bigcup\{Y \in U/R \mid Y \subseteq X\} \\ \overline{R}X &= \bigcup\{Y \in U/R \mid Y \cap X \neq \emptyset\} \end{aligned}$$

The *lower approximation* $\underline{R}X$ is the union of all elementary classes that are completely inside of X, that is: in every elementary class from which $\underline{R}X$ is build, there is no object that lies outside of X. The *upper approximation* $\overline{R}X$ is defined to be the union of all elementary classes that contain at least one element of X. To show the connection to three valued logic, we can state:

- If an element belongs to $\underline{R}X$, this element *surely* is in X.
- If an element belongs to $\overline{R}X - \underline{R}X$,this element *possibly* is in X.

[1] See [Paw91] for an exact algebraic definition.

- If an element belongs to $U - \overline{R}X$, this element *surely* is *not* in X.

The approximation of an arbitrary set by $\underline{R}X$ and $\overline{R}X$ works well, if the set is somewhat close to the borders of the elementary sets. In this case, the upper and lower approximation are quite similar, the upper being only some few elementary classes larger than the lower approximation.

Our knowledge about a given set X is very low if in every elementary set there are some objects belonging to X and some not belonging to X. Then $\underline{R}X = \emptyset$ and $\overline{R}X = U$ and our possibilities to exploit knowledge about X are poor since every object could or could not be in X. From a theoretical point of view, things are clear. Our knowledge about U and X is too small: perhaps we have chosen the wrong attributes to describe our objects, or our instruments for measuring their attribute values are not exact, or maybe there is just too much noise in the data. But practically, what can we do if we have no more attributes or no better instruments?

Rough Set Theory states that an elementary class E with 99 objects belonging to X and only 1 not belonging to X can not be a part of $\underline{R}X$.

An obvious way to make use of the information contained in E is to loosen the definition of $\underline{R}X$ and $\overline{R}X$. We define:

An elementary class belongs to $\underline{R}_\beta X$ iff a $(100\% - \beta)$ majority of its elements belongs to X.

An elementary class does not belong to $\overline{R}_\beta X$ iff a $(100\% - \beta)$ majority of its elements does not belong to X.

The parameter β controls how much we loosen our new definitions. Obviously, β must be restricted to a range of 0%, where the new definitions converge to the original definitions, up to 50% since above this limit one can not reasonably speak of a majority. This extended version of Rough Set Theory was proposed by Ziarko [Ziar93] and is called *Variable Precision Rough Set Model.*

This theory maps to the setting of machine learning as follows: The concepts to be learned correspond to arbitrary sets which are to be described by means of the indiscernibility relation R. R, as a family of equivalence relations, corresponds to the classes of attribute values that are known about the examples. The task is to abstract from the full description of the examples to a few simple rules. There are different ways of exploiting Rough Set Theory for the purposes of machine learning. See [Ziar93, References 3, 7, 12, 15, 17] for algorithms. One of them, [Wong&al86], uses a measure from Rough Set Theory, the 'discriminant index', to replace the 'information gain' measure of ID3 [Qui86].

For the experiments described in this paper, the commercially available software DataLogic/R, the underlying theory of which is the Variable Precision Rough Set Theory [Ziar&al93], was used.

3 German Word Stress

German word stress is an interesting domain to apply symbolic machine learning techniques to. Among linguists, there are differing views on German word stress. Some think that word stress in monomorphemic words is a lexical feature, a property of the word. According to this hypothesis, whenever a speaker learns a new German word, he has to learn the corresponding stress pattern with it. These linguists would tend to say that there are no rules at all for predicting word stress. Among those linguists that assume the existence of rules that control which syllable of a word receives primary stress, several proposals exist on what these stress assignment rules might look like — partly contradicting each other.

A problem with the researchers' work on word stress especially in the case of German is, that most of the authors state their hypothesises in a deductive manner [Jes94]: first some rules are stated which are then followed by some examples that illustrate the rules. As Jessen suggests, an inductive approach would be more adequate, since sufficient objectivity in the phonological analysis is only given with no particular hypothesis in mind. This research follows Jessen in first establishing a corpus of examples. These are phonologically analysed to extract some attributes that are believed to be relevant to word stress. The attributes derived from the analysis are then fed into the machine learning program to generate (hopefully good) rules.

4 Establishing a Corpus

If one attempts to find rules that control German word stress, one has to restrict the type of words under consideration. One problem has to do with the semantic or pragmatic issues manifested in the context of a word. These factors can heavily influence the stress pattern of a word. A frequent example for this is „übersetzen“ `/y:|b6|'zE|ts=n/` 'to translate' vs. „übersetzen“ `/'y:|b6|zE|ts=n/` 'to ferry over'. Semantic influence on word stress can be reduced when only considering monomorphemic words. Most Phonologists concentrate on this kind of word when stating rules about word stress. There are many phonologists who claim that their rules discovered for monomorphemic words also hold for monomorphemic words with additional nonnative suffixes. These are often called class 1 suffixes. For the following experiments the corpus is thus restricted to monomorphemic words or words with class 1 suffixes.

On the other hand, a lot of German words are not very interesting to phonologists because of their simple structure. Most of the native words — e. g. words of Germanic origin — are either monosyllabic such as „Haus“ /haus/ ‘house’ or bisyllabic with the nucleus of the second syllable being a schwa such as „Katze“ /'ka|ts@/ ‘cat’. Since a schwa can never be stressed, there is actually nothing to look for in this kind of word. A few other words considered native such as „Holunder“ /ho|lUn|d6/ ‘elder’ or „Heirat“ /'hai|ra:t/ ‘marriage’ actually have more than one non–schwa syllable and can thus potentially bear stress on different syllables. Some phonologists believe that there are different stress assignment rules for native and nonnative words; but since the distinction between native and nonnative words is quite controversial among phonologists, this has not been made a criterion for the selection of words for the corpus. To filter out the uninteresting cases, the corpus has been restricted to words with at least two non–schwa syllables regardless of being native or nonnative.

The next restriction on the corpus is a more practical one. The corpus was restricted to nouns to cut the number of candidates since a machine readable phonetically transcribed lexicon was not available. The decision towards nouns makes sense because nouns are considered to have the largest diversity of stress patterns [Eis91].

A large database of examples is of course desirable when trying to find rules from sample data. This is especially the case for subsymbolic or statistical learning methods like neural nets. In the symbolic machine learning community, another approach is quite popular: comparatively few but *prototypical* examples are expected to concentrate the algorithm on what it has to learn. The approach taken in this work tries to balance between these two extreme positions by selecting the list of words used in a book about German word stress by Giegerich [Gieg85] as a starting point for the corpus. The list, given in the book’s word index contains 668 words, 242 of these are nouns that are monomorphemic or with only nonnative suffix(es) and have at least two full (i. e. non–schwa) syllables. Since these words are used as examples in a book about German word stress, these words can be considered relevant or prototypical to the stress assignment task. Since Giegerich lists many exceptions to his own rules, it is not believed that the corpus is biased strongly towards Giegerich's analysis of German word stress. Note however that the stress-patterns in this list are likely to be more difficult to predict than those from alternative possible approaches: For example, one could have selected the n most frequently used words that satisfy the above restrictions or n randomly selected words. Evidently, with the latter two approaches there is a much smaller chance of getting the problematic cases (words considered as exceptional by almost every phonological analysis) that are discussed in the phonological literature and thus in [Gieg85]. However, a larger database with, say, the 1000 most frequently used nouns, should be considered when thinking of applications of this work to text–to–speech etc. Such an approach is not taken here, because here we are interested in the methodological value of using Machine Learning techniques in linguistic research rather than in high predictive accuracy which of course would be highly preferable in other domains.

5 Preprocessing the Data

A commonly accepted pronunciation dictionary of German ([Mang90]) was consulted for the phonetic transcription of every selected word. Three words that were not listed have been removed from the corpus. Whenever Mangold states more than one possibility in pronunciation and/or stress assignment, a corresponding number of cases is added to the corpus. If a variant is explicitly stated non–German, it is not included in the corpus. Note that „Establishment“ would be rejected (because it is indicated to be English), but some (at least to the authors) questionable cases like „Arrangement“ (not stated French) are retained in the list. With the variants, the corpus consists of 260 cases (239 different words).

Word stress is not believed to be a binary feature. The syllables of a word are not assigned just two levels of stress (stressed or unstressed) but rather a multitude of levels yielding a net of prominence relations between the syllables. A syllable is said to bear the primary stress, if it is more prominent than the other syllables of the word. A word can have a syllable with secondary stress, that is a syllable that is more accented than all other syllables except that syllable with primary stress etc. In this study we focus on the prediction of primary stress, one reason being that secondary stress is marked very sparingly in the pronunciation dictionaries of German, so that assigning a secondary stress would have been based on our own intuition. In [Mang90], there is not a single word among those selected that has a syllable marked with secondary stress.

The phonetic transcriptions of [Mang90] for the selected words were stored in a slightly modified version of SAM–PA notation along with syllable break information added by the authors. Via a simple definite clause grammar, the following attributes were computationally extracted for every syllable:[2]

1. Typ: the phonological type of the syllable. Possible values are: `@` (schwa-syllable without onset), `C@` (schwa-syllable with nonempty onset), `Vu` (Syllable with „ungespanntem"/lax vowel, coda is empty), `Vu(C)` (same as above but corresponding orthography of next syllable starts with bb, ck, ff, kk, ll, mm, nn, pp, rr, ss, tt or x)[3], `VuC` (same as above but exactly one consonant follows the vowel in the coda), `VuCC1` (same as above but the coda consists of two or more consonants), `V_V` (syllable with nucleus consisting of a diphthong), `Vg` (syllable with „gespanntem"/tense vowel, coda is empty), `VgC1` (as above but with nonempty coda), `konson.` (Nucleus being a syllabic consonant like the final syllable in „Büffel" or „Ensemble").

2. Onset: the length of the onset. Possible values are: `vorh.` (a nonempty onset), `n.vorh.` (an empty onset).

3. Hoehe: the height of the vowel in the nucleus. Possible values are: `hoch` (a high vowel, one of `/I/`, `/U/`, `/Y/`, `/i/`, `/u/`, `/y/`), `mittel` (a vowel of medium height, one of `/E/`, `/O/`, `/9/`, `/e/`, `/o/`, `/2/` and the two schwas `/@/` and `/6/`), `tief` (a low vowel, `/a/`)[4], `steigend` (rising, the nucleus consists of a diphthong, one of `/ai/`, `/au/`, `/Oy/`), `konson.` (consonantic, the nucleus consists of a syllabic consonant).

[2] every syllable consists of two parts: an onset and a rhyme. The latter is further divided into a nucleus and a coda. Onset and coda are build from consonants and unsyllabic vowels and are allowed to be empty. The nucleus must consist of a vowel, a diphthong or a syllabic consonant and must be present. Note that some phonologists count a diphthong partially to the nucleus and partially to the coda.

[3] this is used to model ambisyllabicity or „scharf geschnittene"/sharp cut syllables claimed to be relevant to the stress assignment task by some phonologists

[4] [Mang90] only transcribes a single „a-Laut" for German words

4. Laenge: the length of the nucleus. Possible values are: `lang` (long, the vowel is marked with a colon in [Mang90], or is a diphthong), `kurz` (short, nucleus neither marked long nor a diphthong)

5. Spannung: the tension of the nucleus vowel. Possible values are: `gespannt` (tense vowels, one of `/e/`, `/o/`, `/2/`, `/i/`, `/u/`, `/y/`), `ungesp.` (lax vowels, one of `/a/`, `/E/`, `/O/`, `/9/`, `/I/`, `/U/`, `/Y/` and the diphthongs `/ai/`, `/au/`, `/Oy/`), `Schwa` (one of `/@/`, `/6/`), `konson.` (consonantic, the nucleus consists of a syllabic consonant).

6. Coda: the length of the coda. Possible values are: `offen` (open, the coda is empty), `ambiv.` (ambivalent, or halfclosed, as noted above at the `Vu(C)` category) `einf.g.` (closed, coda consists of one consonant), `mehrf.g.` (closed, coda consists of more than one consonant).

Since these attributes (of which the first can be viewed as computed from the others) were extracted from every syllable and the word with the most syllables in the corpus consists of six syllables, there are $6 \cdot 6 = 36$ attributes that describe a word. The attributes were aligned 'from right to left' so that the six attributes of the last syllable are in the first six columns, the attributes from the syllable before the last syllable fill up the next six columns and so on. Every word (or listed pronunciation variant) makes up one row in the table. The decision attribute makes up another column, having possible values of `final` (the last syllable gets primary stress, 113 cases), `penult` (the syllable before the last syllable gets primary stress, 107 cases), `antepenult` (the syllable before the syllable before the last syllable gets primary stress, 40 cases).

6 Experiments

Three experiments were carried out with the data. The first two experiments were done to examine two assumptions most phonologists make in their analyses. The first assumption is that stress assignment in German operates from right to left. Since there is a wide agreement that primary stress can not be more than three syllables away from the end of a word, (at least for non–compound words), information from the syllables further to the beginning of the word is irrelevant. We would expect that e. g. rules that predict penultimate stress would rely on information from the last two syllables. The second assumption is the choice of classes that are used in the phonological literature. E. g. height of the nucleus vowel of a syllable is irrelevant to its assignment to a phonological type like VuC. It is possible that valuable information is lost through this generalizations (as already mentioned in a comparable study for Dutch, [Dael&al93]).

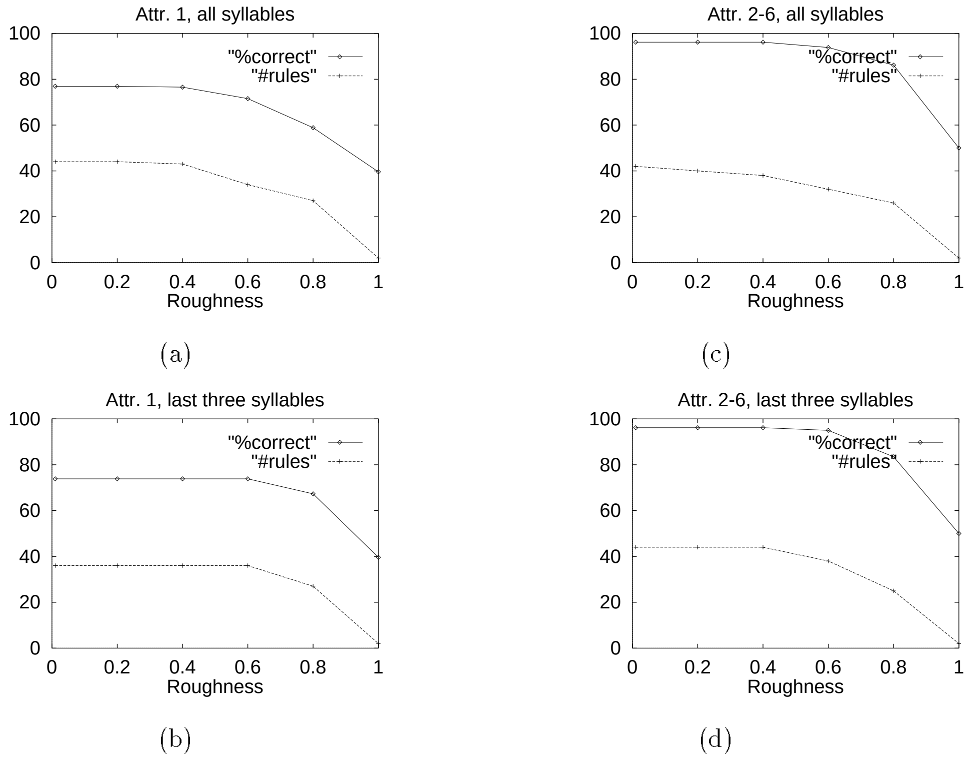
The second experiment was performed to estimate the predictive accuracy that can be achieved using phonological type information on the one hand and the alternative, more detailed information coded in the remaining five attributes on the other hand. It is clear that the second coding is not to be expected to yield worse results since it contains more information than the first (speaking in Rough Set terminology, the indiscernibility relation of the latter attribute coding is more finegrained than the first's, thus allowing more accurate approximations). However, a diminishing advantage of the 'full' coding over the phonological would justify the generalizations made by the phonologists.

In the third experiment we tested if the results of the previous experiment still hold, if length information is not available. [Rapp94] shows empirical evidence that the length of a syllable correlates very strongly with word stress at least in German. It could be that Mangold notes a vowel as being long *because* that syllable is stressed (and thus is produced with a long vowel by most speakers) or, on the other hand, the syllable is long and *thus* it attracts stress.

6.1 First Experiment

Figure 1 shows the results of how well the rules which were generated by the machine learning algorithm can describe the patterns in the data. For different levels of roughness (a parameter that controls the algorithms ability to generalize) there is shown the prediction rate for the data with the rules produced thereof, that is *learn* = *test* and the complexity of the rules that describe it (given in number of produced rules). Figure 1-a shows the results, if only the phonological type information is given to the Rough Set programm. Figure 1-b below shows the results if the available information is further restricted to the phonological type information of only the last three syllables. The right side of the figure shows the corresponding results, if the remaining five attributes (per syllable) are given to the learning algorithm. Again, the upper figure displays results based on the information for all syllables, the lower figure is derived only from information from the last three syllables.

All figures show that in increasing the roughness value, the number of rules decreases. The more the algorithm is urged to generalize (showing in a lower rule



(a)

(c)

(b)

(d)

Figure 1: Results from the first experiment. The figures show the number of rules that were generated and the percentage of correctly classified cases for different levels of roughness (*learn* = *test*).

complexity), the less the rules are able to describe the place of primary stress for our 260 cases (since the more general rules don't cover outliers/exceptional cases). However, if we want to predict new, unseen cases, generalization is absolutely necessary if we don't want to over-fit to the learn data. We will come to this issue in the discussion of the following experiments.

A comparison between the upper two figures with the lower ones gives justification for the phonologists' claims: word stress depends mainly on the structure of the last three syllables; the small differences can be explained with the algorithms better possibilities to model outliers when it has more information available. In other words, this is to be viewed as an artefact of our experiment setting with *learn* = *test* and will diminish in a *learn* ≠ *test* setting. But even

```
Betonung==> antep.        Betonung==> penult        Betonung==> final
======================    ======================    ======================
Attribute  Max.Loc.Str    Attribute  Max.Loc.Str    Attribute  Max.Loc.Str
=========  ===========    =========  ===========    =========  ===========
Laenge_3          0.54    Laenge_2          0.55    Laenge_1          0.47
Spannung_2        0.40    Laenge_1          0.33    Laenge_2          0.42
Hoehe_3           0.39    Coda_2            0.25    Coda_1            0.25
Coda_2            0.38    Onset_3           0.23    Laenge_3          0.14
Coda_3            0.33    Coda_1            0.22    Onset_1           0.13
Onset_3           0.32    Hoehe_3           0.18    Spannung_1        0.12
Laenge_1          0.27    Hoehe_2           0.16    Hoehe_4           0.09
Coda_4            0.23    Spannung_1        0.15    Hoehe_3           0.09
Laenge_2          0.22    Onset_1           0.14    Hoehe_2           0.08
Hoehe_2           0.22    Coda_3            0.14    Hoehe_1           0.07
                          Hoehe_1           0.13    Spannung_3        0.07
```

Figure 2: Attribute strength reports for experiment 1 (attributes 2 to 6 from all syllables, roughness value 0.60)

with this experiment, it doesn't seem to be a strong restriction when phonologists concentrate on the last three syllables for the stress assignment task. Further justification comes from data that is also produced by the machine learning program, the attribute strength report. It shows the importance of individual attributes for the classification task. Figure 2 shows this data for the run with attributes 2 to 6 from all syllables and a roughness value of 0.6. The attributes can have values ranging from 1 (most important) to 0 (not useful). Attributes of the last syllables are more to the top of the lists than those of the others. Only few attributes from syllable 4 are listed among the attributes of the last three syllables. We can say that for all roughness values, the algorithm has clearly chosen information from the last three syllables to be much more important than that from others.

As a result of the first experiment, future experiments will only rely on information from at most the last three syllables.

It is obvious that the rules generated from the full information (that is from attributes 2 to 6, plots to the right in figure 1) are more powerful in describing the data. The following experiment will show, how much this is grounded on the better abilities to overfit the learning set. (Note that attribute 1 can be

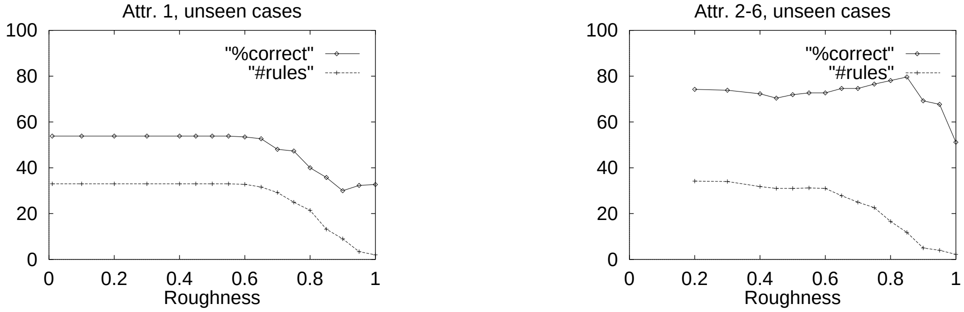


Figure 3: Results from the second experiment. The figures show the number of rules that were generated and the percentage of correctly classified cases (both averaged over five runs) for different levels of roughness (*learn* $\neq$ *test*).

computed from attributes 2 to 6 but not vice versa. In doing so we already made a generalization and thus reduced our ability to overfit to the data.)

6.2 Second Experiment

For the second and third experiment, the 260 cases were partitioned randomly into 5 parts of equal size. For every considered roughness value, there were 5 runs. In every run, the machine learning algorithm had to extract rules from four of the five parts, the learning set. Then the remaining fifth part, the test set, was used to evaluate the predictive accuracy. As can be seen in Figure 3, the full coding (to the right) has substantially better averaged prediction rates than the phonological coding (to the left) at comparable rule complexity.[5] Having a closer look at the individual runs (figure 4), the figures reveal a range of good values for the generalization parameter. Since one can not be sure how the generalization of one or another rule effects the prediction of unseen cases, the curves have some ups and downs in the range of 0.40 to 0.95, especially the runs with atributes 2 to 6. However, good values for the roughness parameter can be chosen from this range, as the averaged curve already suggests. Unfortunately, there is no best value for both codings. Only the 'full' coding runs were able to

[5] Of course, rules with attributes 2 to 6 are somewhat longer than rules with attribute 1. Note that only the number of rules is plotted and not the number of decisions that must be considered.

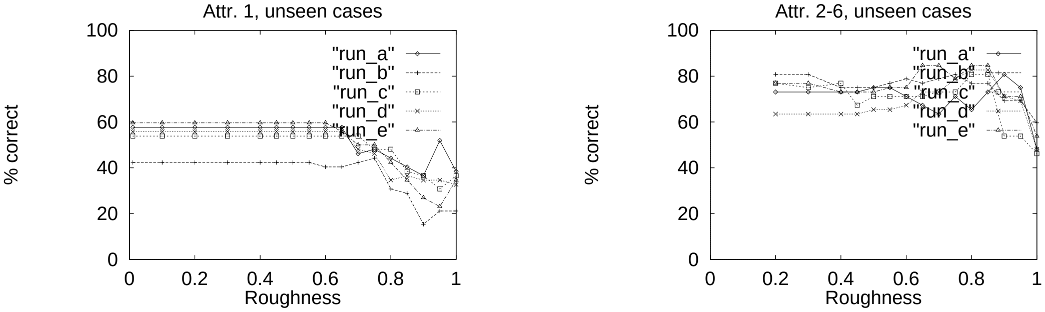


Figure 4: Results from the second experiment. The figures show the percentage of correctly classified cases for the five runs over different levels of roughness ($learn \neq test$).

produce a peak at the point of best performance (79.616% at roughness 0.85 for the averaged curve; individual runs 80.77% to 84.62% at roughness values from a range of 0.75 to 0.90). The phonological coding runs show, that prediction cannot be further improved through the algorithms abilities to generalize, since, with the exception of run b, the best results are achieved at low roughness values (53.848% for the averaged curve; 42.31% to 59.62% for the individual runs). The fact that the phonological coding lacks a peak could be seen as a hint that the phonological categories are already well generalized, and thus the machine learning algorithm is not able to improve prediction rate further through generalization.

Of course there is a methodological problem with a posteriori adjustment of parameters of the ML algorithm. Ideally, there should be experimentation with a small portion of the data to obtain good values for all parameters of the algorithm. These should then be fixed and left unchanged for further evaluation to avoid what is called 'training on the test set'. In this study, however, the amount of data is not very large, so taking a small portion of it is not advisible. Since in this study we bias on the advantages and disadvantages of the different codings and not on a good estimation for predictive accuracy, we think of the choosen procedure as a valid one.

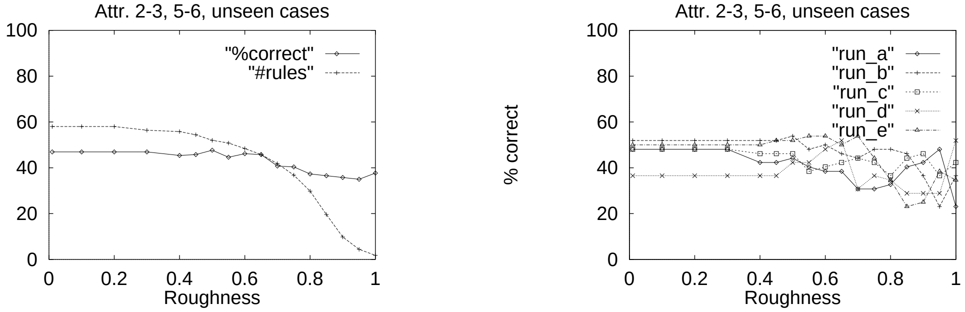


Figure 5: Results from the third experiment. The figures show the number of rules that were generated and the percentage of correctly classified cases (the left shows results averaged over five runs, the right individual runs) for different levels of roughness (*learn* $\neq$ *test*).

6.3 Third Experiment

What arises from the second experiment is the question why the 'full' coding is so much better than the phonological coding. The main difference between the two is that the phonological coding almost completely ignores the length information that is presented as individual attributes in the full coding. From the phonological coding, the only implications that can be made concerning the length attribute are the following: if the syllable is of type `V_V` (a diphthong–syllable), it is long, if it is of type `@` or `C@` (a schwa–syllable), it is short. A distinction e. g. between a syllable with long tense vowel and one with a short tense vowel is not possible with the phonological types. Since the little knowledge that the phonological types have about length is (for the full coding) also coded in the height attribute (where only diphthongs get the value `steigend`) and the tension attribute (where only schwa syllables get the value `Schwa`), it is possible to examine how much of the advantage of the full coding is based on the length attribute. Thus, for the next experiment only attributes 2, 3, 5 and 6 were fed into the machine learning algorithm. Figure 5 shows the results. On the left side there are the averaged curves for number of rules and percentage of correctly predicted cases. Again, we used the *learn* $\neq$ *test* method as described in the second experiment. The individual runs are shown on the right side of figure 5.

The figures show that with removing the length attributes from the input data, the advantage of the separate attribute coding over the phonological classes coding completely disappears. There are even slight disadvantages in predictive accuracy when considering attributes 2, 3, 5 and 6 instead of attribute 1 (47.692% instead of 53.848%). Having a look at the individual runs, shown to the right of figure 5, we see that although the algorithm can improve predictive accuracy through generalization, the results are worse than the results of the phonological coding runs (with the exception of, again, run b). As a result, these numbers seem to justify the classes commonly found in phonological literature. It seems that, if one does not accept length as a valid attribute for the description of syllable structure it makes no sense to look at the individual factors that make up the phonological types. On the other hand, if one accepts vowel length to be a valid attribute in the stress assignment task, the 'full' coding, attributes 2 to 6, is clearly superior to both other codings (reaching a predictive accuracy of roughly 80%).

7 Conclusions

These results enable us to state the following conclusions:

- The experiments showed justification for the assumption that the structure of syllables more than three syllables from the end of a word are not important to stress assignment. (Note that this rule is only stated for monomorphemic words or those with class–1 suffixes)

- The phonological classes as described above make less sense if one accepts the length of the vowel of a syllable to be a valid attribute for the description of the data and not merely a result of stressing that syllable. Note that although in a minority, there are long vowels in words of the corpus that do not bear primary stress. In omitting the length attribute from the set of describing attributes, much of the information is lost as is shown in a strong decreased predictive accuracy.

- If one does not accept vowel length as a valid attribute, the phonological classes as described above are in general better than the individual attributes from which the classes are build: the classes developed by phonologists are a good abstraction for the stress assignment task.

- In general, machine learning techniques have proven to be useful in theory formation and in validation of existing theories. Some assumptions made by researchers have received empirical support through the experiments done in this work.

The results of this study encourage further research in this field. Two things seem to be interesting to the authors: A similar approach should be tried with orthography of a word instead of its phonetic transcription as input to a machine learning program (as an application for text–to–speech). It is believed, that especially in text–to–speech applications, ignoring the length information absolutely makes no sense, since it is sometimes coded in German orthography (e. g. "‘Dehnungs-h"’, "‘Plato"’ vs. "‘Plateau"’). A comparable study should be done with an enlarged database of examples that takes into account word frequency. Rules produced from this along with stress detection from speech signals of some kind might help to cut the number of candidates in the speech recognition task.

References

[Dael&al93] Daelemans, W., Gillis, S., Durieux, G., van den Bosch, A. (1993) Learnability and markedness in data–driven acquisition of stress. *Computational Phonology. Edinburgh Working Papers in Cognitive Science,* **8,** 157–178. (avail. by anonymous FTP from scott.cogsci.ed.ac.uk)

[Eis91] Eisenberg, P. (1991) Syllabische Struktur und Wortakzent: Prinzipien der Prosodik deutscher Wörter. *Zeitschrift für Sprachwissenschaft,* **10,** 37–64.

[Gieg85] Giegerich, H. J. (1985) *Metrical Phonology and phonological structure. German and English.* Cambridge University Press, Cambridge etc.

[Jes94] Jessen, M. (1994) *A survey of German word stress.* Manuscript, University of Stuttgart.

[Mang90] Mangold, M. (1990). *Duden Ausspracheworterbuch: Worterbuch der deutschen Standardaussprache,* 3. Auflage, Bibliographisches Institut & F. A. Brockhaus AG, Mannheim.

[Paw82] Pawlak, Z. (1982). Rough sets. *International Journal of Information and Computer Sciences,* **11,** 145–172.

[Paw91] Pawlak, Z. (1991). *Rough sets: Theoretical Aspects of Reasoning About Data.* Kluwer Academic, Dordrecht, The Netherlands.

[Qui86] Quinlan, J. R. (1986) Induction of decision trees. *Machine Learning,* **1,** 81–106.

[Rapp94] Rapp, S. (1994). Maschinelles Lernen von Aspekten des deutschen Wortakzents. *Diplomarbeit Nr. 1176, Universität Stuttgart, Fakultät Informatik.*

[Wong&al86] Wong, S. K. M., Ziarko, W. and Li Ye, R. (1986). Comparison of rough–set and statistical methods in inductive learning. *International Journal of Man–Machine Studies,* **25,** 53–72.

[Ziar&al93] Ziarko, W. and Edwards, R. G. D. (1993) An application of the rough sets approach to discovery of strong predictive rules in stock market data. Available by anonymous FTP from ftp.cs.uregina.ca (/pub/ebrsc).

[Ziar93] Ziarko, W. (1991). Variable Precision Rough Set Model. *Journal of Computer and System Sciences,* **46,** 39–59.