PROBABILISTIC INDUCTION OF METRICAL TREES FOR WORD STRESS ASSIGNMENT

Uwe D. Reichel

Institute of Phonetics and Speech Processing, University of Munich
reichelu@phonetik.uni-muenchen.de

Abstract: An algorithm for word stress assignment in German compounds is introduced. First, a metrical tree is automatically derived from adjacent morpheme cohesion scores which are based on co-occurrence statistics. This tree is used to identify the stressed compound part by applying the compound stress rule of metrical phonology. Then the stressed syllable is identified within this compound part by means of a k-nearest-neighbor classifier using weighted vowel quantity and syllable coda type features. The accuracy of the metrical compound analysis amounted to 83%. Compound stress assignment was successful in 95%, and stress location within multi-syllable word stems in 84% of all cases.

1 Introduction

For languages like German with a highly productive lexical compounding tendency word stress assignment is to be carried out in two steps: the identification of the stressed compound part, and the localisation of the stressed syllable within this part. The first task can be accomplished by adopting concepts from metrical phonology [7]. In this framework the relations of compound parts can be hierarchically represented by means of metrical trees which are labeled by the *compound stress rule CSR* stating:

> **CSR:** Given constituent [AB], *B* is strong *s* if only *B* is further divisible, else *A*.

An example is given in Figure 1. The stressed compound part is identified by tracing the *s* branches to the corresponding leaf. For the constituent decomposition collocations like *Bahn+hof (station)* are not considered to be further divisible and thus do not attract stress in words like *'Haupt+[bahn+hof] (central station)*.
Stress location within a compound part is determined by stress-attracting affixes as in *pass+'abel (acceptable)* and, if there is no such affix, by several stress constraints for simplex (i.e. monomorphemic) word forms. For German the most important constraints are:

- the 3-syllable window constraint stating that stress is located within the last 3 syllables of a word,
- the Final-schwa constraint: if a final syllable is reduced, then the penult is stressed as in *Ta'pete (wallpaper)*, and
- the Closed-penult constraint saying that the closed penult hinders stress to move further left (*Hi'biskus; hibiscus*).

A more detailed presentation of German word stress constraints including numerous exceptions can be found in [5]. Automatic stress assignment using syllable characteristics is carried out

e.g. by means of neural nets [4] or instance-based learning [2]. The decision tree approach of [10] additionally includes morphologic features.
In the following sections the steps of word stress assignment are introduced: the metrical compound decomposition and the instance based assignment of word stress within the stressed compound part.

2 Metrical compound decomposition

The compound analysis to identify the stressed compound part consists of a morphological segmentation and the induction of a metrical tree for this segmentation.

2.1 Morphological segmentation

Our morphological segmentation algorithm for concatenative morphology has first been introduced in [12] and requires:

- a morpheme lexicon $L = \{< x, m >\}$ containing morphemes x and their classes m,
- a specification of morphotactics $t : m \times m \longrightarrow \{0,1\}$ constraining the morpheme class combinations, and
- a specification of the compatibility $c : w \times m \longrightarrow \{0,1\}$ of a word's part of speech label w and the class m of word-final morpheme to avoid erroneous segmentations like $*$*kombi+niere* (*estate kidney* instead of *combine*).

The recursive splitting function $f : s \longrightarrow x+y$ places morpheme boundaries $+$ within strings s if the following constraints are fulfilled:

1. x is in the lexicon, i.e. $\exists m :< x, m >\in L$,
2. y is further divisible or in lexicon, i.e. $f(y)$ holds or $\exists m :< y, m >\in L$,
3. the morpheme class pair for x and the first segment of y does not violate morphotactics: $t(m_x, m_{y_1}) = 1$, and
4. the morpheme class of the last y-segment is compatible with the word's part of speech: $c(w, m_{y_n}) = 1$.

This procedure results in a concatenative flat morphological segmentation $x_1 \ldots x_n$. From this segmentation a compound decomposition $[x_1 \ldots x_i][x_{i+1} \ldots x_n]$ is deduced if x_i is a linking morpheme or if the morpheme classes x_i and x_{i+1} belong to the following sets respectively:

- $m_{x_i} \in \{$LexicalMorph, InflectionEnding, Suffix, OrdinalEnding$\}$,
- $m_{x_{i+1}} \in \{$LexicalMorph, Prefix, Adverbial, VerbalParticle$\}$.

As an example, this way *bund+es+haus+halt+s+aus+schuss (federal budget committee)* is decomposed into [*bund+es*] [*haus*] [*halt+s*] [*aus+schuss*].

2.2 Metrical tree induction

From the flat compound representation derived from the preceeding segmentation a hierarchic representation is derived by recursively splitting the compounds at coherence minima and by pruning the resulting trees in order to merge collocations.

2.2.1 Coherence-based tree creation

The coherence of adjacent compound parts x and y is measured by means of the Likelihood ratio of two hypotheses:

$$\frac{\text{H0: } x \text{ and } y \text{ are independent}}{\text{H1: } x \text{ and } y \text{ are mutually dependent}} \tag{1}$$

This technique has originally been used to extract collocations [3] and will therefore also be helpful for tree pruning as is explained in the following section. It serves to compare the hypothesis-related likelihoods $L(H0)$ and $L(H1)$ for the observed frequencies k for $x+y$ co-occurrence and $N-k$ for the occurrence of y without x given the hypothesised probabilities:

- **H0 (independence):** $P(y|x) = p = P(y|\neg x)$
- **H1 (dependence):** $P(y|x) = p_1 \neq p_2 = P(y|\neg x)$

The likelihoods $L(H0)$ and $L(H1)$ for k and $N-k$ given the probabilities p, respectively p_1 and p_2, are calculated assuming a binomial distribution, and their ratio is transformed into a χ^2 value by taking $-2ln\frac{L(H0)}{L(H1)}$ as is described in [8]. This transformation allows for a gradual interpretation of coherence: the higher χ^2, the higher the dependency between x and y.

Given the coherence values of all adjacent compound parts a coherence tree is induced by recursively splitting the compound at local coherence minima which is schematically shown in Figure 1.

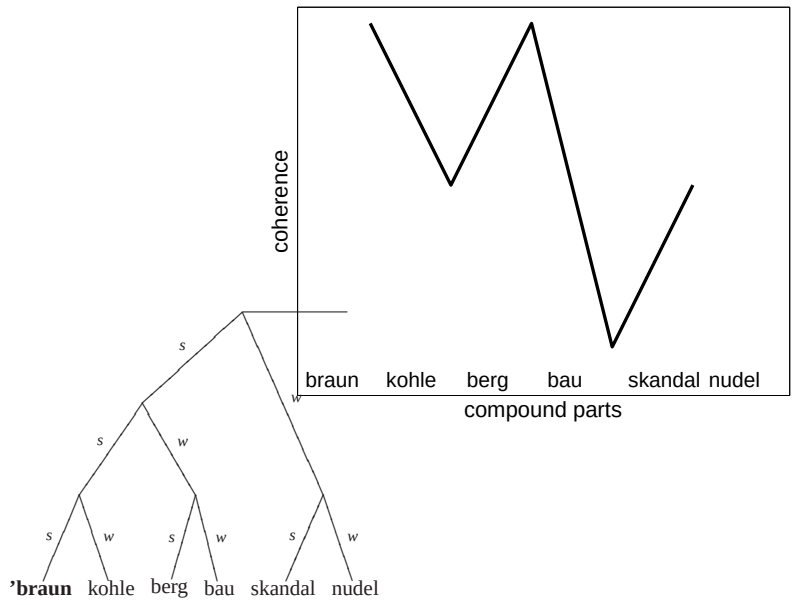

Figure 1 - A metrical tree for the compound [*braun*][*kohle*][*berg*][*bau*][*skandal*][*nudel*] *(brown coal mining sleazebag)* inferred from compound part coherences. **Right:** Coherence values of adjacent compound parts. **Left:** Resulting tree by recursive splitting at local coherence minima and application of the compound stress rule.

2.2.2 Pruning

The pruning of the tree consists in merging adjacent compound parts x and y which are identified as collocations. Generally, collocations are characterised by:

- a high degree of coherence, and
- semi-compositionality, i.e. the meaning of $x+y$ cannot be composed by the meanings of its parts: $||x+y|| \neq ||x|| + ||y||$.

Following [3] the high degree of coherence can simply be expressed by high χ^2 values which have been derived by the preceeding processing step.

For semi-compositionality we propose a distributional measure based on the intuition that the lexical contexts of collocations and their final parts, e.g. of *Bahnhof (station)* and *Hof (yard)* are more distinct than for non-collocative compositions like *Sporttasche (sports bag)* and *Tasche (bag)*, since the latter pair is interchangeable in more contexts than the first. To account for this notion we adopt the information radius *IR* measure, which is already established in measuring semantic similarity [8]. It quantifies the difference between the word probability distribution p in the context of *Bahnhof* as opposed to q in the context of *Hof* as follows:

$$IR(p,q) = D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2}), \text{ where} \quad (2)$$

$$D(p||q) = \sum_i p_i \log_2 \frac{p_i}{q_i}. \quad (3)$$

Here, the context has been defined as the word history in a bigram model trained on a German text corpus. The Relative Entropy $D(p||q)$ gives the number of bits additionally needed to encode events i, for which the distribution p holds, by a code based on q. $IR(p,q)$ is a symmetric version of this divergence measure and thus a proper distance metrics. Following our intuition it is expected that semi-compositionality yields high information radius values.

Indeed, as shown in Figure 2 significant differences for χ^2 and IR values have been found in the expected direction (two-sided Welch tests; for χ^2: $t_{74} = 3.86, \alpha = 0.001$; for *IR*: $t_{318} = 1.83, \alpha = 0.05$. The IR difference is not apparent looking at the boxplots, nevertheless, the mean IR values are 1.83 for collocations and 1.76 for non-collocative parts).

For tree pruning the following thresholds were derived on a small data sample by Simplex optimisation: Adjacent compound parts with an IR > 1.96 and a χ^2 value > 90 are considered as collocations and therefore merged.

2.2.3 From coherence to metrical trees

As shown in Figure 1 the coherence tree is transformed into a metrical tree by assigning s (strong) and w (weak) labels to its branches following the CSR. The stressed compound part is identified by tracing the s branches starting from the tree root.

3 Instance-based learning of word stem stress

Based on the morphological information which is returned by the procedure described in section 2.1, in some cases the stressed syllable can directly be localised within its compound part. Such trivial cases consist of derived word forms containing stress-attracting affixes as *Akzept'anz (acceptance)*, and of one-syllable stems with or without unstressed affixes as *Haus (house)* and *Be+'geh+ung (inspection)*.

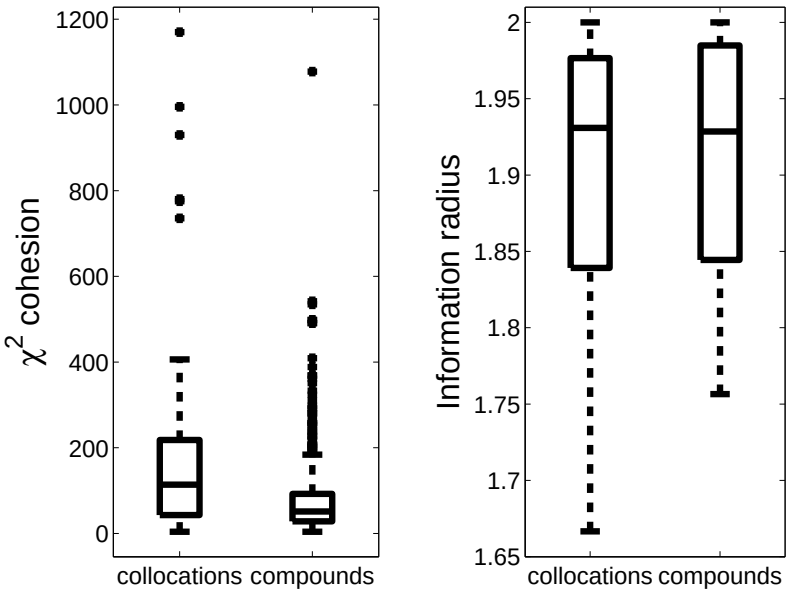

Figure 2 - Left: χ^2 co-occurrence values of collocations (e.g. *Bahn+hof*) and real compound parts (e.g. *Sport+tasche*). **Right:** Information radius values of collocations and real compound parts.

For all other multi-syllable word stems derived by affix stripping, a syllabified canonic transcription is generated by a grapheme-phoneme converter [10], and stress is located within the stem by instance-based learning. Word stems are represented by six features derived from the canonic transcription:

- *vowel quantity* $\in \{$*reduced, short, long,* $0\}$, and
- *coda type* $\in \{$*open, closed,* $0\}$

each extracted for the ultimate (final), penultimate and antepenultimate syllable. 0 is assigned for absent values in words shorter than three syllables. The chosen features capture the stress constraints formulated in section 1. The dependent variable to be predicted from these features is the absolute position of the stressed syllable relative to the ultimate and has one of the following values: $\{0,1,2,3\}$, 0 indicating a stressed ultimate, 1 a stressed penultimate and so on. Our training data did not contain words longer than four syllables. As an example, *Lawine (deluge)* with the transcription [la.'vi:.n@] is stored as the following <feature vector, target> instance in the memory M: $<$ [*short, open, long, open, reduced, open*], **1**$>$.

In application, for an incoming syllabified transcription of a word stem the k nearest neighbors are derived from M and the stress position occuring most often among these k objects is assigned to the input. On a small development set k was set to 15 by Simplex optimisation. Objects were compared by means of the weighted Hamming distance D as follows:

$$D(a,b) \quad = \quad \sum_{a_i \neq b_i} w_i, \tag{4}$$

i ranging over the elements of the objects' feature vectors a and b. w_i is the weight of the underlying feature X, which is set to the mutual information I between X and the dependent word stress position variable Y:

$$w_X \quad = \quad I(Y;X) = H(Y) - H(Y|X). \tag{5}$$

$H(Y)$ and $H(Y|X)$ represent the entropy and conditional entropy of the word stress position, respectively. w_X therefore makes explicit the information gain for predicting word stress given that the value of feature X is known.

4 Results

4.1 Compound level

For a sample of 700 compounds containing more than two parts the accuracy of the hierarchical compound analysis based on the morphological segmentation and the coherence tree induction amounted 83%. Compound stress assignment was successful in 95% of all cases.

The adequacy of the compound stress rule expressed in the conditional probability P(stress correct|compound analysis correct) is 0.96, indicating that for the used data this rule is appropriate.

4.2 Word stem level

As can be seen by the mutual information values in the left plot of Figure 3 vowel quantity is generally more influential for word stress assignment than the presence or absence of the syllable coda. Furthermore, the characteristics of the last syllable is most influential for stress location.

Shown in the right plot of Figure 3, in a 10-fold cross validation task on 1300 non-trivial cases, i.e. multi-syllable word stems, the *k*-nearest-neighbor classifier successfully predicted the stress location in 84% of all cases.

5 Discussion

5.1 Metrical tree induction

In this study a new procedure to automatise the induction of metrical trees has been introduced which is based on a statistical notion of compound part coherence and the well-known compound stress rule from metrical phonology. Initial results of 83% accuracy for tree construction and 95% for stress assignment are encouraging. The CSR turned out to be highly adequate for our data.

Nevertheless, the identification of collocations for tree pruning needs further elaboration. Collocations indeed show significantly higher χ^2 and information radius values as opposed to other compound part pairings, but the discriminative power of these measures is not very high as can be seen by the largely overlapping boxplots in Figure 2, especially for IR. It will be explored in future studies whether a more sophisticated definition of the examined word context will improve the contribution of the IR measure.

Rhythmic constraints as avoiding stress clashes [7] or long sequences of unstressed syllables are not yet implemented in our procedure. The latter constraint, which might favor the pattern *Braunkohlebergbau'skandalnudel* over the predicted *'Braunkohlebergbauskandalnudel* (cf. Figure 1), can be addressed by including the concept of minor stress [5] which can be realised by applying the CSR in parallel for subtrees originating from the tree root, thus applying it separately for *Braunkohlebergbau* and for *Skandalnudel* in the example given above.

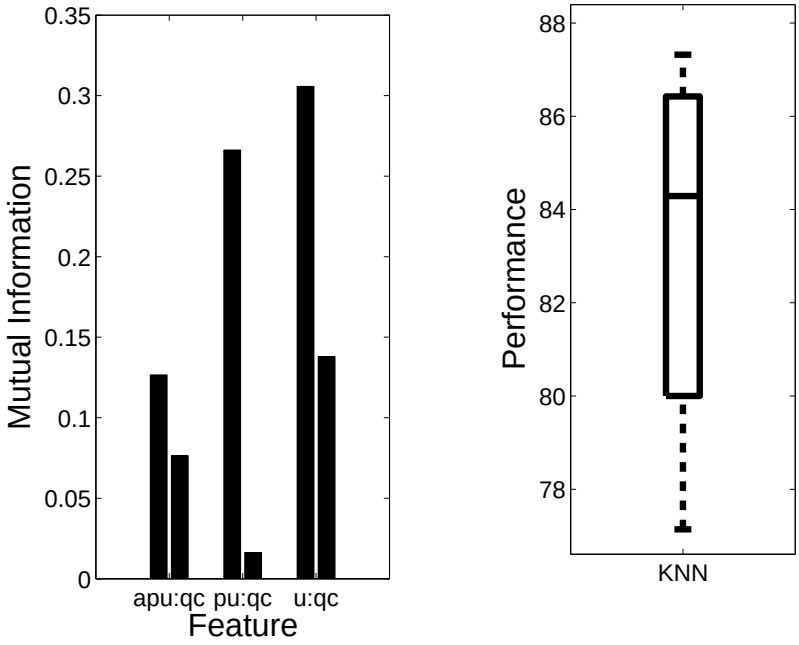

Figure 3 - Left: Mutual information between syllable features and word stress position. apu – antepenult, pu – penult, u – ult, q – vowel quantity, c – coda type. **Right:** 10-fold cross-validation of the KNN classifier.

Another challenge not addressed yet arises from extrinsic factors for stress shift like contrast constructions as *Arbeit'nehmer und Arbeit'geber (employees and employers)* as opposed to *'Arbeitgeber*.

5.2 Instance-based word stem stress assignment

In contrast to previous work of the author [11] the entity chosen for stress assignment for the current machine learning approach is the word stem and not the syllable. This difference also implies different target values to be predicted, namely the syllable index instead of the binary distinction *stressed* vs. *unstressed*. Since in contrast to compounds simplex word forms cannot be arbitrarily long, the set of target values is still finite, and for our data limited to integers from 0 to 3. As opposed to the strictly syllable-based approach in [11] the global word pattern can be taken into consideration for stress localisation.

Among the issues not addressed by the current approach are stress shifts (*'Doktor vs. Dok'toren*) and homographs like *durchlaufen* in transitive (*durch'laufen*) vs. intransitive (*'durchlaufen*) usage.

Even when leaving aside these cases, the word stress constraints listed in section 1 are not sufficient to describe all stress patterns. Furthermore, numerous contradicting cases can be found (see e.g. [5]), as for example *'Abenteuer (adventure)* violating the 3-syllable window constraint. Therefore, a machine learning approach like instance-based learning is expected to be more robust than rule-based algorithms.

In addition, due to intrinsic equivalences the chosen approach of instance-based learning can easily be linked to fundamental research frameworks as Exemplar Theory [9, 6]. This relation allows for generating hypotheses and models for topics like second language acquisition about how speakers might stress unknown words of a foreign language.

Weighting schemes like the mutual information between features and stress location as pro-

posed here, give insight in the relative contribution of features and could be adopted for the fine-tuning of Exemplar Theory models.

6 Acknowledgments

The work of the author has been carried out within the CLARIN-D project [1] (BMBF-funded).

References

[1] *http://eu.clarin-d.de/index.php/en/*. Clarin-D web page.

[2] DAELEMANS, W., S. GILLIS and G. DURIEUX: *The Acquisition of Stress, a data-oriented approach*. Computational Linguistics, 20(3):421–451, 1994.

[3] DUNNING, T.: *Accurate methods for the statistics of surprise and coincidence*. Computational Linguistics, 19:61–74, 1993.

[4] GUPTA, P. and D. TOURETZKY: *Connectionist Models and Linguistic Theory: Investigations of Stress Systems in Language*. Cognitive Science, 18(1):1–50, 1994.

[5] JESSEN, M.: *A survey of German word stress*. In *AIMS*, vol. 2, pp. 115–139. University of Stuttgart, 1995.

[6] JOHNSON, K.: *Speech perception without speaker normalization: An exemplar model*. In JOHNSON, K. and J. W. MULLENNIX (eds.): *Talker Variability in Speech Processing*, pp. 145–166. Academic Press, San Diego, 1997.

[7] LIBERMAN, M. and A. PRINCE: *On Stress and Linguistic Rhythm*. Linguistic Inquiry, 8:249–336, 1977.

[8] MANNING, C. and H. SCHÜTZE: *Foundations of statistical natural language processing*. MIT, Cambridge, Massachusetts, 2001.

[9] NOSOFSKY, R.: *Exemplar-based accounts of relations between classification, recognition, and typicality*. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(4):700–708, 1988.

[10] REICHEL, U.: *PermA and Balloon: Tools for string alignment and text processing*. In *Proc. Interspeech*, Portland, Oregon, to appear in 2012.

[11] REICHEL, U. and F. SCHIEL: *Using Morphology and Phoneme History to improve Grapheme-to-Phoneme Conversion*. In *Proc. Eurospeech*, pp. 1937–1940, Lisboa, 2005.

[12] REICHEL, U. and K. WEILHAMMER: *Automated Morphological Segmentation and Evaluation*. In *Proc. 4th Language Resources & Evaluation Conference*, pp. 503–506, Lisbon, Portugal, 2004.