Morphological and Phonological Learning: Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON), Philadelphia, July 2002, pp. 70-80. Association for Computational Linguistics.

Probabilistic Context-Free Grammars for Phonology

Karin Müller
Department of Computational Linguistics
University of Saarland, Germany
kmueller@coli.uni-sb.de

Abstract

We present a phonological probabilistic context-free grammar, which describes the word and syllable structure of German words. The grammar is trained on a large corpus by a simple supervised method, and evaluated on a syllabification task achieving 96.88% word accuracy on word tokens, and 90.33% on word types. We added rules for English phonemes to the grammar, and trained the enriched grammar on an English corpus. Both grammars are evaluated qualitatively showing that probabilistic context-free grammars can contribute linguistic knowledge to phonology. Our formal approach is multilingual, while the training data is language-dependent.

1 Introduction

In this paper, we present an approach to supervised learning and automatic detection of syllable structure. The primary goal of the paper is to show that probabilistic context-free grammars can be used to gain substantial phonological knowledge about syllable structure. Beyond an evaluation of the trained model on a real-world task documenting the performance of the model, we focus on an extensive qualitative evaluation.

In contrast to other approaches which work with syllable structures extracted from a pronunciation dictionary, our approach focuses on the probability of use of certain syllable structures. Among other approaches that deal with syllable structure, there are example-based approaches (Hall (1992), Wiese (1996), Féry (1995), Kenstowicz (1994), Morelli (1999)), symbolic approaches (Belz, 2000), connectionist phonotactic models (Stoianov and Nerbonne, 1998), stochastic models describing partial structures (Pierrehumbert (1994), Coleman and Pierrehumbert (1997)), or application-based approaches for syllabification (Van den Bosch, 1997) or text-to-speech systems (Kiraz and Möbius, 1998).

Our method builds on two resources. The first one is a large written text corpus, which is looked-up in a pronunciation dictionary resulting in a large transcribed and syllabified corpus. The second resource is a manually written context-free grammar describing German and English syllable structure. We code the assumptions (similar to Goldsmith (1995)) that the phonological material that can occur in the onsets or codas might differ depending on the syllable positions: word-initial, word-final, word-medial, versus monosyllabic words.

We train the context-free grammar for German on the transcribed and syllabified training corpus with a simple supervised training method (Müller, 2001a). The main idea of the training method is that after a grammar transformation step, the grammar together with a parser can predict syllable boundaries of unknown phoneme strings. The trained model is evaluated on a syllabification task showing a high precision on a test corpus. We exemplify that the method can be easily transferred to related languages (here English) by adding rules for missing phonemes to the grammar. In an qualitative evaluation, we compare German and English syllable structure by interpreting the probability weights of the preterminal

grammar rules.

In sum, we aim to show that our method (i) models all possible words of a language, (ii) models how likely certain structures are used (in comparison to pure dictionary-based approaches), (iii) yields good results in an application-oriented evaluation, (iv) is able to disambiguate competing structures, (v) can be easily applied to other languages, (vi) produces mathematically well-defined models.

The paper is organized as follows. We present our method in Section 2, the experiments in Section 3, and our evaluation in Section 4. In Section 5, we discuss the results, and in Section 6, we conclude.

2 Method

We build on the novel approach of Müller (2001a) which aims to combine the advantages of treebank and bracketed corpora training. In general, this approach consists of four steps: (i) writing a (symbolic i.e. non-probabilistic) context-free phonological grammar with syllable boundaries, (ii) training this grammar on a large automatically transcribed and syllabified corpus, (iii) transforming the resulting probabilistic phonological grammar by dropping the syllable boundaries, and (iv) predicting syllable boundaries of unseen phoneme strings by choosing their most probable phonological tree according to the transformed probabilistic grammar.

The advantages of this approach are, that simple and efficient supervised training on bracketed corpora can be used (the brackets guarantee that all syllabified words of the training corpus receive only one single analysis), and that raw phoneme strings can be parsed and syllabified after the grammar transformation.

Preserving these advantages, our approach differs in several important details. First, we write a more advanced phonological grammar for German, yielding a more fine-grained probabilistic model of syllable structure. Second, it is easily possible to enrich our phonological grammar by adding grammar rules for missing phonemes to adapt our phonological grammar to other languages (here English). Third, in addition to an evaluation on a real-world task (syllabification for German), we qualitatively evaluate the resulting probabilistic versions of our phonological grammar for German and English.

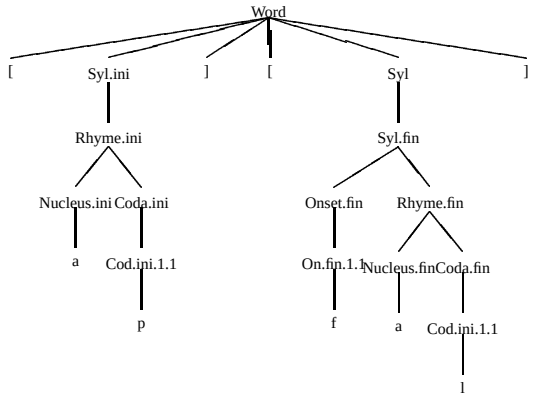


Figure 1: Syllable structure of the word "Abfall" ([ap][fal]) according to our phonological grammar for German.

Our phonological grammar divides a word into syllables, which are in turn rewritten by onset, nucleus, and coda. Furthermore, the phonological grammar differentiates between monosyllabic and polysyllabic words. In polysyllabic words, the syllables are divided into syllables appearing word-initially, word-medially, and word-finally. Additionally, the grammar distinguishes between consonant clusters of different sizes (ranging from one to five consonants), as well as between consonants occurring in different positions within a cluster. Figure 1 displays the structure of the German word "Abfall" (*waste*) according to our phonological grammar.
In the following sections, we especially focus on the rewriting rules involving phonemic terminal nodes: $\mathbf{X}.r.s.t \rightarrow C$ and $\mathbf{Y}.r \rightarrow V$. The rules of the first type bear three of the above mentioned features for a consonant C inside an onset or a coda (**X**=On, Cod), namely: the position of the syllable in the word (r=ini, med, fin, one), the cluster size ($t = 1 \ldots 5$), and the position of a consonant within a cluster ($s = 1 \ldots 5$). Obviously, vowels or diphthongs V of a nucleus (**Y**=Nucleus) do not need the position and size features (t and s). The probabilities of these phonological rules (after supervised training) are exactly the basis for our description and evaluation of the syllable parts in Section 4.

3 Experiments

In the following, we describe two experiments. The first experiment investigates syllable structure for German. In the second experiment, we generalize and apply the method to another language (English).

Experiment with German data.

First, we manually write a phonological grammar for German consisting of 2,394 context-free rules. If compared to the most successful grammar constructed by Müller (2001a), our grammar is enriched with an additional feature: the size of the onsets and codas. Second, we extracted a training corpus of 2,127,798 words (3,961,982 syllables) from a German newspaper, the Stuttgarter Zeitung, and an additional corpus of 242,047 words for testing. All words are looked up in the German part of the CELEX (Baayen et al., 1993) yielding transcribed and syllabified corpora. As phoneme set, we used the symbols from the English and German SAMPA alphabet (Wells, 1997). In contrast to Müller (2001a), we did not investigate smaller training corpora, since we are interested in maximal phonological knowledge about internal word structure. Third, we train the phonological context-free grammar on the training corpus using the supervised method presented in Section 2. Additionally, due to events not occurring in the training data, we use the implemented smoothing procedure of the LoPar system (Schmid, 2000) producing rules with positive probabilities.

Experiment with English data.

In this experiment, we show that our method can be easily applied to other languages. We create a second training corpus of the same size of 2,123,081 words from the British National Corpus. The words are looked-up in the English part of the CELEX. Furthermore, the context-free grammar is extended by rules for all possible English phonemes. This means, we add preterminal rules for phonemes not occurring in German words, e.g., rules for the apicodental phoneme [T] (appearing in the word *this*). This (semi-automatic) procedure yields an English phonological grammar consisting of 4,418 rules, which is trained on the new corpus.

4 Evaluation

First, we evaluate on a syllabification task for German. Second, we analyze linguistically the errors made by the German phonological parser on the evaluation corpus (word types). Third, and more important, we concentrate on a qualitative evaluation of syllable structure for German and English.

4.1 Evaluation on Syllabification

The resulting probabilistic phonological grammar of German (Section 3) is evaluated on a syllabification task by comparing the maximum-probability parses of all raw phoneme strings in the test corpus (242,047 word tokens, 24,735 word types) with their annotated bracketed variants. As evaluation measure, we used "word accuracy" which computes the rate of words with all predicted syllable brackets exactly matching the annotated syllable brackets. The evaluation shows that our phonological grammar for German achieves 96.88% word accuracy on word tokens, and 90.33% on word types.

Error Analysis

We analyze the results of the German phonological parser on the evaluation corpus consisting of word types. Out of 24,735 words types 2391 words contained wrongly predicted syllable boundaries. We analyze every tenth word of the incorrect words, which means we look at 239 items. There are 243 errors found in the analyzed words. Due to the fact that there can occur more than one error in one word, the number of errors is higher than the number of items. We find that 72.42% of the errors are made for consonants uncorrectly assigned to the onset, whereas only 27.57% errors are made for consonants wrongly assigned to the coda. The tendency that more errors occur in predicting the onset agrees with the findings of Müller (2001b). The main errors appear when a [t], [R], or an [n] is predicted to be an onset consonant. The main errors found in predicting the coda consonants mainly occur with [k], [s], and [p].

If we investigate the errors made on the linguistic level, the most frequent error occurs with word boundaries (98 cases). However, a further error source is found with syllable boundaries occuring in conjunction with prefixes and suffixes. The most frequent error appears with syllable boundaries after prefixes like /ver-/, /er-/, and /un-/, whereas with suffixes the most frequent error appears with /-lich/. Foreign words, which are subject to different phonotactic constraints, seem to be a minor source of errors (20 cases). Thus, we can see that most of the errors are found in conjunction with morphological

English	initial	medial	final	monosyl	sum
0C onset	191,888	17,630	32,350	453,117	694,985
1C onset	407,380	246,802	556,954	852,866	2,064,002
2C onset	88,692	58,888	96,604	122,295	366,479
3C onset	3,332	3,577	5,368	4,511	16,788
4C onset	-	1	16	-	17
sum	691,292	326,898	691,292	1,432,789	3,142,271
0C coda	485,968	257,595	122,058	435,247	1,300,868
1C coda	196,409	65,594	440,176	777,347	1,479,526
2C coda	8,872	3,706	112,597	206,941	332,116
3C coda	42	3	16,361	13,014	29,420
4C coda	1	-	100	240	341
5C coda	-	-	-	-	-
sum	691,292	326,898	691,292	1,432,789	3,142,271

German	initial	medial	final	monosyl	sum
0C onset	294,428	54,388	50,508	273,407	672,731
1C onset	658,986	577,358	963,600	687,429	2,887,373
2C onset	127,632	103,353	75,721	68,050	374,756
3C onset	10,879	7,160	2,096	6,987	27,122
4C onset	-	-	-	-	-
sum	1,091,925	742,259	1,091,925	1,035,873	3,961,982
0C coda	633,196	489,576	236,905	170,845	1,530,522
1C coda	404,616	223,590	708,651	645,447	1,982,304
2C coda	48,399	26,807	130,712	200,472	406,390
3C coda	4,988	2,255	14,945	16,028	38,216
4C coda	726	31	712	3,078	4,547
5C coda	-	-	-	3	3
sum	1,091,925	742,259	1,091,925	1,035,873	3,961,982

Table 1: **Frequency of occurrence of onsets and codas** for English (left) and German (right), displayed for different complexities (ranging from 0 to 5 consonants) and different positions of the syllable in the word (initial, medial, final, monosyl).

English	word initial	word medial	word final	monosyl
3:	0.024	0.013	0.009	0.014
9	<0.001	-	-	-
A:	0.030	0.012	0.005	0.017
E	0.126	0.105	0.030	0.044
E@	0.008	0.002	0.003	0.010
I	0.166	0.282	0.367	0.171
I@	0.009	0.010	0.019	0.003
O	0.072	0.026	0.006	0.082
O:	0.028	0.021	0.014	0.041
OI	0.002	0.004	0.001	0.004
U	0.028	0.020	0.005	0.015
U@	<0.001	0.003	0.001	<0.001
V	0.069	0.021	0.008	0.029
Y	<0.001	-	-	-
&	0.070	0.032	0.008	0.095
@	0.167	0.268	0.386	0.220
@U	0.037	0.020	0.022	0.029
aI	0.036	0.030	0.022	0.042
aU	0.015	0.003	0.014	0.014
eI	0.039	0.077	0.031	0.069
i:	0.045	0.017	0.027	0.029
u:	0.018	0.024	0.012	0.064

German	word initial	word medial	word final	monosyl
2:	0.008	0.008	0.002	<0.001
9	0.010	0.003	<0.001	<0.001
@	0.082	0.180	0.693	-
A∼:	0.001	0.091	<0.001	-
E	0.131	0.015	0.023	0.061
E:	0.011	0.129	0.005	0.001
I	0.064	0.030	0.049	0.163
O	0.051	-	0.008	0.052
OI	-	0.008	<0.001	-
OY	0.020	-	0.001	0.001
O∼	-	<0.001	-	-
O∼:	<0.001	<0.001	<0.001	-
U	0.047	0.033	0.044	0.082
Y	0.019	0.012	0.002	0.002
a	0.107	0.092	0.036	0.103
a:	0.066	0.056	0.032	0.037
aI	0.105	0.048	0.026	0.060
aU	0.031	0.009	0.008	0.051
e:	0.075	0.054	0.010	0.171
eI	<0.001	-	-	-
i:	0.054	0.135	0.022	0.124
o:	0.061	0.052	0.018	0.021
u:	0.026	0.027	0.007	0.039
y:	0.023	0.010	0.003	0.021

Table 2: **Nuclei** for English (left) and German (right).

entities. This might point out that a further morphological level could help to disambiguate syllabification alternatives.

4.2 Qualitative Evaluation

The evaluation is carried out for both English and German. First, we compare the complexity of words, syllables, and syllable parts. Second, we analyze the probabilities of grammar rules involving phonemic terminal nodes. Unfortunately, due to space constraints and the large size of our derived probabilistic phonological grammars, only preliminary results can be presented.

Table 1 displays the occurrence frequencies of onsets and codas (of different size and different syllable position), counted on the basis of the training corpora for English and German (see Section 3). The following three complexity analyses are carried out on the basis of this table.

Word complexity. German words tend to be more complex than English words. The German training corpus comprises 48.7% [1] monosyllabic words, whereas 67.4% are found in the English training corpus. The high frequency of occurrence of monosyllabic words justifies the separate treatment of those words.

Syllable complexity. *German syllables* usually consist of onset and rhyme. An onset is observed in initial syllables (73%) [2], in medial syllables (94%), in final syllables (94%), and in monosyllabic words (73%). A coda is found in initial syllables (42%), in medial syllables (34%), in final syllables (78%), and

[1] (1,035,873 / 2,127,798)=48.7%

[2] ((658,986 + 127,632 + 10,879) / 1,091,925) = 73%

English	initial	medial	final	monosyl
D	0.002	0.002	0.022	0.250
N	<0.001	<0.001	<0.001	-
S	0.006	0.017	0.073	0.019
T	0.011	0.006	0.007	0.008
Z	<0.001	0.004	0.010	-
b	0.068	0.025	0.042	0.072
d	0.084	0.059	0.075	0.024
f	0.057	0.052	0.029	0.055
g	0.018	0.016	0.018	0.017
h	0.048	0.004	0.003	0.063
j	0.019	0.001	0.003	0.032
k	0.113	0.046	0.052	0.032
l	0.048	0.097	0.095	0.034
m	0.087	0.078	0.057	0.045
n	0.042	0.091	0.054	0.035
p	0.080	0.065	0.037	0.021
r	0.098	0.080	0.065	0.020
s	0.114	0.097	0.067	0.052
t	0.033	0.125	0.185	0.105
v	0.026	0.080	0.052	0.003
w	0.034	0.002	0.012	0.105
z	0.003	0.041	0.031	<0.001

German	initial	medial	final	monosyl
N	-	0.003	0.021	-
R	0.040	0.074	0.082	0.009
S	0.011	0.024	0.033	0.004
Z	<0.001	0.001	<0.001	-
b	0.096	0.056	0.062	0.031
d	0.073	0.072	0.106	0.454
f	0.115	0.052	0.029	0.099
g	0.098	0.103	0.065	0.016
h	0.062	0.031	0.016	0.013
j	0.030	0.005	0.001	0.011
k	0.064	0.032	0.023	0.015
l	0.042	0.121	0.076	0.011
m	0.078	0.055	0.042	0.072
n	0.037	0.078	0.120	0.076
p	0.024	0.018	0.006	0.002
s	-	0.017	0.030	-
t	0.026	0.116	0.184	0.007
v	0.115	0.055	0.019	0.064
x	<0.001	0.009	0.036	-
z	0.078	0.069	0.039	0.107

Table 3: **Onsets consisting of 1 consonant** for English (left) and German (right)

English	initial	medial	final	monosyl
D	0.010	<0.001	<0.001	0.019
N	0.043	0.018	0.138	0.006
S	0.011	0.012	0.007	0.002
T	0.005	<0.001	<0.001	0.007
Z	-	<0.001	<0.001	<0.001
b	0.019	<0.001	<0.001	0.002
d	0.022	0.008	0.106	0.068
f	0.037	0.001	0.003	0.102
g	0.022	0.008	<0.001	0.003
h	-	-	-	<0.001
k	0.118	0.176	0.020	0.029
l	0.074	0.105	0.119	0.046
m	0.111	0.066	0.023	0.052
n	0.435	0.426	0.166	0.168
p	0.014	0.039	0.005	0.015
r*	<0.001	<0.001	0.178	0.123
s	0.026	0.080	0.046	0.038
t	0.031	0.035	0.056	0.162
v	0.005	0.016	0.016	0.026
x	<0.001	-	<0.001	<0.001
z	0.009	0.002	0.108	0.124

German	initial	medial	final	monosyl
N	0.011	0.014	0.055	0.002
R	0.356	0.315	0.183	0.271
S	<0.001	<0.001	0.004	0.001
b	-	-	-	<0.001
f	0.028	0.020	0.005	0.036
k	0.038	0.046	0.021	0.009
l	0.067	0.082	0.026	0.026
m	0.030	0.020	0.038	0.050
n	0.269	0.287	0.511	0.273
p	0.031	0.019	0.001	0.008
s	0.084	0.091	0.047	0.156
t	0.040	0.018	0.055	0.064
v	<0.001	-	-	-
x	0.041	0.082	0.047	0.096

Table 4: **Codas consisting of 1 consonant** for English (left) and German (right).

in monosyllabic words (83%).

English syllables. An onset is observed in initial syllables (72.2%), in medial syllables (94%), in final syllables (95%), and in monosyllabic words (68.3%). A coda is found in initial syllables (29.7%), in medial syllables (21.2%), in final syllables (82.3%), and in monosyllabic words (69%).

Onset and coda complexity. English and German syllables prefer simple onsets and codas.

English onsets. For both initial and medial syllables, a single consonant is found (80%), two consonants (18%), and three consonants (less than 1%). For both final syllables and monosyllabic words, one consonant is observed (85%), two consonants (13%), and three consonants (less than 0.9%).

German onsets. For both initial and medial syllables, one consonant is found (82%), two consonants (15%), and three consonants (1%). For both monosyllabic words and final syllables, a single consonant is found (90%), two consonants (7%), and three consonants (less than 1%).

English codas. For both initial and medial syllables, one consonant is observed (95%), and two consonants (5%). For both final syllables and monosyllabic words, one consonant is found (77%), two consonants (20%), and three consonants (2%).

German codas. For both initial and medial syllables, one consonant is observed (88%), two consonants

(10%), and three consonants (about 1%). In final syllables, one consonant is found (82.8%), two consonants (15.2%), three consonants (1.7%), and four consonants (0.08%). In monosyllabic words, one consonant occurs (74.6%), two consonants (23.1%), three consonants (1.8%), and four consonants (0.3%).

Nuclei. Table 2 displays the nuclei found for English and German. The symbol "-" indicates that the phoneme has not been observed in the training corpus. However, the corresponding grammar rules receive a (very small) positive probability according to our smoothing procedure. Note, we do not display those phonemes in the tables which are marked with "-" for all positions . Moreover, the symbol "< 0.001" indicates a probability of less than 0.001 (which means a occurrence frequency of less than 0.1%).

English nuclei. The most likely nuclei in initial syllables are [@, I, E, O, &, V] (16.7%, 16.6%, 12.6%, 7.2%, 7%, 6.9%), in medial syllables [I, @, E, eI] (28.2%, 26.8%, 10.5%, 7.7%), in final syllables [@, I] (38.6%, 36.7%), and in monosyllabic words [@, I,&, O, eI, u:] (22%, 17.1%, 9.5%, 8.2%, 6.9%, 6.4%). Furthermore, in monosyllabic words, 33.6% of the nuclei are long vowels/diphthongs, 29.1% in initial syllables, 23.6% in medial syllables, and 18% in final syllables.

German nuclei. In initial syllables, the most probable nuclei are [E, a, aI, @, e:, a:, I, o:, i:, O,] (13.%, 10.7%, 10.5%, 8.2%, 7.5%, 6.6%, 6.4%, 6.1%, 5.4%, 5.1%), in medial syllables, [@, i:, I, a, E, a:, e:, o:] (18%, 13.5%, 12.9%,9.2%, 9.1%, 5.6%, 5.4%, 5.2%), in final syllables [@, I, U, a] (69.3%, 4.9%, 4.4%, 3.6%), and in monosyllabic words [e:, I, i:, a, U, E, aI, O, aU] (17.1%, 16.3%, 12.4%, 10.3%, 8.2%, 6.1%, 6%, 5.2%, 5.1%). Generally, we can observe that long vowels/diphthongs are more likely in monosyllabic words (52.6%) than in all other syllable positions (48.3% in initial syllables, 42% in medial syllables, and 13.4% in final syllables).

Mono-consonantal onsets and codas. Table 3 and Table 4 display the onsets and codas consisting of 1 consonant.

German onsets. The most probable consonants in initial syllables are [f, v, g, b,z, m, d, k, h], (11.5%, 11.5%, 9.8%, 9.6%, 7.8%, 7.8%, 7.3%, 6.4%, 6.2%), in medial syllables [l, t, g, n, R, d, z] (12.1%, 11.6%, 10.3%, 7.8%, 7.4%, 7.2%, 6.9%), in final syllables [t, n, d, R, l] (18.4%, 12%, 10.6%, 8.2%, 7.6%), and in monosyllabic words [d, z, f, n, m] (45.4%, 10.7%, 9.9%, 7%).

English onsets. In initial syllables, the most probable consonants are [s, k, r, m, d, p] (11.4%, 11.3%, 9.8%, 8.7%, 8.4%, 8%), in medial syllables [t, s, l, n, r, v] (12.5%, 9.8%, 9.7%, 9.1%, 8%, 8%), in final syllables [t, l, d, S, s, r] (18.5%, 9.5%, 7.5%, 7.3%, 6.7%, 6.5%), and in monosyllabic words [D, t, w, b, h] (25%, 10%, 10%, 6%, 6%).

German codas. In initial position, the most likely consonants are [R, n] (35.6%, 26.9%), in medial syllables [R, n] (31.5%, 28.7%), in final syllables [n, R] (50.1%, 18.3%), and in monosyllabic words [n, R, s, x] (27.3%, 27.1%, 15.6%, 9.6%).

English codas. In initial syllables, the most dominant consonants are [n, k, m] (43.5%, 11.8%, 11.1%), in medial syllables [n, k, l] (42.6%, 17.6%, 10.5%), in final syllables [r*, n, N, l, z, d] (17,8%-10%), and in monosyllabic words [n, t, z, r*, f] (16.8%- 11%).

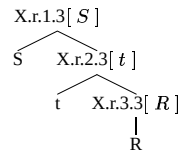
Onset and coda clusters. Clusters of 2-3 consonants are displayed and described in Table 5 and Table 7 (onsets), and in Table 6 and Table 8 (codas). Due to space constraints, we omitted to display clusters of 4-5 consonants, but our analysis can be found elsewhere (Müller, to appear 2002). Clusters with more than 5 consonants have not been found in our corpora. Furthermore, for German, no onsets comprising 4 consonants, and for English, no codas occur comprising 5 consonants. Last, for German, there is only one consonant cluster [Rnsts] appearing in words like "Ernsts", the genitive case of the proper name "Ernst".

5 Discussion

As Müller (2001a), we presented a method for prediction of syllable boundaries using phonological probabilistic context-free grammars for German. However, our approach performs slightly better (96.88% word accuracy on word tokens versus 96.49%). Van den Bosch (1997) reports a word error rate of 2.22% on English syllabification using inductive learning. Due to the feature "cluster size", which was not used by Müller (2001a), we are able to give an extensive qualitative evalu-

ation of syllable structure considering syllable positions, as well as the complexities of consonant clusters, and the position of a consonant within a cluster. Since our approach is multilingual (only training is language dependent), we evaluated two languages (German and English) showing that probabilistic context-free grammars add linguistic knowledge to phonology. In contrast to theoretical approaches, we focus on syllable structures that are preferred in a certain language. Theoretical phonotactic approaches like (Hall (1992), Wiese (1996), Féry (1995)) describe possible syllable structures or German, and Kenstowicz (1994), Morelli (1999) for English. There are many more approaches dealing with syllable structure. For instance, Kiraz and Möbius (1998) develop multilingual syllabification models on the basis of a pronunciation dictionary. Partial English syllable structure is described by Pierrehumbert (1994), who also used a dictionary. A more general model was introduced by Müller et al. (2000), who used a clustering algorithm to induce English and German syllable classes. However, the approach treats the onsets and codas as one string. In our method, we describe in more detail the internal structure of onsets and codas. Our model has several advantages. (i) We believe that the syllable structure of all words occurring in a certain language can be described by an elaborated context-free grammar. (ii) Moreover, using probabilistic context-free grammars, alternative syllabifications of phoneme strings can be disambiguated. (iii) Our model is able to analyze unexpected phoneme strings. For instance, the onset [bRd] of the proper name "Brdaric" ([bRda:RItS]) is not allowed according to German phonotactics. Table 7 correctly displays that neither [b] occurs as a first, nor [R] as a second, or [d] as a third consonant in initial triconsonantal German onset clusters. Due to the smoothing procedure, our model syllabifies this name as [bRda:][RItS] although the onset [bRd] has never been observed in the German training corpus. (iv) The syllable structure of nonsense words can be predicted. The model can be exploited in two ways: first, it predicts the most probable syllabification, second, it can be used to model the lexical decision task. An example of English words is mentioned by Pierrehumbert (1994). She compared "bistro" ([bIstr@U]) as a possible word, with a good word "bimplo" ([bImpl@U]), and a bad word "bilflo" ([bIlfl@U]). The four possible syllabifications are [bIm][pl@U], [bI][mpl@U], [bImp][l@U], and [bImpl][@U]. The most probable syllable structure for "bimplo" is [bIm][pl@U] (1.8403e-13). For the twosyllabic word "bilflo", the model assigns the highest probability to a syllable boundary between [l] and [fl]. Out of the four possible syllabifications for the real word "bistro" [bIstr][@U] is the most probable syllable structure. Although the triconsonantal cluster should be rather an onset cluster than a coda cluster, the model prefers [str] as a cluster, whereas a syllable boundary is added between [m] and [pl], and [l] and [fl]. A further example mentioned in the literature is [brIk], [blIk], and [bnIk]. The first one is a possible word of English [brIk], which receives a probability of 1.16e-08, the second non-occurring word [blIk] (7.2391e-09), and the non-English word [bnIk] (4.3249e-09). The least probable one is the non-English word, followed by the non-occurring one, and the highest probability is assigned to the real word brick.

Beside the good performance of the current models in applications, further improvements of the present approach can possibly be achieved by embedding more prior phonotactic knowledge. For example, it might be useful to model the distribution of a consonant dependent on the previous one (Menzel, 2001). In future work, we will investigate this issue by using head-lexicalized probabilistic context-free grammars, like those suggested by Carroll and Rooth (1998), where the consonant cluster [StR] would be analyzed as:



Here, the lexical choice events express the desired phonotactic feature. Moreover, it would be interesting to incorporate the stress feature.

Our linguistic evaluation of the errors point out that most errors of the phonological parser occur in conjunction with morphological phenomena like prefixes, suffixes and word boundaries. This might point out that a further morphological layer could improve word accuracy (see also Meng (2001)).

6 Conclusion

We introduced a multilingual approach to probabilistic modeling of syllable structure using probabilistic context-free grammars. We exemplified our approach for two languages, German and English. Evaluation on a syllabification task shows a small improvement in word accuracy rate compared to other state-of-the-art systems for German. Additionally, we presented an extensive qualitative evaluation of German and English syllable, onset, and coda structure (section 4) showing which structures are preferably used. However, the presented work is a starting point of analyzing in detail the huge amount of data with regard to phonological regularities. For instance, we found evidence that in consonant clusters sonorous consonants (like [R], [l]) are more restricted to appear next to the nucleus than less sonorous consonants. Clearly, this is promising work for future investigations. We believe that our method can be easily transferred to a variety of other languages, and aim in future work at embedding more fine-grained phonotactic constraints to our grammar.

References

Harald R. Baayen, Richard Piepenbrock, and H. van Rijn. 1993. The CELEX lexical database—Dutch, English, German. LDC. Univ. of Pennsylvania.

Anja Belz. 2000. Multi-syllable phonotactic modelling. In *Proc. of SIGPHON*.

Glenn Carroll and Mats Rooth. 1998. Valence induction with a head-lexicalized pcfg. In *Proc. of EMNLP*.

John Coleman and Janet Pierrehumbert. 1997. Stochastic Phonological Grammars and Acceptability. In *Proc. of SIGPHON*.

Carolin Féry. 1995. Alignment, syllable and metrical Structure in German. SFS-report, Univ. of Tübingen.

John A. Goldsmith. 1995. Phonological Theory. In *Handbook of Phonological Theory*.

Tracy Hall. 1992. *Syllable structure and syllable related processes in German*. Niemeyer, Tübingen.

M. Kenstowicz. 1994. *Phonology in Generative Grammar*. Blackwell, Cambridge, MA.

George Anton Kiraz and Bernd Möbius. 1998. Multilingual Syllabification Using Weighted Finite-State Transducers. In *Proc. 3rd ESCA Workshop on Speech Synthesis (Jenolan Caves)*, pages 59–64.

Helen Meng. 2001. A hierarchical lexical representation for bi-directional spelling-to-pronunciation/pronunciation-to-spelling generation. *Speech Communication*, 33:213–239.

Wolfgang Menzel. 2001. Personal communication.

Frida Morelli. 1999. *The Phonotactics and Phonology of Obstruent Clusters in Optimality Theory*. Ph.D. thesis, University of Maryland.

Karin Müller, Bernd Möbius, and Detlef Prescher. 2000. Inducing Probabilistic Syllable Classes Using Multivariate Clustering. In *Proc. of ACL*.

Karin Müller. 2001a. Automatic Detection of Syllable Boundaries Combining the Advantages of Treebank and Bracketed Corpora Training. In *Proc. of ACL*.

Karin Müller. 2001b. Probabilistic Context-Free Grammars for Syllabification and Grapheme-to-Phoneme Conversion. In *Proc. of EMNLP*, Pittsburgh, PA.

Karin Müller. to appear 2002. *Probabilistic Syllable Modeling Using Supervised and Unsupervised Learning Methods*. Ph.D. thesis, University of Stuttgart.

Janet Pierrehumbert. 1994. Syllable structure and word structure. In *Phonological Structure and Phonetic Form*. University Press, Cambridge.

Helmut Schmid. 2000. LoPar. Design and Implementation. IMS, University of Stuttgart.

Ivelin Stoianov and John Nerbonne. 1998. Exploring Phonotactics with Simple Recurrent Networks. In *Proc. of Comp. Linguistics in the Netherlands*.

Antal Van den Bosch. 1997. *Learning to Pronounce Written Words: A Study in Inductive Language Learning*. Ph.D. thesis, Univ. Maastricht, Maastricht, The Netherlands.

John Wells. 1997. Spoken language reference materials. In *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin.

Richard Wiese. 1996. *The Phonology of German*. Clarendon Press, Oxford.

English	initial		medial		final		monosyl	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd
D	-	-	<0.001	-	<0.001	-	-	-
S	<0.001	0.059	0.010	0.038	0.030	0.087	0.001	0.041
T	0.007	-	0.001	-	<0.001	-	0.053	-
Z	-	0.080	<0.001	0.088	0.003	0.115	-	0.066
b	0.055	-	0.036	-	0.062	-	0.044	-
d	0.106	-	0.119	-	0.216	0.001	0.093	-
f	0.044	<0.001	0.025	-	0.018	0.001	0.132	<0.001
g	0.036	-	0.051	-	0.037	-	0.050	-
h	0.025	-	0.001	-	0.007	-	0.192	-
j	-	0.071	<0.001	0.258	<0.001	0.112	-	0.058
k	0.104	0.016	0.164	0.013	0.061	0.011	0.067	0.025
l	<0.001	0.094	0.023	0.097	0.018	0.113	<0.001	0.126
m	0.013	0.005	0.031	<0.001	0.002	<0.001	<0.001	0.016
n	0.009	0.001	0.022	0.004	0.030	0.023	0.031	0.002
p	0.273	0.045	0.138	0.078	0.058	0.028	0.070	0.024
r	-	0.466	<0.001	0.249	<0.001	0.309	-	0.341
s	0.160	0.001	0.216	-	0.203	<0.001	0.168	<0.001
t	0.157	0.068	0.131	0.116	0.212	0.149	0.085	0.079
v	0.001	-	0.021	-	0.032	-	0.006	-
w	-	0.084	-	0.053	-	0.045	-	0.216
z	<0.001	0.002	0.003	-	0.001	-	-	-

German	initial		medial		final		monosyl	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd
N	-	-	<0.001	-	<0.001	-	-	-
R	-	0.428	-	0.257	-	0.166	-	0.225
S	0.261	0.001	0.229	<0.001	0.153	0.040	0.218	-
Z	-	<0.001	-	<0.001	-	0.002	-	0.004
b	0.045	-	0.055	-	0.058	-	0.053	-
d	0.018	-	0.029	-	0.036	-	0.054	-
f	0.101	0.003	0.034	0.008	0.016	0.027	0.043	0.002
g	0.094	-	0.056	<0.001	0.054	-	0.038	-
j	-	<0.001	-	0.009	-	0.003	-	<0.001
k	0.087	0.003	0.070	<0.001	0.038	-	0.063	<0.001
l	-	0.126	-	0.138	-	0.129	-	0.105
m	<0.001	0.003	<0.001	0.002	0.005	0.002	-	0.002
n	-	0.009	-	0.020	-	0.023	-	0.019
p	0.155	0.046	0.044	0.014	0.053	0.011	0.035	0.020
s	0.005	0.191	0.019	0.367	0.116	0.366	<0.001	0.457
t	0.228	0.143	0.453	0.169	0.463	0.220	0.492	0.143
v	-	0.041	0.007	0.012	0.003	0.006	<0.001	0.019
x	-	-	<0.001	-	<0.001	-	-	-
z	-	-	<0.001	<0.001	<0.001	-	-	-

Table 5: **Onsets consisting of 2 consonants** for English (top) and German (bottom)

English: In *initial* syllables, the most dominant first consonants are [p, s, t, d, k] (27.3%, 16%, 15.7%, 10%, 10%), the second consonants are [r, l, w, Z, j] (46.6%, 9.4%, 8.4%, 8%, 7.1%). In *medial* syllables, the most likely first consonants are [s, k, p, t, d] (21.6%, 16.4%, 13.8%, 11.9%), and the second consonants [j, r, t, l, Z] (25.8%, 24.9%, 11.6%, 9.7%, 8.8%). In *final* syllables, the most probable first consonants are [d, t, s], (21.6%, 21.2%, 20.3%), and the most probable second consonants are [r, t, Z, l, j, S] (30.9%, 14.9%, 11.5%, 11.3%, 11.2%, 8.7%). *Monosyllabic* words favor on the first position [h, s, f, d, t, p], (19.2%, 16.8%, 13.2%, 9.3%, 8.5%, 7%), and on the second position [R, w, l, t] (34.1%, 21.6%, 12.6%, 7.9%).
The consonants which only appear as first consonants are [S,b,d,g,k], and as second consonants [R,Z,j,l,n]. The consonants which are restricted to the first position are less sonorous than the consonants restricted to the second position. This observation corresponds with the assumption that the sonority increases towards the nucleus.

German: In *initial* syllables, the first consonants are [S, t, p, f, g, k] (26.1%, 22.8%, 15.5%, 10%, 9.4%, 8.7% with descending probability), and the second consonants are [R, s, t, l] (42.8%, 19%, 14%, 12.6%). In *medial* syllables, the first consonants are [t,S] (45.3%, 22.9%), and the second consonant are [s, R, t, l] (36.7%, 25.7%, 16.9%, 13.8%). In *final* syllables, the first consonants are [t, S, s] (46.3%, 15.3%, 11.6%), and the second consonants are [s, t, R, l] (36.6%, 22%, 16.6%, 12.9%). In *monosyllabic* words, the most probable first consonants are [t, S] (49.2%, 21.8%), and the second consonants are [s, R, t, l] (45.7%, 22.5%, 14.3%, 10.5%).
Summarizing, the most probable first consonants are [S, t, p, f, g, h], and the most probable second ones are [s, R, t, l].

English	initial		medial		final		monosyl	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd
D	-	-	-	-	<0.001	-	<0.001	-
N	0.044	-	0.011	-	0.016	-	0.025	-
S	-	0.033	-	0.069	0.008	0.007	0.002	0.087
T	-	0.006	-	0.001	<0.001	0.005	<0.001	0.007
Z	-	0.036	-	0.224	-	0.041	<0.001	0.023
b	0.057	-	<0.001	-	0.005	<0.001	0.003	<0.001
d	0.035	0.107	0.161	0.042	0.057	0.172	0.035	0.400
f	0.028	0.011	-	0.010	0.003	0.017	0.010	0.002
g	<0.001	0.007	-	-	0.001	-	0.004	-
h	-	-	-	-	-	-	-	<0.001
k	0.271	0.036	0.029	0.009	0.071	<0.001	0.061	0.025
l	0.072	-	0.026	-	0.101	-	0.087	-
m	0.137	-	0.011	-	0.038	<0.001	0.032	0.001
n	0.266	-	0.678	-	0.501	-	0.429	-
p	<0.001	0.140	0.008	<0.001	0.017	<0.001	0.022	0.006
r	<0.001	-	-	-	-	-	-	-
s	0.050	0.393	0.028	0.026	0.083	0.149	0.110	0.141
t	0.033	0.124	0.042	0.584	0.052	0.379	0.147	0.201
v	<0.001	-	-	0.025	0.024	0.001	0.011	0.002
w	-	-	-	-	-	-	<0.001	-
x	-	-	-	-	-	-	<0.001	-
z	-	0.100	-	0.003	0.016	0.223	0.013	0.100

German	initial		medial		final		monosyl	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd
N	0.036	-	0.130	-	0.023	-	0.009	-
R	0.205	-	0.145	-	0.292	-	0.131	-
S	-	0.046	-	0.003	0.002	0.001	<0.001	0.004
b	-	-	-	-	-	-	<0.001	-
f	0.019	0.058	0.063	0.019	0.049	0.004	0.011	0.017
k	0.035	0.077	0.013	0.093	0.059	0.014	0.027	0.036
l	0.051	<0.001	0.079	<0.001	0.147	-	0.118	-
m	0.019	0.005	0.009	0.003	0.025	0.003	0.012	0.006
n	0.378	0.019	0.336	0.012	0.203	0.125	0.405	0.003
p	0.022	0.009	0.011	0.004	0.019	0.016	0.020	0.001
s	0.065	0.181	0.009	0.357	0.031	0.166	0.118	0.105
t	0.125	0.557	0.150	0.487	0.049	0.663	0.017	0.801
x	0.039	0.043	0.050	0.018	0.093	0.002	0.125	0.023

Table 6: **Codas consisting of 2 consonants** for English (top) and German (bottom).

English: In *initial* syllables, the most likely first consonants are [k, n, m] (27%, 26.6%, 13.7%), and the second consonants are [s, p, t, d, z] (39.3%, 14%, 12.4%, 10.7%, 10%). In *medial* syllables, the most prominent first consonants are [n, d] (67,8%, 16.1%), and the second consonants are [t, Z] (58,4%, 22.4%). In *final* syllables, the most dominant first consonants are [n, l, s] (50%, 10%, 8%), and the most dominant second ones are [t,z,d,s] (37.9%, 22.3%, 17.2%, 14.9%). In *monosyllabic* words, the most likely first consonants are [n, t, s, l] (42.9%, 14.7%, 11%, 8.7%), and the second consonants are [d, t, s, z, S] (40%, 20%, 14%, 10%, 8.7%). Generally, [n] is the most dominant first consonant, whereas [t] is the dominant second consonant in medial, and final syllables.

German: In *initial* syllables, the most likely first consonants are [n, R, t] (37.8%, 20.5%, 12.5%), and the second consonants are [t, s] (55.7%, 18.1%). In *medial* syllables, the first consonants are [n, t, R, N], (33.6%, 15%, 14.5%, 13%), and the second consonants are [t, s] (48.7%, 35.7%). In *final* syllables, the first consonants are mainly [R, n, l, x] (29.2%, 20%, 14.7%, 12.9%), and the second ones are [t, s, n] (66%, 16.6%, 12.5%). In *monosyllabic* words, [n, R, x, l, s] are the most probable first consonants (40.5%, 13.1%, 12.5%, 11.8%, 11.8%), and the second consonants are [t, s] (80%,10.5%).
In general, the first position is more variable than the second one. Despite this fact, the most dominant first consonant is [n], whereas the second position is mainly restricted to [t,s].

English	initial			medial			final			monosyl		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
S	-	-	-	-	0.008	-	-	0.030	-	-	-	-
Z	-	-	-	0.004	<0.001	-	0.011	0.060	-	-	-	-
d	-	-	-	0.001	-	-	0.062	-	-	-	-	-
f	-	-	-	0.012	-	-	0.055	-	-	-	-	-
g	-	-	-	0.006	-	-	0.005	-	-	-	-	-
j	-	-	0.318	-	0.015	0.039	-	0.049	0.096	-	-	0.004
k	0.001	0.111	-	0.001	0.206	-	0.040	0.160	-	-	0.176	-
l	-	-	0.022	-	0.012	0.209	-	0.089	0.161	-	-	0.031
n	-	-	-	-	-	0.004	0.022	-	0.031	-	-	-
p	-	0.053	-	-	0.248	-	-	0.253	-	-	0.109	-
r	-	0.001	0.614	-	<0.001	0.707	-	0.001	0.603	-	-	0.868
s	0.995	-	-	0.959	-	-	0.766	-	-	0.997	-	-
t	-	0.830	-	0.011	0.503	-	0.034	0.352	-	-	0.710	-
w	-	-	0.040	-	-	0.036	-	<0.001	0.105	-	-	0.093

German	initial			medial			final			monosyl		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
R	-	-	0.447	-	-	0.729	-	-	0.918	-	-	0.285
S	0.446	-	-	0.626	-	-	0.879	-	-	0.285	-	-
f	-	0.040	-	-	0.056	-	-	0.040	-	-	0.016	-
j	-	-	<0.001	-	-	-	-	-	-	-	-	-
k	-	0.004	-	-	<0.001	-	-	0.006	-	-	0.001	-
l	-	-	0.043	-	-	0.059	-	-	0.045	-	-	0.016
p	0.040	0.166	-	0.056	0.424	-	0.040	0.657	-	0.016	0.121	-
s	0.035	0.476	0.030	0.139	0.175	0.034	0.039	0.036	<0.001	0.001	0.696	-
t	0.476	0.310	-	0.175	0.341	-	0.036	0.256	-	0.696	0.164	-
v	-	-	0.476	-	-	0.175	-	-	0.031	-	-	0.697

Table 7: **Onsets consisting of 3 consonants** for English (top) and German (bottom)

For English, the consonants occurring as third consonants are more sonorous than all other ones ([r,j,l,w]), which also applies for German ([v,R]). The sonority decreases in the direction to the syllable edge, which corresponds to the sonority sequencing principle.

English	initial			medial			final			monosyl		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
N	0.113	-	-	-	-	-	0.001	-	-	0.097	-	-
S	-	-	-	-	-	-	-	0.012	<0.001	-	0.082	<0.001
T	-	-	0.018	-	-	-	-	<0.001	<0.001	-	0.054	0.038
Z	-	-	-	-	-	-	-	0.063	<0.001	-	0.067	<0.001
b	-	-	-	-	-	-	-	-	-	<0.001	<0.001	-
d	-	0.056	-	-	-	-	0.053	0.064	0.098	0.041	0.279	0.071
f	0.528	-	-	0.222	-	-	<0.001	0.001	-	0.024	<0.001	-
k	-	0.113	-	-	-	-	0.112	0.001	-	0.051	0.170	-
l	0.018	-	-	-	-	-	0.106	-	<0.001	0.076	<0.001	-
m	-	0.018	-	-	-	-	0.008	0.001	-	0.029	0.009	-
n	0.132	-	-	0.555	-	-	0.589	-	-	0.453	-	<0.001
p	-	-	-	-	-	-	0.010	<0.001	-	0.003	0.054	-
s	-	-	0.691	-	0.222	0.555	0.104	0.160	0.624	0.154	0.057	0.352
t	-	0.603	-	-	0.555	0.222	0.012	0.623	0.174	0.066	0.214	0.242
v	-	-	-	-	-	-	-	0.070	-	-	0.007	-
z	-	-	0.074	-	-	-	-	<0.001	0.100	-	<0.001	0.293

German	initial			medial			final			monosyl		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
N	0.018	-	-	0.002	-	-	0.066	-	-	0.065	-	-
R	0.344	-	-	0.182	-	-	0.203	-	-	0.286	-	-
S	-	-	<0.001	-	-	-	-	0.013	-	-	0.015	<0.001
f	0.005	0.013	0.034	0.237	0.004	0.017	0.004	0.093	0.012	<0.001	0.018	0.018
k	<0.001	0.089	-	-	0.020	-	0.012	0.152	-	0.016	0.104	-
l	0.087	-	0.002	0.055	-	-	0.035	<0.001	-	0.039	<0.001	<0.001
m	0.073	-	-	0.017	-	-	0.019	0.003	-	0.026	0.007	-
n	0.322	-	-	0.246	<0.001	-	0.288	0.049	-	0.193	0.007	-
p	0.004	0.034	-	-	0.029	-	0.016	0.018	-	0.006	0.027	-
s	-	0.358	0.568	-	0.010	0.964	<0.001	0.366	0.328	0.001	0.485	0.350
t	0.038	0.493	0.392	-	0.930	0.014	0.220	0.300	0.658	0.257	0.329	0.629
x	0.102	0.008	-	0.255	<0.001	-	0.129	<0.001	-	0.105	0.003	-

Table 8: **Codas consisting of 3 consonants** for English (top) and German (bottom).

It is remarkable that in English initial and medial syllables, triconsonantal coda clusters are avoided. A similar tendency can be observed for German. Moreover, sonorous consonants rather occur next to the nucleus (for English [n,N,l], for German [R,n]).