Improving Syllabification Models with Phonotactic Knowledge

Karin Müller
Institute of Phonetic Sciences
University of Amsterdam
kmueller@science.uva.nl

Abstract

We report on a series of experiments with probabilistic context-free grammars predicting English and German syllable structure. The treebank-trained grammars are evaluated on a syllabification task. The grammar used by Müller (2002) serves as point of comparison. As she evaluates the grammar only for German, we reimplement the grammar and experiment with additional phonotactic features. Using bi-grams within the syllable, we can model the dependency from the previous consonant in the onset and coda. A 10-fold cross validation procedure shows that syllabification can be improved by incorporating this type of phonotactic knowledge. Compared to the grammar of Müller (2002), syllable boundary accuracy increases from 95.8% to 97.2% for English, and from 95.9% to 97.2% for German. Moreover, our experiments with different syllable structures point out that there are dependencies between the onset on the nucleus for German but not for English. The analysis of one of our phonotactic grammars shows that interesting phonotactic constraints are learned. For instance, unvoiced consonants are the most likely first consonants and liquids and glides are preferred as second consonants in two-consonantal onsets.

1 Introduction

In language technology applications, unknown words are a continuous problem. Especially, Text-to-speech (TTS) systems like those described in Sproat (1998) depend on the correct pronunciation of those words. Most of these systems use large pronunciation dictionaries to overcome this problem. However, the lexicons are finite and every natural language has productive word formation processes. Thus, a TTS system needs a module which converts letters to sounds and a second module which syllabifies these sound sequences. The syllabification information is important to assign the stress status of the syllable, to calculate the phone duration (Van Santen et al. (1997)), and to apply phonological rules (Kahn (1976), Blevins (1995)). Many automatic syllabification methods have been suggested e.g., (Daelemans and van den Bosch, 1992; Van den Bosch, 1997; Kiraz and Möbius, 1998; Vroomen et al., 1998; Müller, 2001; Marchand et al., to appear 2006). Müller (2001) shows that incorporating syllable structure improves the prediction of syllable boundaries. The syllabification accuracy increases if the onset and coda is more fine-grained (Müller, 2002). However, she only incorporates partial phonotactic knowledge in her approach. For instance, her models cannot express that the phoneme /l/ is more likely to occur after an /s/ than after a /t/ in English. The information that a phoneme is very probable in a certain position (here, the /l/ appears as second consonant in a two-consonantal onset cluster) will not suffice to express English phonotactics of an entire consonant cluster. Moreover, she

Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology at HLT-NAACL 2006, pages 11–20,
New York City, USA, June 2006. ©2006 Association for Computational Linguistics

only reports the performance of the German grammar. Thus, we are interested if the detection of syllable boundaries can be improved for both English and German by adding further phonotactic knowledge to a grammar.

Phonotactic constraints within the onset or coda seem to be important for various tasks. Listeners indeed use phonotactic knowledge from their mother language in various listening situations. Vitevitch and Luce (1999), e.g., showed if English speakers have to rate nonsense words how “English-like” the stimuli are, highly probable phonotactic stimuli were rated more “English-like” than stimuli with a lower probability. Speakers make also use of their phonotactic knowledge when they have to segment a sequence into words. In a words spotting task, Weber and Cutler (2006) found evidence that speakers of American English can segment words much easier when the sequence contains phonotactic constraints of their own language.

Beside many perception experiments which show that phonotactic constraints are useful information, many different methods have been suggested to model phonotactic constraints for language technology applications. Krenn (1997), for instance, uses Hidden Markov Models to tag syllable structure. The model decides whether a phoneme belongs to the onset, nucleus or coda. However, this model does not incorporate fine-grained phonotactics. Belz (2000) uses finite state automatons (FSA) to model phonotactic structure of different syllable types. We use similar positional features of syllables. Moreover, Carson-Berndsen (1998) and Carson-Berndsen et al. (2004) focus on automatically acquiring feature-based phonotactics by induction of automata which can be used in speech recognition. In our approach, we concentrate on explicit phonotactic grammars as we want to test different suggestions about the internal structure of words from phonological approaches (e.g. Kessler and Treiman (1997)). We assume, for instance, that codas depend on the previous nucleus and that onsets depend on the subsequent nucleus.

In this paper, we present experiments on a series of context-free grammars which integrate step by step more phonological structure. The paper is organized as follows: we first introduce our grammar development approach. In section 3, we describe our experiments and the evaluation procedure. The subsequent section 4 shows what kind of phonotactic information can be learned from a phonotactic grammar. Last, we discuss our results and draw some conclusions.

2 Method

We build on the approach of Müller (2001) which combines the advantages of treebank and bracketed corpora training. Her method 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 pronunciation dictionary which contains markers for syllable boundaries (see Example 1; the pre-terminals “X[” and “X]” denote the beginning and end of a syllable such that syllables like [strIN] can be unambiguously processed during training), (iii) transforming the resulting probabilistic phonological grammar by dropping the syllable boundary markers[1] (see Example 2), and (iv) predicting syllable boundaries of unseen phoneme strings by choosing their most probable phonological tree according to the transformed probabilistic grammar. The syllable boundaries can be extracted from the *Syl* node which governs a whole syllable.

(1) Word → X[Syl_{one}]X
(2) Word → Syl_{one}

We use a grammar development procedure to describe the phonological structure of words. We expect that a more fine-grained grammar increases the precision of the prediction of syllable boundaries as more phonotactic information can be learned. In the following section, we describe the development of a series of grammars.

2.1 Grammar development

Our point of comparison is (i) the *syllable complexity grammar* which was introduced by Müller (2002). We develop four different grammars: (ii) the *phonotactic grammar*, (iii) the *phonotactic on-nuc grammar* (iv) the *phonotactic nuc-coda grammar* and (v) the *phonotactic on-nuc-coda grammar*. All five grammars share the following features: The grammars describe a word which is composed of one

[1] We also drop rules with zero probabilities

to n syllables which in turn branch into onset and rhyme. The rhyme is re-written by the nucleus and the coda. Onset or coda could be empty. Furthermore, all grammar versions differentiate between monosyllabic and polysyllabic words. In polysyllabic words, the syllables are divided into syllables appearing word-initially, word-medially, and word-finally. Additionally, the grammars distinguish between consonant clusters of different sizes (ranging from one to five consonants).

We assume that phonotactic knowledge within the onset and coda can help to solve a syllabification task. Hence, we change the rules of the *syllable complexity grammar* (Müller, 2002) such that phonotactic dependencies are modeled. We express the dependencies within the onset and coda as well as the dependency from the nucleus by bi-grams.

2.1.1 Grammar generation

The grammars are generated automatically (using perl-scripts). As all possible phonemes in a language are known, our grammar generates all possible re-write rules. This generation process naturally over-generates, which means that we receive rules which will never occur in a language. There are, for instance, rules which describe the impossible English onset /tRS/. However, our training procedure and our training data make sure that only those rules will be chosen which occur in a certain language.

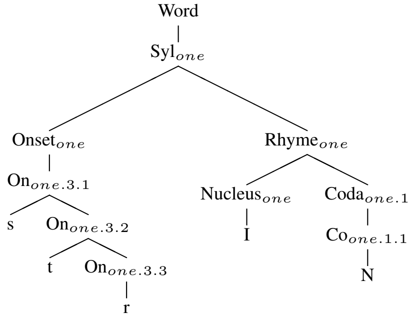
The monosyllabic English word *string* is used as a running example to demonstrate the differences of the grammar versions. The word *string* is transcribed in the pronunciation dictionary CELEX as ([strIN]) (Baayen et al., 1993). The opening square bracket, "[", indicates the beginning of the syllable and the closing bracket, "]", the end of the syllable. The word consists of the tri-consonantal onset $[str]$ followed by the nucleus, the short vowel $[I]$ and the coda $[N]$.

In the following paragraphs, we will introduce the different grammar versions. For comparison reasons, we briefly describe the grammar of Müller (2002) first.

2.1.2 Syllable complexity grammar (Müller, 2002)

The syllable complexity grammar distinguishes between onsets and codas which contain a different number of consonants. There are different rules which describe zero to n-consonantal onsets. Tree (3) shows the complete analysis of the word *string*.

(3)



(4) $\text{On}_{one.3.1} \rightarrow \text{s} \quad \text{On}_{one.3.2}$

(5) $\text{On}_{one.2.1} \rightarrow \text{s} \quad \text{On}_{one.2.2}$

Rule 4, e.g., describes a tri-consonantal onset, e.g., $[str]$. This rule occurs in example tree 3 and will be used for words such as ***str**ing* or ***spr**ay*. Rule (5) describes a two-consonantal onset occurring in the analysis of words such as ***sn**ake* or ***st**and*. However, this grammar cannot model phonotactic dependencies from the previous consonant.

2.1.3 Phonotactic grammar

Thus, we develop a phonotactic grammar which differs from the previous one. Now, a consonant in the onset or coda depends on the preceding one. The rules express bi-grams of the onset and coda consonants. The main difference to the previous grammars can be seen in the re-writing rules involving phonemic preterminal nodes (rule 6) as well as terminal nodes for consonants (rule 7).

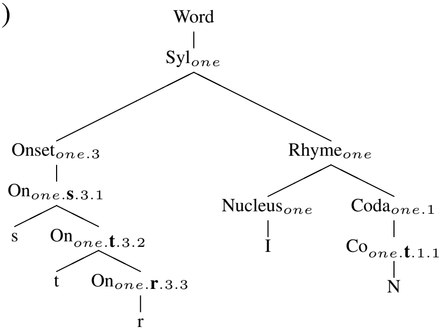
(6) $\mathbf{X}\text{.r.C.s.t} \rightarrow \text{C} \quad \mathbf{X}\text{.r.C}^{+}\text{.s.t}$

(7) $\mathbf{X}\text{.r.C.s.t} \rightarrow \text{C}$

Rules of this type bear four 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 current terminal node ($C = consonant$), the succeeding consonant (C^{+}), the cluster size ($t = 1 \ldots 5$), and the position of a consonant within a cluster ($s = 1 \ldots 5$).

The example tree (8) shows the analysis of the word *string* with the current grammar version. The rule (9) comes from the example tree showing that the onset consonant $[t]$ depends on the previous consonant $[s]$.

(8)

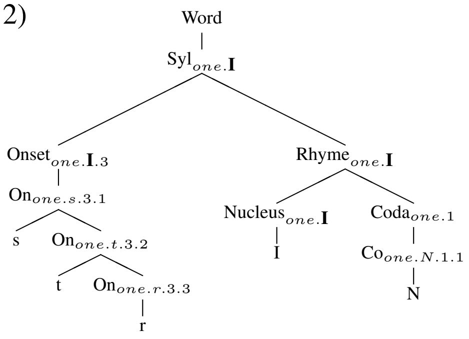

Word
Syl$_{one}$
Onset$_{one.3}$
Rhyme$_{one}$
On$_{one.\mathbf{s}.3.1}$
s
On$_{one.\mathbf{t}.3.2}$
t
On$_{one.\mathbf{r}.3.3}$
r
Nucleus$_{one}$
I
Coda$_{one.1}$
Co$_{one.\mathbf{t}.1.1}$
N

(9) $\text{On}_{one.s.3.1} \rightarrow \text{s} \quad \text{On}_{one.t.3.2}$

2.1.4 Phonotactic on-nuc grammar

We also examine if there are dependencies of the first onset consonant on the succeeding nucleus. The dependency of the whole onset on the nucleus is indirectly encoded by the bi-grams within the onset. The phonotactic onset-nucleus grammar distinguishes between same onsets with different nuclei. In example tree (12), the triconsonantal onset starting with a phoneme $[s]$ depends on the Nucleus $[I]$. Rule (10) occurs in tree (12) and will be also used for words such as *strict* or *strip* whereas rule (11) is used for words such as *strong* or *strop*.

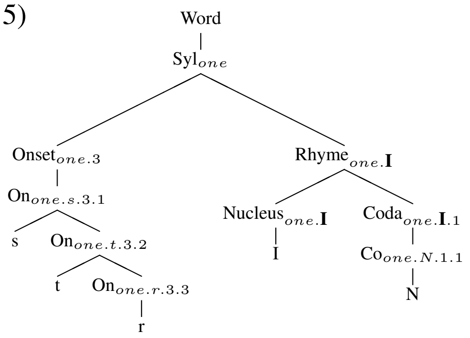
(10) $\text{Onset}_{one.\mathbf{I}.3} \rightarrow \text{On}_{one.s.3.1}$
(11) $\text{Onset}_{one.\mathbf{O}.3} \rightarrow \text{On}_{one.s.3.1}$
(12)



2.1.5 Phonotactic nuc-coda grammar

The phonotactic nucleus-coda grammar encodes the dependency of the first coda consonant on the nucleus. The grammar distinguishes between codas that occur with various nuclei. Rule 13 is used, for instance, to analyze the word *string*, shown in Example tree 15. The same rule will be applied for words such as *bring, king, ring* or *thing*. If there is a different nucleus, we get a different set of rules. Rule 14, e.g., is required to analyze words such as *long, song, strong* or *gong*.

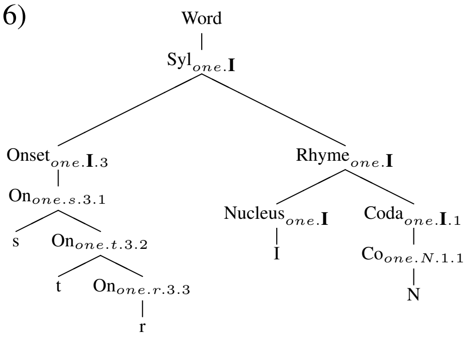
(13) $\text{Coda}_{one.\mathbf{I}.1} \rightarrow \quad \text{N Co}_{one.t.1.1}$
(14) $\text{Coda}_{one.\mathbf{O}.1} \rightarrow \quad \text{N Co}_{one.t.1.1}$
(15)



2.1.6 Phonotactic on-nuc-coda grammar

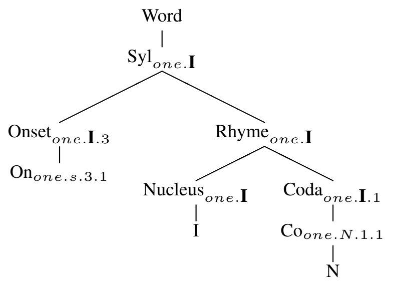
The last tested grammar is the phonotactic onset-nucleus-coda grammar. It is a combination of grammar 2.1.4 and 2.1.5. In this grammar, the first consonant of the onset and coda depend on the nucleus. Tree 16 shows the full analysis of our running example word *string*.

(16)



The rules of the subtree (17) are the same for words such as *string* or *spring*. However, words with a different nucleus such as *strong* will be analyzed with a different set of rules.

(17)


Word
$\text{Syl}_{one.\mathbf{I}}$
$\text{Onset}_{one.\mathbf{I}.3}$
$\text{Rhyme}_{one.\mathbf{I}}$
$\text{On}_{one.s.3.1}$
$\text{Nucleus}_{one.\mathbf{I}}$
$\text{Coda}_{one.\mathbf{I}.1}$
I
$\text{Co}_{one.N.1.1}$
N

3 Experiments

In this section, we report on our experiments with four different phonotactic grammars introduced in Section 2.1 (see grammar 2.1.3-2.1.6), as well as with a re-implementation of Müller's less complex grammar (Müller, 2002). All these grammars are trained on a corpus of transcribed words from the pronunciation lexicon CELEX. We use the full forms of the lexicon instead of the lemmas. The German lexicon contains 304,928 words and the English lexicon 71,493 words. Homographs with the same pronunciation but with different part of speech tags are taken only once. We use for our German experiments 274,435 words for training and 30,492 for testing (evaluating). For our English experiments, we use 64,343 for training and 7,249 for testing.

3.1 Training procedure

We use the same training procedure as Müller (2001). It is a kind of treebank training where we obtain a probabilistic context-free grammar (PCFG) by observing how often each rule was used in the training corpus. The brackets of the input guarantee an unambiguous analysis of each word. Thus, the formula of treebank training given by (Charniak, 1996) is applied: r is a rule, let $|r|$ be the number of times r occurred in the parsed corpus and $\lambda(r)$ be the non-terminal that r expands, then the probability assigned to r is given by

$$p(r) = \frac{|r|}{\sum_{r' \in \{r' | \lambda(r') = \lambda(r)\}} |r'|}$$

After training, we transform the PCFG by dropping the brackets in the rules resulting in an analysis grammar. The bracket-less analysis grammar is used for parsing the input without brackets; i.e., the phoneme strings are parsed and the syllable boundaries are extracted from the most probable parse.

In our experiments, we use the same technique. The advantage of this training method is that we learn the distribution of the grammar which maximizes the likelihood of the corpus.

3.2 Evaluation procedure

We evaluate our grammars on a syllabification task which means that we use the trained grammars to predict the syllable boundaries of an unseen corpus. As we drop the explicit markers for syllable boundaries, the grammar can be used to predict the boundaries of arbitrary phoneme sequences. The boundaries can be extracted from the *syl*-span which governs an entire syllable.

Our training and evaluation procedure is a 10-fold cross validation procedure. We divide the original (German/English) corpus into ten parts equal in size. We start the procedure by training on parts 1-9 and evaluating on part 10. In a next step, we take parts 1-8 and 10 and evaluate on part 9. Then, we evaluate on corpus 8 and so forth. In the end, this procedure yields evaluation results for all 10 parts of the original corpus. Finally, we calculate the average mean of all evaluation results.

3.2.1 Evaluation Metrics

Our three evaluation measures are word accuracy, syllable accuracy and syllable boundary accuracy. Word accuracy is a very strict measure and does not depend on the number of syllables within a word. If a word is correctly analyzed the accuracy increases. We define **word accuracy** as

$$\frac{\#\ \ of\ \ correctly\ \ analyzed\ \ words}{total\ \ \#\ \ of\ \ words}$$

Syllable accuracy is defined as

$$\frac{\#\ \ of\ \ correctly\ \ analyzed\ \ syllables}{total\ \ \#\ \ of\ \ syllables}$$

The last evaluation metrics we used is the **syllable boundary accuracy**. It expresses how reliable the boundaries were recognized. It is defined as

$$\frac{\#\ \ of\ \ correctly\ \ analyzed\ \ syllable\ boundaries}{total\ \ \#\ \ of\ \ syllable\ boundaries}$$

The difference between the three metrics can be seen in the following example. Let our evaluation corpus consist of two words, *transferring* and *wet*. The transcription and the syllable boundaries are displayed in table 1. Let our trained grammar predict the boundaries shown in table 2. Then the word accuracy will be 50%

transferring	trA:ns–f3:–rIN
wet	wEt

Table 1: Example: evaluation corpus

transferring	trA:n–sf3:–rIN
wet	wEt

Table 2: Example: predicted boundaries

$(\frac{1\ correct\ word}{2\ words})$, the syllable accuracy will be 50% $(\frac{2\ correct\ syllables}{4\ syllables})$, and the syllable boundary accuracy is 75% $(\frac{3\ correct\ syllable\ boundaries}{4\ syllable\ boundaries})$. The difference between syllable accuracy and syllable boundary accuracy is that the first metric punishes the wrong prediction of a syllable boundary twice as the complete syllable has to be correct. The syllable boundary accuracy only judges the end of the syllable and counts how often it is correct. Mono-syllabic words are also included in this measure. They serve as a baseline as the syllable boundary will be always correct. If we compare the baseline for English and German (tables 3 and 4, respectively), we observe that the English dictionary contains 10.3% monosyllabic words and the German one 1.59%.

Table 3 and table 4 show that phonotactic knowledge improves the prediction of syllable boundaries. The syllable boundary accuracy increases from 95.84% to 97.15% for English and from 95.9% to 96.48% for German. One difference between the two languages is if we encode the nucleus in the onset or coda rules, German can profit from this information compared to English. This might point at a dependence of German onsets from the nucleus. For English, it is even the case that the on-nuc and the nuc-cod grammars worsen the results compared to the phonotactic base grammar. Only the combination of the two grammars (the on-nuc-coda grammar) achieves a higher accuracy than the phonotactic grammar. We suspect that the on-nuc-coda grammar encodes that onset and coda constrain each other on the repetition of liquids or nasals between /s/C onsets and codas. For instance, *lull* and *mam* are okey, whereas *slull* and *smame* are less good.

4 Learning phonotactics from PCFGs

We want to demonstrate in this section that our phonotactic grammars does not only improve syllabification accuracy but can be used to reveal interesting phonotactic[2] information at the same time. Our intension is to show that it is possible to augment symbolic studies such as e.g., Hall (1992), Pierrehumbert (1994), Wiese (1996), Kessler and Treiman (1997), or Ewen and van der Hulst (2001) with extensive probabilistic information. Due to time and place constraints, we concentrate on two-consonantal clusters of grammar 2.1.3.

grammar version	word accuracy	syllable accuracy	syll bound. accuracy
baseline	10.33%		
(Müller, 2002)	89.27%	91.84%	95.84%
phonot. grammar	92.48%	94.35%	97.15%
phonot. on-nuc	92.29%	94.21%	97.09%
phonot. nuc-cod	92.39%	94.27%	97.11%
phonot. on-nuc-cod	92.64%	94.47%	**97.22%**

Table 3: Evaluation of four English grammar versions.

grammar version	word accuracy	syllable accuracy	syll bound. accuracy
baseline	1.59%		
(Müller, 2002)	86.06%	91.96%	95.90%
phonot. grammar	87.95%	93.09%	96.48%
phonot. nuc-cod	89.53%	94.09%	97.01%
phonot. on-nuc	89.97%	94.35%	97.15%
phonot. on-nuc-cod	90.45%	94.62%	**97.29%**

Table 4: Evaluation of four German grammar versions.

Phonotactic restrictions are often expressed by tables which describe the possibility of combination of consonants. Table 5 shows the possible combinations of German two-consonantal onsets (Wiese, 1996). However, the table cannot express differences in frequency of occurrence between certain clusters. For instance, it does not distinguish between onset clusters such as [pfl] and [kl]. If we consider the frequency of occurrence in a German dictionary then there is indeed a great difference. [kl] is much more common than [pfl].

4.1 German

Our method allows additional information to be added to tables such as shown in table 5. In what follows, the probabilities are taken from the rules of grammar 2.1.3. Table 6 shows the probability of

[2] Note that we only deal with phonotactic phenomena on the syllable level and not on the morpheme level.

mono		l	R	n	m	s	v	f	t	ts	p	k	j	z	g
0.380	S	0.160	0.093	0.056	0.074		0.165		0.318		0.131				
0.158	k	0.351	0.322	0.175			0.151								
0.090	b	0.489	0.510												
0.086	t		0.955				0.044								
0.083	f	0.620	0.364										0.015		
0.066	g	0.362	0.617	0.019											
0.042	p	0.507	0.400	0.030		0.061									
0.033	d		1.000												
0.019	s	0.200		0.066	0.100		0.133	0.033			0.133	0.333			
0.019	ts						1.000								
0.011	pf	0.882	0.117												
0.007	v		1.000												

Table 6: German two-consonantal onsets in monosyllabic words - sorted by probability of occurrence

mono		l	r	n	m	s	v	f	t	ts	p	k	j	z	g	w	S	d
0.322	s	0.157	0.001	0.099	0.060		0.001	0.004	0.223		0.150	0.174	0.006			0.120		
0.148	k	0.375	0.390		0.003		0.003						0.030			0.196		
0.093	b	0.420	0.574										0.004					
0.083	f	0.591	0.333										0.075					
0.079	p	0.480	0.457										0.056				0.005	
0.072	g	0.283	0.709													0.006		
0.068	t		0.686										0.039			0.274		
0.048	d		0.822										0.112			0.065		
0.035	h												0.089			0.910		
0.018	T		0.857										0.047			0.095		
0.014	S		0.878	0.030	0.030											0.060		
0.004	m												1.000					
0.003	n												1.000					
0.002	l												1.000					
0.002	v												1.000					

Table 7: English two-consonantal onsets in monosyllabic words - sorted by probability of occurrence

	Sonorants				Obstruents	
	l	R	n	m	s	v
Obstruents						
p	+	+	(+)	-	+	-
t	-	+	-	-	-	(+)
k	+	+	+	(+)	(+)	+
b	+	+	-	-	-	-
d	-	+	-	-	-	-
g	+	+	+	(+)	-	-
f	+	+	-	-	-	-
v	(+)	+	-	-	-	
ts	-	-	-	-	-	+
pf	+	+	-	-	-	-
S	+	+	+	+	-	+

Table 5: (Wiese, 1996) German onset clusters

occurrence of German obstruents ordered by their probability of occurrence. [S] occurs very often in German words as first consonant in two-consonantal onsets word initially. In the first row of table 6, the consonants which occur as second consonants are listed. We observe, for instance, that [St] is the most common two-consonantal onset in monosyllabic words. This consonant cluster appears in words such as *Staub* (dust), *stark* (strong), or *Stolz* (pride). We believe that there is a threshold indicating that a certain combination is very likely to come from a loanword. If we define the probability of a two-consonantal onset as

$$p(onset\_ini\_2) =_{def} p(C_1) \times p(C_2)$$

where $p(C_1)$ is the probability of the rule

$$\mathbf{X}.\mathrm{r.C}_1.\mathrm{s.t} \rightarrow \mathrm{C}_1\ \ \mathbf{X}.\mathrm{r.C}_2.\mathrm{s.t}$$

and $p(C_2)$ is the probability of the rule

$$\mathbf{X}.\mathrm{r.C}_2.\mathrm{s.t} \rightarrow \mathrm{C}_2,$$

then we get a list of two-consonantal onsets ordered by their probabilities:

$$p(St) > ... > p(sk) > p(pfl) > p(sl) > ... > p(sf)$$

These onsets occur in words such as *Steg* (footbridge), *stolz* (proud), *Staat* (state), *Skalp* (scalp), *Skat* (skat) *Pflicht* (duty), *Pflock* (stake), or *Slang* (slang) and *Slum* (slum). The least probable combination is [sf] which appears in the German word *Sphäre* (sphere) coming from the Latin word *sphaera*. The consonant cluster [sl] is also a very uncommon onset. Words with this onset are usually loanwords from English. The onset [sk], however, is an onset which occur more often in German words. Most of the words are originally from Latin and are translated into German long ago. Interestingly, the onset [pfl] is also a very uncommon onset. Most of these onsets result from the second sound shift where in certain positions the simple onset consonant /p/ became the affricate /pf/. The English translation of these words shows that the second sound shift was not applied to English. However, the most probable two-consonantal onset is [St]. The whole set of two-consonantal onsets can be seen in Table 8.

4.2 English

English two-consonantal onsets show that unvoiced first consonants are more common than voiced ones. However, two combinations are missing. The alveolar plosives /t/ and /d/ do not combine with the lateral /l/ in English two-consonantal onsets. Table 8 shows the most probable two-consonantal onsets sorted by their joint probability.

4.3 Comparison between English and German

The fricatives /s/ and /S/ are often regarded as extra syllabic. According to our study on two-consonantal onsets, these fricatives are very probable first consonants and combine with more second consonants than all other first consonants. They seem to form an own class. Liquids and glides are the most important second consonants. However, English prefers /r/ over /l/ in all syllable positions and /j/ over /w/ (except in monosyllabic words) and /n/ as second consonants. Nasals can only combine with very little first consonants. In German, we observe that /R/ is preferred over /l/ and /v/ over /n/ and /j/. Moreover, the nasal /n/ is much more common in German than in English as second consonants which applies especially to medial and final syllables.

When we compare the phonotactic restrictions of two languages, it is also interesting to observe which combinations are missing. If certain consonant clusters are not very likely or never occur in a language, this might have consequences for language understanding and language learning. Phonotactic gaps in one language might cause spelling mistakes in a second language. For instance, a typical Northern German name is *Detlef* which is often misspelled in English as *Deltef*. The onset cluster /tl/ can occur in medial and final German syllables but not in English. The different phonetic realization of /l/ may play a certain role that /lt/ is more natural than /tl/ in English.

Mono-syllabic: /st/ > /kr/ > /sk/ > /kl/ > /br/ > /gr/ > /sl/ > /fl/ > /sp/ > /tr/ > /dr/ > /bl/ > /sw/ > /pl/ > /pr/ > /sn/ > /hw/ > /kw/ > /fr/ > /gl/ > /sm/ > /tw/ > /Tr/ > /Sr/ > /fj/ > /dj/ > /kj/ > /pj/ > /mj/ > /dw/ > /hj/ > /nj/ > /tj/ > /vj/ > /lj/ > /sj/ > /Tw/ > /sf/ > /Tj/ > /Sw/ > /km/ > /kv/ > /gw/ > /Sn/ > /Sm/ > /pS/ > /bj/ > /sr/ > /sv/ **Initial** /pr/ > /st/ > /tr/ > /kr/ > /sp/ > /sk/ > /br/ > /gr/ > /fl/ > /kl/ > /fr/ > /bl/ > /pl/ > /sl/ > /kw/ > /dr/ > /sn/ > /sw/ > /gl/ > /hw/ > /nj/ > /sm/ > /sj/ > /pj/ > /Tr/ > /mj/ > /kj/ > /dj/ > /tw/ > /tj/ > /fj/ > /hj/ > /lj/ > /bj/ > /ps/ > /Sr/ > /dw/ > /sf/ > /vj/ > /gj/ > /gw/ > /pw/ > /mn/ > /Sm/ > /Tj/ > /Tw/ > /Sn/ > /tsw/ > /zj/ > /pt/ > /mw/ > /kn/ > /gz/ **Medial:** /st/ > /tr/ > /pr/ > /sp/ > /gr/ > /kj/ > /kr/ > /kw/ > /pl/ > /br/ > /tj/ > /lj/ > /dj/ > /dr/ > /kl/ > /nj/ > /sk/ > /mj/ > /fr/ > /pj/ > /bl/ > /fl/ > /bj/ > /gl/ > /gj/ > /fj/ > /Sn/ > /sj/ > /vj/ > /Sj/ > /Tr/ > /vr/ > /gw/ > /sl/ > /nr/ > /sw/ > /mr/ > /sn/ > /hj/ > /hw/ > /sm/ > /zj/ > /tSr/ > /rj/ > /sr/ > /dw/ > /Zr/ > /Sr/ > /jw/ > /tSw/ > /tSn/ > /vw/ > /Dr/ > /dZr/ > /dn/ > /Tj/ > /tw/ > /Sw/ > /Zj/ > /zr/ > /zn/ > /zw/ > /Zw/ > /dZj/ > /dZn/ > /dZw/ **Final:** /st/ > /tr/ > /kl/ > /bl/ > /gr/ > /dr/ > /pl/ > /br/ > /sk/ > /sp/ > /pr/ > /kr/ > /tj/ > /fr/ > /nj/ > /fl/ > /lj/ > /kw/ > /dj/ > /sj/ > /kj/ > /sl/ > /gl/ > /hw/ > /Sn/ > /vr/ > /Sj/ > /vj/ > /bj/ > /pj/ > /fj/ > /Tr/ > /mj/ > /gw/ > /sr/ > /sw/ > /sm/ > /nr/ > /sn/ > /tSr/ > /mr/ > /tw/ > /dZr/ > /zj/ > /gj/ > /dZj/ > /Sr/ > /Zr/ > /sf/ > /nw/ > /zr/ > /Tj/ > /rj/ > /Dr/ > /vw/ > /dw/ > /dn/ > /tSj/ > /pw/ > /jw/ > /hj/ > /St/ > /Zw/ > /tSn/ > /Zj/ > /pn/ > /Dj/ > /dZn/ > /zn/ > /Sw/ > /Zn/ > /tSw/ > /Tw/ > /bd/ > /tsj/ > /Dw/
Monosyllabic: /St/ > /tR/ > /Sv/ > /Sl/ > /kl/ > /fl/ > /kR/ > /Sp/ > /bR/ > /bl/ > /gR/ > /SR/ > /dR/ > /fR/ > /Sm/ > /kn/ > /gl/ > /kv/ > /pl/ > /Sn/ > /tsv/ > /pR/ > /pfl/ > /vR/ > /sk/ > /sl/ > /tv/ > /ps/ > /sp/ > /sv/ > /sm/ > /pfR/ > /pn/ > /gn/ > /sn/ > /fj/ > /sf/ **Initial:** /St/ > /tR/ > /pR/ > /Sp/ > /kR/ > /Sv/ > /gR/ > /Sl/ > /fR/ > /kl/ > /bR/ > /bl/ > /fl/ > /Sm/ > /gl/ > /tsv/ > /pl/ > /kv/ > /kn/ > /Sn/ > /dR/ > /SR/ > /sk/ > /pfl/ > /ps/ > /gn/ > /sl/ > /sm/ > /sts/ > /sf/ > /sv/ > /ks/ > /tv/ > /vR/ > /sn/ > /mn/ > /st/ > /pn/ > /sp/ > /fj/ > /pfR/ > /mj/ **Medial:** /St/ > /tR/ > /bR/ > /fR/ > /Sl/ > /gR/ > /kR/ > /bl/ > /dR/ > /Sp/ > /kl/ > /fl/ > /pR/ > /gl/ > /Sv/ > /SR/ > /st/ > /pl/ > /ks/ > /kv/ > /gn/ > /Sn/ > /Sm/ > /kn/ > /tsv/ > /pfl/ > /dl/ > /dn/ > /gm/ > /sp/ > /sn/ > /fn/ > /bn/ > /vj/ > /xR/ > /tn/ > /sl/ > /vR/ > /sk/ > /pj/ > /ps/ > /sts/ > /xn/ > /xl/ > /ml/ > /Rn/ > /Nn/ > /NR/ > /zn/ > /zl/ > /mn/ > /tl/ > /sf/ > /ln/ > /tsR/ > /tsl/ > /sR/ > /ft/ > /zR/ > /pfR/ > /pt/ > /nR/ > /sg/ > /pn/ > /dm/ > /tz/ > /sv/ > /zv/ > /tv/ **Final:** /St/ > /tR/ > /bl/ > /Sl/ > /bR/ > /fl/ > /kl/ > /dR/ > /gR/ > /Sp/ > /kR/ > /Sv/ > /fR/ > /SR/ > /gl/ > /ks/ > /dl/ > /pl/ > /gn/ > /pR/ > /Sn/ > /Sm/ > /kn/ > /dn/ > /kv/ > /tsv/ > /tl/ > /ml/ > /xl/ > /tsl/ > /gm/ > /pfl/ > /Nl/ > /zl/ > /tn/ > /xR/ > /vR/ > /fn/ > /bn/ > /vj/ > /zn/ > /Nn/ > /pn/ > /RR/ > /mn/ > /xn/ > /zR/ > /NR/ > /lR/ > /dZm/ > /tsR/ > /nl/ > /gv/ > /ps/ > /ft/ > /pfR/ > /tZl/ > /nR/ > /sp/ > /st/ > /sv/ > /sk/ > /sR/ > /sn/ > /sl/ > /sm/ > /sts/

Table 8: Two-consonantal onsets ordered by joint probability (top: English, bottom:German)

5 Discussion

Comparison of the syllabification performance with other systems is difficult: (i) different approaches differ in their training and evaluation corpus; (ii) comparisons across languages are hard to interpret; (iii) comparisons across different approaches require cautious interpretations. Nevertheless, we want to refer to several approaches that examined the syllabification task. Van den Bosch (1997) investigated the syllabification task with five inductive learning algorithms. He reported a generalization error for words of 2.22% on English data. However, the evaluation procedure differs from ours as he evaluates each decision (after each phoneme) made by his algorithms. Marchand et al. (to appear 2006) evaluated different syllabification algorithms on three different pronunciation dictionaries. Their best algorithm (SbA) achieved a word accuracy of 91.08%. The most direct point of comparison are the results presented by Müller (2002). Her approach differs in two ways. First, she only evaluates the German grammar and second she trains on a newspaper corpus. As we are interested in how her grammars perform on our corpus, we reimplemented her grammars and tested both in our 10-fold cross evaluation procedure. We find that the first grammar (Müller, 2001) achieves 85.45% word accuracy, 88.94% syllable accuracy and 94.37% syllable boundary accuracy for English and 84.21%, 90.86%, 95.36% for German respectively. The results show that the syllable boundary accuracy increases from 94,37% to 97.2% for English and from 95.3% to 97.2% for German. The experiments point out that phonotactic knowledge is a valuable source of information for syllabification.

6 Conclusions

Phonotactic restrictions are important for language perception and production. They influence the ability of children to segment words, and they help to recognize words in nonsense sequences. In this paper, we presented grammars which incorporate phonotactic restrictions. The grammars were trained and tested on a German and an English pronunciation dictionary. Our experiments show that English and German profit from phonotactic knowledge to predict syllable boundaries. We find evidence that German codas depend on the nucleus which does not apply for English. The English grammars which model the dependency of part of the onset or coda on the nucleus worsen the syllabification accuracy. However, the combination of both show a better performance than the base phonotactic grammar. This suggests that there are constrains in the selection of the onset and coda consonants.

7 Acknowledgments

I would like to thank Paul Boersma who invited me as a guest researcher at the Institute of Phonetic Sciences of the University of Amsterdam. Special thanks also to Detlef Prescher as well as to the three anonymous reviewers, whose comments were very useful while preparing the final version of this paper.

References

Harald R. Baayen, Richard Piepenbrock, and H. van Rijn. 1993. The CELEX lexical database—Dutch, English, German. (Release 1)[CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, Univ. Pennsylvania.

Anja Belz. 2000. Multi-syllable phonotactic modelling. In *Proceedings of SIGPHON 2000: Finite-State Phonology*, Luxembourg.

Juliette Blevins. 1995. The Syllable in Phonological Theory. In John A. Goldsmith, editor, *Handbook of Phonological Theory*, pages 206–244, Blackwell, Cambridge MA.

Julie Carson-Berndsen, Robert Kelly, and Moritz Neugebauer. 2004. Automatic Acquisition of Feature-Based Phonotactic Resources. In *Proceedings of the Workshop of the ACL Special Interest Group on Computational Phonology (SIGPHON)*, Barcelona, Spain.

Julie Carson-Berndsen. 1998. *Time Map Phonology. Finite State Models and Event Logics in Speech Recognition*, volume 5 of *Text, Speech and Language Technology*. Springer.

Eugene Charniak. 1996. Tree-bank grammars. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, AAAI Press/MIT Press, Menlo Park.

Walter Daelemans and Antal van den Bosch. 1992. Generalization performance of backpropagation learning on a syllabification task. In M.F.J. Drossaers and A Nijholt, editors, *Proceedings of TWLT3: Connectionism and Natural Language Processing*, pages 27–37, University of Twente.

Colin J. Ewen and Harry van der Hulst. 2001. *The Phonological Structure of Words. An Introduction*. Cambridge University Press, Cambridge, United Kingdom.

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

Daniel Kahn. 1976. *Syllable-based Generalizations in English Phonology*. Ph.D. thesis, Massachusetts Institute of Technology, MIT.

Brett Kessler and Rebecca Treiman. 1997. Syllable Structure and the Distribuation of Phonemes in English Syllables. *Journal of Memory and Language*, 37:295–311.

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.

Brigitte Krenn. 1997. Tagging syllables. In *Proceedings of the 5th European Conference on Speech Communication and Technology, Eurospeech 97*, pages 991–994.

Yannick Marchand, Connie A. Adsett, and Robert I. Damper. to appear 2006. Automatic syllabification in English: A comparison of different algorithms. *Language and Speech*.

Karin Müller. 2001. Automatic Detection of Syllable Boundaries Combining the Advantages of Treebank and Bracketed Corpora Training. In *Proc. 39th Annual Meeting of the ACL*, Toulouse, France.

Karin Müller. 2002. Probabilistic Context-Free Grammars for Phonology. In *Proceedings of the Workshop on Morphological and Phonological Learning at ACL 2002*.

Janet Pierrehumbert. 1994. Syllable structure and word structure: a study of triconsonantal clusters in English. In Patricia A. Keating, editor, *Phonological Structure and Phonetic Form*, volume III of *Papers in Laboratory Phonology*, pages 168–188. University Press, Cambridge.

Richard Sproat, editor. 1998. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer Academic, Dordrecht.

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

Jan P.H. Van Santen, Chilin Shih, Bernd Möbius, Evelyne Tzoukermann, and Michael Tanenblatt. 1997. Multilingual duration modeling. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 5, pages 2651–2654, Rhodos, Greece.

Michael S. Vitevitch and Paul A. Luce. 1999. Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *Journal of Memory and Language*, (40):374–408.

Jean Vroomen, Antal van den Bosch, and Beatrice de Gelder. 1998. A Connectionist Model for Bootstrap Learning of Syllabic Structure. *Language and Cognitive Processes. Special issue on Language Acquisition and Connectionism*, 13(2/3):193–220.

Andrea Weber and Anne Cutler. 2006. First-language phonotactics in second-language listening. *Journal of the Acoustical Society of America*, 119(1):597–607.

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