Inducing Probabilistic Syllable Classes Using Multivariate Clustering - GOLD -

Karin Müller, Bernd Möbius, Detlef Prescher

Institute of Natural Language Processing

University of Stuttgart

Abstract

An approach to automatic detection of syllable structure is presented. We demonstrate a novel application of EM-based clustering to multivariate data, exemplified by the induction of 3- and 5-dimensional probabilistic syllable classes. The 3-dimensional models were subjected to a pseudo-disambiguation task, the result of which shows that the onset is the most variable, or least predictable, part of the syllable. An extensive qualitative evaluation shows that the method yields phonologically meaningful syllable classes. We then propose a novel approach to grapheme-to-phoneme conversion and show that syllable structure represents valuable information for pronunciation systems.

1 Introduction

In this paper we present an approach to unsupervised learning and automatic detection of syllable structure. The primary goal of the paper is to demonstrate the application of EM-based clustering to multivariate data. The suitability of this approach is exemplified by the induction of 3- and 5-dimensional probabilistic syllable classes. A secondary goal is to outline a novel approach to the conversion of graphemes to phonemes (g2p) which uses a context-free grammar (cfg) to generate all sequences of phonemes corresponding to a given orthographic input word and then ranks the hypotheses according to the probabilistic information coded in the syllable classes.

Our approach builds on two resources. The first resource is a cfg for g2p conversion that was constructed manually by a linguistic expert (Müller, 2000). The grammar describes how words are composed of syllables and how syllables consist of parts that are conventionally called onset, nucleus and coda, which in turn are composed of phonemes, and corresponding graphemes. The second resource consists of a multivariate clustering algorithm that is used to reveal syllable structure hidden in unannotated training data. In a first step, we collect syllables by going through a large text corpus, looking up the words and their syllabifications in a pronunciation dictionary and counting the occurrence frequencies of the syllable types. Probabilistic syllable classes are then computed by applying maximum likelihood estimation from incomplete data via the EM algorithm. Two-dimensional EM-based clustering has been applied to tasks in syntax (Rooth et al., 1999), but so far this approach has not been used to derive models of higher dimensionality and, to the best of our knowledge, this is the first time that it is being applied to speech. Accordingly, we have trained 3- and 5-dimensional models for English and German syllable structure.

The obtained models of syllable structure were evaluated in three ways. Firstly, the 3-dimensional models were subjected to a pseudo-disambiguation task, the result of which shows that the onset is the most variable, or least predictable, part of the syllable. Secondly, the resulting syllable classes were qualitatively evaluated from a phonological and phonotactic point of view. Thirdly, a 5-dimensional syllable model for German was tested in a g2p conversion task. The results compare well with the best currently available data-driven approaches to g2p conversion (e.g., (Damper et al., 1999)) and suggest that syllable structure represents valuable information for pronunciation systems. Such systems are critical components in text-to-speech

(TTS) conversion systems, and they are also increasingly used to generate pronunciation variants in automatic speech recognition.

The rest of the paper is organized as follows. In Section 2 we introduce the multivariate clustering algorithm. In Section 3 we present four experiments based on 3- and 5-dimensional data for German and English. Section 4 is dedicated to evaluation and in Section 5 we discuss our results.

2 Multivariate Syllable Clustering

EM-based clustering has been derived and applied to syntax (Rooth et al., 1999). Unfortunately, this approach is not applicable to multivariate data with more than two dimensions. However, we consider syllables to consist of at least three dimensions corresponding to parts of the internal syllable structure: onset, nucleus and coda. We have also experimented with 5-dimensional models by adding two more dimensions: position of the syllable in the word and stress status. In our multivariate clustering approach, classes corresponding to syllables are viewed as hidden data in the context of maximum likelihood estimation from incomplete data via the EM algorithm. The two main tasks of EM-based clustering are (i) the induction of a smooth probability model on the data, and (ii) the automatic discovery of class structure in the data. Both aspects are considered in our application. We aim to derive a probability distribution $p(y)$ on syllables y from a large sample. The key idea is to view y as conditioned on an unobserved class $c \in C$, where the classes are given no prior interpretation. The probability of a syllable $y = (y_1, .., y_d) \in \mathcal{Y}_1 \times .. \times \mathcal{Y}_d$, $d \geq 3$, is defined as:

$$p(y) \;=\; \sum_{c \in C} p(c, y) = \sum_{c \in C} p(c) p(y|c) = \sum_{c \in C} p(c) \prod_{i=1}^{d} p(y_i|c)$$

Note that conditioning of y_i on each other is solely made through the classes c via the independence assumption $p(y|c) = \prod_{i=1}^{d} p(y_i|c)$. This assumption makes clustering feasible in the first place; later on (in Section 4.1) we will experimentally determine the number $|C|$ of classes such that the assumption is optimally met. The EM algorithm (Dempster, Laird, and Rubin, 1977) is directed at maximizing the incomplete data log-likelihood $L = \sum_y \tilde{p}(y) \ln p(y)$ as a function of the probability distribution p for a given empirical probability distribution $\tilde{p}$. Our application is an instance of the EM-algorithm for context-free models (Baum et al., 1970), from which simple re-estimation formulae can be derived. Let $f(y)$ the frequency of syllable y, and $|f| = \sum_{y \in \mathcal{Y}} f(y)$ the total frequency of the sample (i.e. $\tilde{p}(y) = \frac{f(y)}{|f|}$), and $f_c(y) = f(y)p(c|y)$ the estimated frequency of y annotated with c. Parameter updates $\hat{p}(c), \hat{p}(y_i|c)$ can thus be computed by ($c \in C, y_i \in \mathcal{Y}_i, i = 1, .., d$):

$$\hat{p}(c) = \frac{\sum_{y \in Y} f_c(y)}{|f|}, \text{ and}$$

$$\hat{p}(y_i|c) = \frac{\sum_{y \in \mathcal{Y}_1 \times .. \times \mathcal{Y}_{i-1} \times \{y_i\} \times \mathcal{Y}_{i+1} \times .. \times \mathcal{Y}_d} f_c(y)}{\sum_{y \in \mathcal{Y}} f_c(y)}$$

As shown by Baum et al. (1970), every such maximization step increases the log-likelihood function L, and a sequence of re-estimates eventually converges to a (local) maximum.

3 Experiments

A sample of syllables serves as input to the multivariate clustering algorithm. The German data were extracted from the Stuttgarter Zeitung (STZ), a newspaper corpus of about 31 million words. The English data came from the British National Corpus (BNC), a collection of samples of written and spoken language from a wide range of sources containing about 100 million words. For both languages, syllables

example	English Celex	modified to	example	English Celex	modified to
bacon	[N,]	[@n]	not	[n]	[nOt]
idealism	[m,]	[z@m]	on	[n]	[On]
burden	[n,]	[@n]	to	[t]	[tu:]
dangle	[l,]	[g@l]	us	[s]	[Vs]
father	[r*]	[@r]	would	[d]	[wUd]
the	[D]	[D@]	do	[d]	[du:]
them	[@m]	[D@m]	had	[d]	[h&d]
and	[N,]	[&nd]	shall	[S]	[S&l]
have	[v]	[h&v]	has	[z]	[h&z]
've	[v]	[h&v]	ma'am	[m]	[m&][@m]

Figure 1: Modifications of the English pronunciation lexicon.

were collected by going through the corpus, looking up the words and their syllabifications in a pronunciation dictionary (Baayen, Piepenbrock, and van Rijn, 1993) and counting the occurrence frequencies of the syllable types. We slightly modified the English pronunciation lexicon to obtain non-empty nuclei, e.g. /idealism/ [aI][dI@][lIzm,] was modified to [aI][dI@][lI][z@m] (SAMPA transcription). Figure 1 shows some modifications to prevent empty nuclei.

German Celex collapses the fricatives [C] and [x] into [x], the motivation being that the respective allophone is entirely predictable from the preceding vowel: [C] would appear after front vowels and [x] after back vowels. Moreover, the vocalized /r/ (Celex [6]) is coded as [@R], which yields syllable classes with [R] in the coda. We did not alter these conventions for our experiments.

Figures 2 and Figure 3 show the phonetic character sets of English and German, respectively. The columns show the phonemes together with an example word. The relevant graphemes are underlined.

p	pat	T	thin	I	pit	u:	boon
b	bad	D	then	E	pet	3:	burn
t	tack	s	sap	&	pat	eI	bay
d	dad	z	zap	V	putt	aI	buy
k	cad	S	sheep	O	pot	OI	boy
g	game	Z	measure	U	put	@U	no
N	bang	j	yank	@	another	aU	brow
m	mad	x	loch	i:	bean	I@	peer
n	nad	h	had	A:	barn	E@	pair
l	lad	w	why	O:	born	U@	p*oor*
r	rat	tS	cheap				
f	fat	dZ	jeep				
v	vat						

Figure 2: English phonetic character set.

p	Pakt	s	Glas	i:	Lied	E	Bett
b	Bad	z	Suppe	a:	klar	9	Götter
t	Tag	x	Bach, ich	u:	Hut	a	hat
d	dann	S	Schiff	y:	für	O	Glocke
k	kalt	h	Hand	E:	Käse	U	Pult
g	Gast	Z	Genie	e:	Mehl	@	Beginn
N	Klang	j	Jacke	2:	Möbel		
m	Mass	ts	Zahl	o:	Boot		
n	Naht	pf	Pferd	aI	weit		
l	Last	tS	Matsch	aU	Haut		
R	Ratte	dZ	Gin	OY	freut		
f	falsch			I	Mitte		
v	Welt			Y	Pfütze		

Figure 3: German phonetic character set.

class 0 0.212	NOP[I]	0.282	I	0.999	NOP[I]	0.460
	t	0.107			n	0.121
	l	0.074			N	0.096
	d	0.071			z	0.079
	b	0.065			d	0.049
	s	0.060			t	0.042
	r	0.040			m	0.017
	m	0.040			k	0.015
	n	0.033			ts	0.013
	S	0.018			f	0.012
	f	0.017			s	0.011
	w	0.017			v	0.009
	v	0.016			l	0.009
	st	0.015			ks	0.007
	z	0.014			dZ	0.006

Figure 4: Class #0 of a 3-dimensional English model with 12 classes

In two experiments, we induced 3-dimensional models based on syllable onset, nucleus, and coda. We collected 9327 distinct German syllables and 13,598 distinct English syllables. The number of syllable classes was systematically varied in iterated training runs and ranged from 1 to 200.

Figure 4 shows a segment of class #0 from a 3-dimensional English model with 12 classes[1]. The first column displays the class index 0 and the class probability $p(0)$. The most probable onsets and their probabilities are listed in descending order in the second column, as are nucleus and coda in the third and fourth columns, respectively. Empty onsets and codas are labeled “NOP[nucleus]”. The nucleus is considered obligatory and may therefore not be empty.

In two further experiments we induced 5-dimensional models, augmented by the additional parameters of position of the syllable in the word and stress status. We collected 16,595 distinct German syllables and 24,365 distinct English syllables. The number of syllable classes ranged from 1 to 200. Figure 5 illustrates (part of) class

[1]The unabridged syllable classes obtained from the 3- (12 classes) and 5-dimensional models (50 classes) for English and German are available on the World Wide Web at http://www.ims.uni-stuttgart.de/phonetik/g2p/

class 46 0.007	NOP[E]	0.630	E	0.990	nt	0.602	INI	0.627	STR	0.596
	ts	0.256	e:	0.004	t	0.128	FIN	0.331	USTR	0.403
	d	0.074	i:	0.003	n	0.092	MED	0.040		
	k	0.024			l	0.053				
	m	0.007			m	0.037				
	t	0.002			Rn	0.02				
	kv	0.001			N	0.013				
	n	0.001			s	0.012				
					pt	0.010				
					Rnst	0.005				
					ks	0.004				
					Rns	0.003				
					nts	0.003				
					Rt	0.002				
					Rts	0.001				
					Rp	0.001				

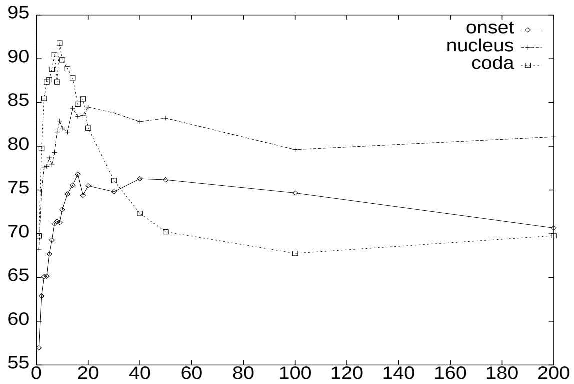
Figure 5: Class #46 of a 5-dimensional German model with 50 classes

#46 from a 5-dimensional German model with 50 classes. Syllable position and stress are displayed in the last two columns.

Two further experiments were performed on 3- and 5-dimensional German syllable types by setting the occurrence frequencies of distinct syllables to an artificial value of 1 (i.e. define $f(y) = 1$ for each syllable y (Section 2)). In these experiments, we used the previously collected 9,327 3-dimensional and 16,595 5-dimensional German distinct syllables. The number of syllable classes again ranged from 1 to 200. Our intention was to show that frequency counts represent valuable information for our clustering task. For the results of these experiments we refer to Section 4.4.

4 Evaluation

In the following sections, (i) the 3-dimensional models are subjected to a pseudo-disambiguation task (Section 4.1); (ii) the syllable classes are qualitatively evaluated (4.2); (iii) the 5-dimensional syllable model for German is tested in a g2p task (4.3); and (iv) the German syllable models are compared to models trained on syllable types (4.4).



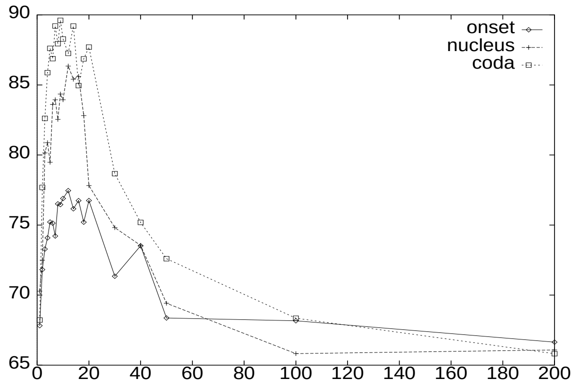


Figure 6: Evaluation on pseudo-disambiguation task for English (left) and German (right): accuracy for onset, nucleus and coda (y-axis) as a function of number of clusters (x-axis).

4.1 Pseudo-Disambiguation

We evaluated our 3-dimensional clustering models on a pseudo-disambiguation task similar to the one described by Rooth et al. (1999), but specified to onset, nucleus, and coda ambiguity. The first task is to judge which of two onsets on and on' is more likely to appear in the context of a given nucleus n and a given coda cod. For this purpose, we constructed an evaluation corpus of 3000 syllables (on, n, cod) selected from the original data. Then, randomly chosen onsets on' were attached to all syllables in the evaluation corpus, with the resulting syllables (on', n, cod) appearing neither in the training nor in the evaluation corpus. Furthermore, the elements on, n, cod, and on' were required to be part of the training corpus.

Clustering models were parameterized in (up to 10) starting values of EM-training and in the number of classes of the model (varying between 1 and 200), resulting in a sequence of 10×20 models. Accuracy was calculated as the number of times the model decided $p(on, n, cod) \geq p(on', n, cod)$ for all choices made. Two similar tasks were designed for nucleus and coda.

Results for the best starting values are shown in Figure 6. Models of 12 classes show the highest accuracy rates. For German we reached accuracy rates of 88–90% (nucleus and coda) and 77% (onset). For English we achieved accuracy rates of 92% (coda), 84% (nucleus), and 76% (onset).

One interpretation of these findings is that the coda is quite predictable for are given onset-nucleus combination. Slightly less predictable is the identity of the nucleus if onset and coda a given. The least restricted syllable part is the onset, which is hard to predict for a given nucleus-coda combination.

The results of the pseudo-disambiguation agree with intuition: in English and German, the onset is the most variable part of the syllable in terms of which phones may occur there for a given nucleus and coda; in other words, it is easy to find minimal pairs that vary in the onset; and it is easier to predict the nucleus and especially the coda for a given onset, as their choices are more restricted.

The pseudo-disambiguation results and linguistic intuition are further corroborated by observations in the context of automatic speech recognition that onsets appear to be more important for word discriminability than nuclei and codas are—and being less predictable, onset consonants are therefore also acoustically less variable and pronounced canonically more often than other phones in both read and spontaneous speech (Fosler-Lussier, Greenberg, and Morgan, 1999).

4.2 Qualitative Evaluation

In this section the syllable classes obtained from training 3- and 5-dimensional models for English and German are evaluated from a phonological point of view. Before discussing the results, we first review the basic properties of German and English syllable structure.

4.2.1 Syllable Structure

The phonotactics of English and German allow complex consonant clusters in the onset and coda of syllables. The maximum number of consonants in the onset is 3 in both languages. In German codas, clusters of up to 5 (or 6, if contractions such as *du schrumpfst's* [du: SrUmpfsts] "you shrink it" are considered too) consonants are observed, whereas English allows up to 4 coda consonants. Thus, the maximum number of consecutive consonants across syllable boundaries is 9 in German and 7 in English.

Certain restrictions exist as to which (classes of) consonants may occur in which position within the onset or coda of a syllable. For instance, in both languages there are only a few possible onset clusters with three consonants, and no phones other than obstruents may occur *before* an obstruent in the onset. In codas, only combinations and alternations of voiceless dental stops and fricatives are possible in positions 2 through 4 in English, and 3 through 5 (or 6) in German; after the first obstruent no phones other than obstruents may occur in the coda. Examples of the longest consonant clusters in English and German onsets and codas are given in Figures 7 and 8, respectively.

Sonorants (nasals, liquids, and glides) can only occur adjacent to the syllable nucleus. This pattern is sometimes referred to as the *sonority principle*, which ranks phone classes according to their natural acoustic sonority, which in turn is a correlate of the degree of constriction of the vocal tract. We have argued elsewhere (Möbius, 1998) that this ranking is not entirely consistent with actual acoustic measurements and is certainly not a valid descriptor of syllable structure across languages[2].

[2]See also the discussion of definitions (and measurements) of *sonority* in (Dogil and Luschützky, 1990). The authors reinterpret sonority as being derivable from a hierarchical segmental feature structure. This configurationally defined universal sonority scale can be extended to stating preference laws for syllabification.

Onsets		
Class	Clusters	Examples
SPL	sp + l/r/j	*split sprite spurious*
	st + r/j	*street studious*
	sk + l/r/j/w	*sclerosis script skewer squid*

Codas		
Class	Clusters	Examples
LPPS	lpts lkts	*sculpts mulcts*
LPSP	ltst	*waltzed*
LSSS	lfTs	*twelfths*
NPPS	mpts Nkts	*prompts adjuncts*
NPSP	mpst Nkst	*glimpsed jinxed*
PSPS	ksts	*texts*
PSSS	ksTs	*sixths*

Figure 7: English allows up to 3 consonants in the onset and up to 4 consonants in the coda of a syllable. The longest consonant clusters are restricted to a small number of distinct types. Phone class symbols: P = voiceless stop, S = voiceless fricative, L = liquid, N = nasal.

Onsets		
Class	Clusters	Examples
SPL	Sp + l/r	*Splitter Spritze*
	St + r	*Streit*
	sk + l/r	*Sklerose Skrupel*
Codas		
Class	Clusters	Examples
NPSSPS	mpfsts	*schrumpfst's*
NPSSP	mpfst	*schrumpfst*
	ntSst	*plantschst*
NPPSP	mptst	*promptst*
NSPSP	nftst	*?sanftst*

Figure 8: Similar to English, German allows up to 3 consonants in the onset of a syllable, but up to 6 consonants may occur in the coda. The longest consonant clusters are represented by only a small number of distinct types. Phone class symbols: P = voiceless stop, S = voiceless fricative, L = liquid, N = nasal.

The complexity of syllable onset and coda structures poses problems for a syllabification algorithm because, despite of the above-mentioned restrictions, ambiguous and multiple alternative syllable boundary locations are usually observed in polysyllabic words.

Determining the syllable boundary is important because the pronunciation of most phonemes is a function of their position in the syllable. This is most evident in the case of phonologically voiced obstruents in German: the voicing opposition for stops and fricatives is neutralized in the syllable coda. For instance, the phonological minimal pair *Bund* "union" – *bunt* "colorful" is in fact homophonic: [bUnt]. In English, voiceless stops are aspirated when they constitute the onset of a stressed syllable (e.g., [t] in *top*). They are not aspirated, however, if they are preceded in the onset by [s] (e.g., [t] in *stop*), followed in the onset by [l] or [r] (e.g., [t] in *stress*), or if they occur in the coda (e.g., [t] in *pot*).

4.2.2 Three-dimensional models

The 3-dimensional models are based on the three syllable positions, viz. onset, nucleus, and coda. The number of syllable classes (clusters) was systematically varied in iterated training runs and ranged from 1 to 200. The following discussion focusses on the results obtained for 12 classes. Terminating the clustering process at 12 classes yielded the cumulative maximum accuracy for onsets, nuclei and codas in the pseudo-disambiguation task (see Section 4.1).

Note that the accuracy peaks for the three syllable parts do not necessarily co-occur at the same number of classes, and that more than one local cumulative maximum may be observed. In such cases we selected the local cumulative maximum that coincides with the maximum accuracy for nuclei. As we have shown in Section 4.1, the absolute number of clusters does not appear to be crucial. Informal inspection of the syllable classes obtained for different but adjacent numbers of clusters revealed that the pertinent results are qualitatively similar to each other. This observation holds for 3-dimensional and 5-dimensional models of both German and English.

English and German. The 3-dimensional English and German models do not split the classes as clearly as the 5-dimensional models do. For instance, class #0 in Figure 4 contains the highly frequent function words *in, is, it, its* as well as the suffixes *-ing ,-ting, -ling.* Notice that these functions words and suffixes appear to be separated in the 5-dimensional model (classes #1 and #3 in Figure 9). Evidently, the position and stress features sharpen the distinction between different types of syllable structure and reveal structural information that otherwise remains hidden in the raw data.

4.2.3 Five-dimensional models

Our main observation was that the quality of the output dramatically increased when more dimensions were added. The following discussion is restricted to models with 50 classes, which yielded meaningful and interpretable syllable classes.

The 5-dimensional models used as factors the three syllable positions, augmented by the position of the syllable in the word and by stress status. The factor syllable position has four levels, viz. monosyllabic (ONE), initial (INI), medial (MED), and final (FIN). Stress is a binary factor with the levels stressed (STR) and unstressed (USTR), as annotated in the pronunciation dictionary.

We can look at the results from different angles. For instance, we can verify if any of the classes are mainly representatives of a syllable class pertinent to a particular nucleus. Another interesting aspect is whether there are syllable classes that represent parts of lexical content words, as opposed to high-frequency function words. Finally, some syllable classes may correspond to productive affixes.

English. In 24 out of the 50 syllable classes obtained for English one dominant

class 0 0.071	D 0.745 NOP[@] 0.166 fr 0.046 s 0.0353 j 0.006	@ 1	NOP[@] 0.877 m 0.0792 r* 0.04 n 0.002	ONE 0.999	STR 1
class 1 0.049	NOP[I] 0.914 h 0.071 b 0.010 sk 0.001	I 1	n 0.387 z 0.360 t 0.180 f 0.042 ts 0.02	ONE 0.916 INI 0.069 FIN 0.013 MED 0.001	STR 1
class 3 0.040	t 0.206 s 0.106 d 0.104 NOP[I] 0.101 k 0.059 n 0.052 r 0.052 v 0.036 st 0.035 z 0.033 l 0.028 T 0.026 dZ 0.02 m 0.019 p 0.017 S 0.014 tS 0.013 f 0.009 w 0.009 tr 0.008 Z 0.007 dr 0.006	I 0.993 @U 0.005	N 0.466 d 0.167 z 0.152 t 0.034 s 0.032 v 0.029 st 0.019 k 0.014 Nz 0.012 l 0.011 ts 0.01 m 0.007 ks 0.006 S 0.005 sts 0.003 dZ 0.003 p 0.002 vz 0.002 mz 0.002 n 0.002 lz 0.002 nz 0.002 dz 0.001	FIN 0.997 MED 0.002	USTR 0.999
class 4 0.037	t 0.211 v 0.115 D 0.102 d 0.095 NOP[@] 0.072 b 0.052 st 0.045 tS 0.043 l 0.032 s 0.026 p 0.023 f 0.022 n 0.021 g 0.021 k 0.019 dZ 0.017 Z 0.012 m 0.012 w 0.011 z 0.007 T 0.007 r 0.004	@ 0.978 O: 0.009 E@ 0.006 U@ 0.002 I@ 0.002	r* 0.597 z 0.115 d 0.057 l 0.054 n 0.045 ns 0.023 NOP[@] 0.022 m 0.012 t 0.012 lz 0.010 nt 0.009 ld 0.008 dz 0.007 nd 0.004 s 0.003 md 0.003 kt 0.002 kts 0.002 nts 0.002 nz 0.001 mz 0.001	FIN 0.996 MED 0.003	USTR 0.999

Figure 9: Classes #0, #1, #3, #4, of the 5-dimensional English model

class 10 0.028	S 0.257 m 0.227 d 0.063 t 0.059 r 0.059 z 0.039 l 0.038 v 0.037 Z 0.036 s 0.029 dZ 0.02 dr 0.016 n 0.016 tS 0.016 lj 0.015 st 0.013 k 0.008 NOP[@] 0.007 p 0.006 nj 0.004 kw 0.004 gr 0.001 w 0.001	@ 0.926 I 0.031 I@ 0.015 i: 0.006 E 0.005 U@ 0.003 U 0.002 O 0.002 & 0.002 aU 0.002 aI 0.001	n 0.388 nt 0.191 nz 0.088 l 0.066 nts 0.049 ns 0.048 s 0.031 r* 0.027 nd 0.018 p 0.016 z 0.013 m 0.011 st 0.011 lz 0.011 pt 0.003 ndz 0.003 nZ 0.002 mz 0.002 nst 0.002 ps 0.001 k 0.001 t 0.001	FIN 0.999	USTR 0.999
class 14 0.026	m 0.116 p 0.108 k 0.090 g 0.088 t 0.080 pl 0.052 st 0.051 l 0.050 r 0.034 tS 0.031 j 0.028 fr 0.025 gr 0.024 f 0.022 br 0.019 tr 0.017	eI 0.426 A: 0.165 E 0.140 O: 0.110 & 0.043	t 0.162 s 0.131 n 0.088 d 0.079 k 0.079 nd 0.052 ts 0.037 st 0.037 nt 0.029 dZ 0.028 ks 0.026 z 0.021 nst 0.020 m 0.019 nz 0.019 nZ 0.016 dz 0.015	ONE 0.696 FIN 0.276 MED 0.015 INI 0.011	STR 0.984 USTR 0.015
class 17 0.023	NOP[@] 0.973 p 0.017 b 0.002 n 0.002 t 0.001 j 0.001	@ 1	NOP[@] 0.325 r* 0.317 t 0.188 n 0.117 l 0.027 d 0.023	ONE 0.944 INI 0.050 FIN 0.005	STR 1
class 39 0.009	s 0.247 t 0.125 dZ 0.072 f 0.064 sp 0.052 z 0.047 gr 0.046 NOP[E] 0.045 v 0.036	E 0.721 & 0.120	kt 0.191 s 0.116 nt 0.107 lf 0.08 l 0.067 st 0.062 kts 0.054 ns 0.036 pt 0.033	FIN 0.982	STR 0.713 USTR 0.286

Figure 10: Classes #10, #14, #17, #39 of the 5-dimensional English model

nucleus per syllable class is observed. In all of these cases the probability of the nucleus is larger than 99% and in 7 classes the nucleus probability is 100%. Besides several diphthongs only the relatively infrequent vowels /V/, /A:/ and /3:/ do not dominate any class. Figures 9 and 10 show the classes that are described as follows.

High-frequency function words are represented by 10 syllable classes. For example, classes #0 and #17 are dominated by the determiners *the* and *a*, respectively, and class #1 contains function words that involve the short vowel /I/, such as *in, is, it, his, if, its*.

Productive word-forming suffixes are found in class #3 (*-ing*), and common inflectional suffixes in class #4 (*-er, -es, -ed*). Class #10 is particularly interesting in that it represents a comparably large number of common suffixes, such as *-tion, -ment, -al, -ant, -ent, -ence* and others. Similarly, class #39 contains stressed suffixes involving the vowel /E/, such as [Ekt] and [Es], as in *to protect* or *to progress*.

The majority of syllable classes, viz. 31 out of 50, contains syllables that are likely to be found in initial, medial and final positions in the open word classes of the lexicon. For example, class #14 represents mostly stressed syllables involving the vowels /eI, A:, e:, O:/ and others, in a variety of syllable positions in nouns, adjectives or verbs.

German. The majority of syllable classes obtained for German is dominated by one particular nucleus per syllable class. In 24 out of 50 classes the probability of the dominant nucleus is greater than 99%, and in 9 cases it is indeed 100%. The only syllable nuclei that do not dominate any class are the front rounded vowels /y:, Y, 2:, 9/, the front vowel /E:/ and the diphthong /OY/, all of which are among the least frequently occurring nuclei in the lexicon of German. Figure 11 depicts the classes that will be discussed now.

Almost one third (28%) of the 50 classes are representatives of high-frequency function words. For example, class #7 is dominated by the function words *in, ich,*

class 0 0.088	t 0.115 d 0.111 R 0.107 g 0.107 x 0.076 NOP[@] 0.067 S 0.061 s 0.056 n 0.054 l 0.052 b 0.032 f 0.021 st 0.02 z 0.017	@ 0.997	n 0.65 NOP[@] 0.239 R 0.069 s 0.016 nt 0.009 Rn 0.005 ns 0.003 Rt 0.001	FIN 0.977 MED 0.022	USTR 1
class 4 0.032	NOP[aI] 0.624 z 0.163 k 0.043 v 0.029 fR 0.021 m 0.016	aI 1	NOP[aI] 0.689 n 0.303 nst 0.002 ns 0.001	INI 0.755 ONE 0.226	STR 0.999
class 7 0.029	NOP[I] 0.730 z 0.259	I 1	n 0.533 x 0.204 st 0.150 nt 0.067 ns 0.007 m 0.003	ONE 0.867 INI 0.128	STR 0.915 USTR 0.084
class 24 0.019	NOP[i:] 0.163 v 0.105 l 0.089 f 0.073 m 0.07 d 0.065 t 0.046 n 0.044 z 0.034 g 0.034 b 0.024 S 0.019	i: 1	NOP[i:] 0.915 R 0.040 l 0.013 t 0.005 f 0.003	INI 0.613 MED 0.286 FIN 0.1	STR 0.69 USTR 0.309
class 26 0.017	f 0.573 NOP[E] 0.351 ts 0.009 h 0.006	E 0.987 o: 0.007 O 0.001	R 0.983	INI 0.906 MED 0.093	USTR 0.994
class 34 0.011	l 0.408 t 0.175 d 0.133	I 0.905	x 0.690 xt 0.108 k 0.047	FIN 0.936 MED 0.063	USTR 0.999
class 40 0.009	b 0.144 R 0.128 t 0.119 v 0.095 ts 0.090 gl 0.022	aI 0.999	NOP[aI] 0.706 n 0.103 x 0.077 ts 0.057 s 0.016 l 0.015	MED 0.876 FIN 0.119	USTR 0.596 STR 0.403

Figure 11: Classes #0 #4, #7, # 24, #26, #34, #40 of the 5-dimensional German model

ist, im, sind, sich, all of which contain the short vowel /I/. Class #24, on the other hand, includes function words containing the long vowel /i:/, such as *wie, die, wir, mir, dir, ihr, sie*.

Another 32% of the 50 classes represent syllables that are most likely to occur in initial, medial and final positions in the open word classes of the lexicon, i.e. nouns, adjectives, and verbs. Class #4 covers several lexical entries involving the diphthong /aI/, mostly in stressed word-initial syllables. Class #40 provides complimentary information, as it also includes syllables containing /aI/, but here mostly in word-medial position.

We also observe syllable classes that represent productive prefixes (e.g., *ver-, er-, zer-, vor-, her-* in class #26) and suffixes (e.g., *-lich, -ig* in class #34). Finally, there are two syllable classes (e.g. class #0) that cover the most common inflectional suffixes involving the vowel /@/ (schwa).

Class numbers are informative insofar as the classes are ranked by decreasing probability. Lower-ranked classes tend (i) not to be dominated by one nucleus; (ii) to contain vowels with relatively low frequency of occurrence; and (iii) to yield less clear patterns in terms of word class or stress or position. For illustration, class #46 (Figure 5) represents the syllable *ent* [Ent], both as a prefix (INI) and as a suffix (FIN), the former being unstressed (as in *Entwurf* [Ent"vURf] "design") and the latter stressed (as in *Dirigent* [di:Ri:"gEnt] "conductor").

4.3 Evaluation by g2p Conversion

In this section, we present a novel method of g2p conversion, (i) using a cfg to produce all possible phonemic correspondences of a given grapheme string, (ii) applying a probabilistic syllable model to rank the pronunciation hypotheses, and (iii) predicting pronunciation by choosing the most probable analysis.

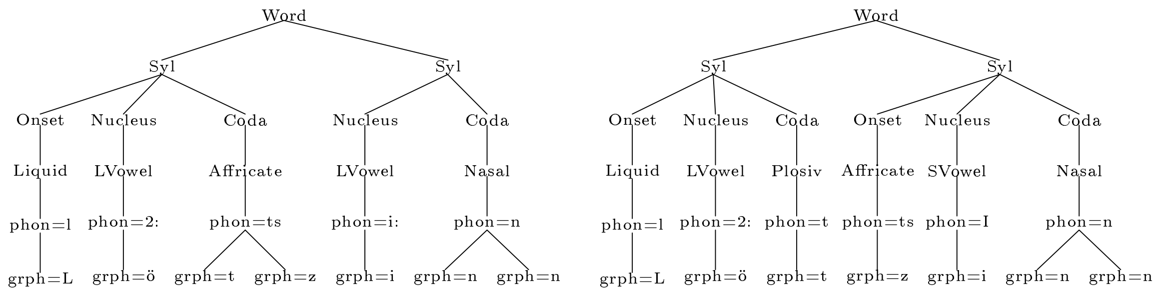


Figure 12: An incorrect (left) and a correct (right) cfg analysis of *Lötzinn.*

We used a cfg for generating transcriptions, because grammars are expressive and writing grammar rules is easy and intuitive. Our grammar describes how words are composed of syllables and syllables branch into onset, nucleus and coda. These syllable parts are rewritten by the grammar as sequences of natural phone classes, e.g. stops, fricatives, nasals, liquids, as well as long and short vowels, and diphthongs. The phone classes are then reinterpreted as the individual phonemes that they are made up of. Finally, for each phoneme all possible graphemic correspondences are listed.

Figure 12 illustrates two analyses (out of 100) of the German word *Lötzinn (tin solder).* The phoneme strings (represented by non-terminals named "phon=...") and the syllable boundaries (represented by the non-terminal "Syl") can be extracted from these analyses. Figure 12 depicts both an incorrect analysis [l2:ts][i:n] and its correct counterpart [l2:t][tsIn].

The next step is to rank these transcriptions by assigning probabilities to them. The key idea is to take the product of the syllable probabilities. Using the 5-dimensional[3] German syllable model yields a probability of $7.5 \cdot 10^{-7} \cdot 3.1 \cdot 10^{-7} = 2.3 \cdot 10^{-13}$ for the incorrect analysis and a probability of $1.5 \cdot 10^{-7} \cdot 6.5 \cdot 10^{-6} = 9.8 \cdot 10^{-13}$

[3] Position can be derived from the cfg analyses, stress placement is controlled by the most likely distribution.

g2p system	3-dim baseline	3-dim classes	5-dim baseline	5-dim classes
word accuracy	66.8 %	67.4 %	72.5 %	75.3 %

Figure 13: Evaluation of g2p systems using probabilistic syllable models

for the correct one. Thus we achieve the desired result of assigning the higher probability to the correct transcription.

We evaluated our g2p system on a test set of 1835 unseen words. The ambiguity expressed as the average number of analyses per word was 289. The test set was constructed by collecting 295,102 words from the German Celex dictionary that were *not seen* in the STZ corpus. From this set we manually eliminated (i) foreign words, (ii) acronyms, (iii) proper names, (iv) verbs, and (v) words with more than three syllables. The resulting test set is available on the World Wide Web[4].

Figure 13 shows the performance of four g2p systems. The second and fourth columns show the accuracy of two baseline systems: g2p conversion using the 3- and 5-dimensional empirical distributions (Section 2), respectively. The third and fifth columns show the word accuracy of two g2p systems using 3- and 5-dimensional syllable models, respectively.

The g2p system using 5-dimensional syllable models achieved the highest performance (75.3%), which is a gain of 3% over the performance of the 5-dimensional baseline system and a gain of 8% over the performance of the 3-dimensional models[5].

[4]http://www.ims.uni-stuttgart.de/phonetik/g2p/

[5]45 resp. 95 words could not be disambiguated by the 3- resp. 5-dimensional empirical distributions. The reported relatively small gains can be explained by the fact that our syllable models were applied only to this small number of ambiguous words.

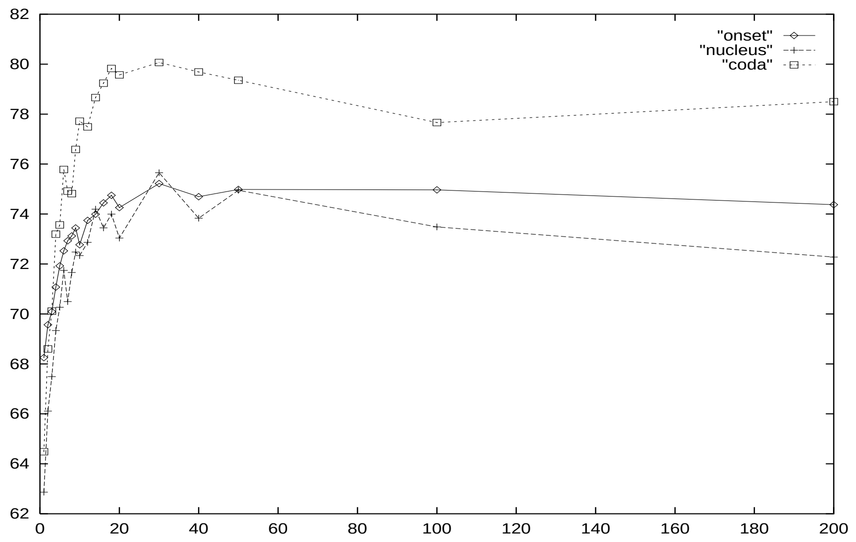


Figure 14: Evaluation on pseudo-disambiguation task for German models trained on 3-dimensional syllable types: accuracy for onset, nucleus and coda (y-axis) as a function of number of clusters (x-axis).

4.4 Type/Token Evaluation

This section is dedicated to the question whether frequency counts represent valuable information for our clustering task. We try to answer this question by comparing our previous evaluation results for German syllable models with results of so-called "normalized models", i.e. models trained on 3- and 5-dimensional syllable types.

Pseudo-Disambiguation. We evaluated the 3-dimensional normalized clustering models (trained on German syllable types) on the pseudo-disambiguation task described in Section 4.1. Clustering models were parameterized in up to 10 starting values of EM-training, and in the number of classes of the model (up to 200), resulting in a sequence of 200 models. Results for the best starting values are shown in Figure 14. Models with 25 classes show the highest accuracy rates of about 80% (coda) and 75% (onset and nucleus). These results are less impressive than the results obtained by models trained on syllable tokens: firstly, these accuracy rates are

g2p system	5-dim baseline	5-dim classes
types	52.8 %	66.0 %
tokens	72.5 %	75.3 %

Figure 15: Evaluation of g2p systems using syllable models trained on types/tokens

about 10% lower, and secondly, the accuracy rates of onset and nucleus are highly correlated and thus do not agree with intuition that the onset is the most variable part of the syllable. The results of the pseudo-disambiguation task clearly show that frequency counts represent valuable information for our clustering task.

Evaluation by g2p Conversion. The 5-dimensional normalized clustering models trained on German syllable types were evaluated on the g2p conversion task described in Section 4.3. Figure 15 shows the performance of four g2p systems. The second column shows the accuracy rates of two baseline systems: g2p conversion using the 5-dimensional empirical distribution[6]. The third column shows the accuracy rates of two g2p systems using the 5-dimensional models trained on types and tokens. The g2p system using a 5-dimensional syllable model trained on tokens achieved the highest performance (75.3%), which is a gain of about 10% over the g2p system using a syllable model trained on types. Furthermore, the g2p baseline system using frequency counts of syllables clearly outperforms the g2p baseline system using the uniform distribution on syllable types. The results of the g2p conversion task impressively show that frequency counts are a valuable information source.

[6]Note that the empirical distribution on a typenized training corpus is equal to the uniform distribution, i.e. the g2p system using this distribution completely disambiguates by random choice. Thus, the relatively high accuracy rate of 52.8% achieved by this system can be attributed to the quality of our hand-crafted cfg.

5 Discussion

We have presented an approach to unsupervised learning and automatic detection of syllable structure, using EM-based multivariate clustering. The method yields phonologically meaningful syllable classes. These classes are shown to represent valuable input information in a g2p conversion task.

In contrast to the application of two-dimensional EM-based clustering to syntax (Rooth et al., 1999), where semantic relations were revealed between verbs and objects, the syllable models cannot *a priori* be expected to yield similarly meaningful properties. This is because the syllable constituents (or phones) represent an inventory with a small number of units which can be combined to form meaningful larger units, viz. morphemes and words, but which do not themselves carry meaning. Thus, from a phonological point of view it is not to be expected that a particular instantiation of a given syllable structure occurs significantly more often than others; for example, the syllable /mad/ is not a priori expected to be more frequent than /mag/, both being representatives of the structure CVC. Yet, some syllable types do occur more often than others simply because certain morphemes and words have a higher frequency count than others in a given text corpus.

As discussed in Section 4.2, however, we do find some interesting properties of syllable classes, some of which apparently represent high-frequency function words and productive affixes, while others are typically found in lexical content words. Subjected to a pseudo-disambiguation task (Section 4.1), the 3-dimensional models confirm the intuition that the onset is the least predictable part of the syllable.

In a feasibility study we applied the 5-dimensional syllable model obtained for German to a g2p conversion task. Automatic conversion of a string of characters, i.e. a word, into a string of phonemes, i.e. its pronunciation, is essential for applications such as speech synthesis from unrestricted text input, which can be expected to contain words that are not in the system's pronunciation dictionary.

In the g2p experiment, orthographic words were processed by a parser that uses a context-free grammar to produce all possible phonemic correspondences of the grapheme string. The pronunciation hypotheses are then ranked by applying a probabilistic syllable model. The model was trained on a text corpus whose words were looked up in a pronunciation dictionary. The trained trees include information on syllable boundaries and parts of syllables (onset, nucleus, coda). Syllabic stress is not coded explicitly; it is hypothesized as part of the probabilistic tree. Pronunciation is finally predicted by taking the most probable analysis, computed from the combined probabilities of the onset, nucleus, coda, syllable part, stress, and syllable class.

The main purpose of the feasibility study was to demonstrate the relevance of the phonological information on syllable structure for g2p conversion. Therefore, information and probabilities derived from an alignment of grapheme and phoneme strings, i.e. the lowest two levels in the trees displayed in Figure 12, was deliberately ignored.

Data-driven pronunciation systems usually rely on training data that include an alignment of graphemes and phonemes. Damper et al. (1999) have shown that the use of unaligned training data significantly reduces the performance of g2p systems. In our experiment, with training on unannotated text corpora and without an alignment of graphemes and phonemes, we obtained a word accuracy rate of 75.3% for the 5-dimensional German syllable model.

Comparison of this performance with that other systems is difficult: (i) hardly any quantitative g2p performance data are available for German; (ii) comparisons across languages are hard to interpret; (iii) comparisons across different approaches require cautious interpretations.

The most direct point of comparison is the method presented by Müller (2000).

In one of her experiments, the standard probability model was applied to the hand-crafted cfg presented in this paper, yielding 42% word accuracy as evaluated on our test set. Running the test set through the pronunciation rule system of the IMS German Festival TTS system (Möhler, 1999) resulted in 55% word accuracy. The Bell Labs German TTS system (Möbius, 1999) performed at better than 95% word accuracy on our test set. This TTS system relies on an annotation of morphological structure for the words in its lexicon and it performs a morphological analysis of unknown words (Möbius, 1998); the pronunciation rules draw on this structural information.

These comparative results emphasize the value of phonotactic knowledge and information on syllable structure and morphological structure for g2p conversion.

In a comparison across languages, a word accuracy rate of 75.3% for our 5-dimensional German syllable model is slightly higher than the best data-driven method for English with 72% (Damper et al., 1999). The data-driven systems for English that were evaluated by Damper et al. produced g2p translation results of 72% word accuracy for Pronunciation by Analogy (Damper et al., 1999), 65% for IB1-IG (Daelemans, van den Bosch, and Weijters, 1997; van den Bosch, 1997), and 50% for NETspeak (McCulloch, Bedworth, and Bridle, 1987); a rule-based system (Elovitz et al., 1976) achieved less than 26% correct word pronunciations.

Recently, Bouma (2000) has reported a word accuracy of 92.6% for Dutch, using a 'lazy' training strategy on data aligned with the correct phoneme string, and a hand-crafted system that relied on a large set of rule templates and a many-to-one mapping of characters to graphemes preceding the actual g2p conversion.

We are confident that a judicious combination of phonological information of the type employed in our feasibility study with standard techniques such as g2p alignment of training data will produce a pronunciation system with a word accuracy that matches the one reported by Bouma (2000).

We believe, however, that for an optimally performing system as it is desired for TTS, an even more complex design will have to be adopted. In many languages, including English, German and Dutch, access to morphological and phonological information is required to reliably predict the pronunciation of words; this view is further evidenced by the performance of the Bell Labs system, which relies on precisely this type of information. We agree with Sproat (1998, p. 77) that it is unrealistic to expect optimal results from a system that has no access to this type of information or is trained on data that are insufficient for the task.

References

Baayen, Harald R., 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.

Baum, Leonard E., Ted Petrie, George Soules, and Norman Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Math. Statistics*, 41(1):164–171.

Bouma, Gosse. 2000. A finite state and data-oriented method for grapheme to phoneme conversion. In *Proc. 1st Conf. North American Chapter of the ACL (NAACL)*, Seattle, WA.

Daelemans, Walter, Antal van den Bosch, and Ton Weijters. 1997. IGTree: using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review*, 11:407–423.

Damper, Robert I., Yannick Marchand, Martin J. Adamson, and Kjell Gustafson. 1999. Evaluating the pronunciation component of text-to-speech systems for

English: a performance comparison of different approaches. *Computer Speech and Language*, 13:155–176.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *J. Royal Statistical Soc.*, 39(B):1–38.

Dogil, Grzegorz and Hans Christian Luschützky. 1990. Notes on sonority and segmental strength. *Rivista di Linguistica*, 2(2):3–54.

Elovitz, Honey S., Rodney Johnson, Astrid McHugh, and John E. Shore. 1976. Letter-to-sound rules for automatic translation of English text to phonetics. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24:446–459.

Fosler-Lussier, Eric, Steven Greenberg, and Nelson Morgan. 1999. Incorporating contextual phonetics into automatic speech recognition. In *Proceedings of the 14th International Congress of Phonetic Sciences*, volume 1, pages 611–614, San Francisco.

McCulloch, Neil, Mark Bedworth, and John Bridle. 1987. NETspeak—a re-implementation of NETtalk. *Computer Speech and Language*, 2:289–301.

Möbius, Bernd. 1998. Word and syllable models for German text-to-speech synthesis. In *Proc. 3rd ESCA Workshop on Speech Synthesis (Jenolan Caves)*, pages 59–64.

Möbius, Bernd. 1999. The Bell Labs German text-to-speech system. *Computer Speech and Language*, 13:319–358.

Möhler, Gregor. 1999. IMS Festival. [http://www.ims.uni-stuttgart.de/phonetik/synthesis/index.html].

Müller, Karin. 2000. Probabilistische kontext-freie Grammatiken für Silbifizierung und Graphem-Phonem-Konvertierung. In *AIMS Report 6(2)*, IMS, Univ. Stuttgart.

Rooth, Mats, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proc. 37th Ann. Meeting of the ACL*, College Park, MD.

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

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