BAStat : New Statistical Resources at the Bavarian Archive for Speech Signals

Florian Schiel

Bavarian Archive for Speech Signals c/o Institute of Phonetics and Speech Processing, LMU München
Schellingstr. 3, 80799 München, Germany
schiel@bas.uni-muenchen.de

Abstract

A new type of language resource 'BAStat' has been released by the Bavarian Archive for Speech Signals. In contrast to primary resources like speech and text corpora BAStat comprises statistical estimates based on a number of primary resources: first and second order occurrence probability of phones, syllables and words, duration statistics, probabilities of pronunciation variants of words and probabilities of context information. Unlike other statistical speech resources BAStat is based solely on recordings of conversational German and therefore models spoken language. It consists of 7-bit ASCII tables and matrices to maximize inter-operability between different platforms and can be downloaded from the BAS web-site. This paper gives a detailed description about the empirical basis, the contained data types, some interesting interpretations and a brief comparison to the text-based statistical resource CELEX.

1. Introduction

In this contribution we describe a new type of language resource *BAStat* published by the Bavarian Archive for Speech Signals (BAS): While a speech corpus may be considered as a primary type of language resource, BAStat is of secondary nature, that is it contains statistical information derived from a (growing) number of different primary language resources (LR). Similar LRs of secondary nature are for example lexica, HMMs, statistical rule sets or grammars.

BAStat provides statistical information about the phonetic structure of German conversational speech, that is what types of phonemes, syllables and words are produced in which surface form, with which duration and in which context in a large number of recordings of spoken German. In contrast to existing statistical resources such as CELEX (Baayen et al., 1995) BAStat is based solely on spoken conversational speech. Hence it provides real occurrence probabilities of linguistic entities in spoken dialogues and the corresponding duration statistics.

Aside from technical applications we would like to encourage researchers to utilize the statistics from BAStat in the context of psycholinguistic experiments that require the knowledge of a-priori probabilities of linguistic entities (e.g. Levelts production model (Levelt, 1989) or Exemplar Theory (Pierrehumbert, 2001)). This often regards the proper selection of lingustic units with high/low/equal probabilities for perceptual experiments.

This contribution is roughly structured into three parts: First we define the empirical basis of BAStat, secondly we describe the contents as being down-loadable from the BAS web-site at the time of writing and present a selection of interesting results that can be drawn from the resource. Finally we briefly compare BAStat to the text-based statistical resource CELEX.

2. Empirical Basis of BAStat

In its present stage[1] BAStat contains data based solely on spoken conversational speech. The reason to deal with conversational speech as opposed to read speech is that we deem spontaneous speech a more challenging and interesting subject for most scientific investigations, especially if we consider human - machine dialog systems that are increasingly allowing connected speech input.

In this section we will present the underlying speech resources, the annotation and automatic segmentation applied to these and a method to derive syllable information for statistical analysis.

2.1. Corpora

The BAS maintains and distributes a multitude of LRs, mostly corpora of spoken German (BAS, 2010). Some BAS corpora consist entirely or partly of recordings of unimpeded conversational speech, that is either two humans talking to each other or a single human talking to a virtual machine in a Wizard-of-Oz experiment or a triad situation where two human speakers interact which each other and simultaneously with a machine.

Table 1 lists the BAS corpora included in the BAStat analysis in its present stage; the maximum number of analyzed word tokens is 689966 representing 16426 word types. We plan to extend this set in the course of 2010 by the spontaneous parts of the *ALC*, *RVG-J* and *SmartWeb* corpora (BAS, 2010). In the following we briefly describe the main properties of the different speech corpora.

Corpus	RVG1	VM1	VM2	SK
Speakers	450	780	259	233
Setting	interview	dialog	dialog	WOZ
Word tokens	63162	285280	153438	55681

Table 1: BAS corpora exploited for BAStat analysis

2.1.1. Regional Variants of German 1 (RVG1)

The RVG1 corpus comprises speech recordings in 4 channels of approx. 450 speaker recorded in the main European areas of spoken German, that is Germany, Austria and Switzerland. A small part of these recordings were done

[1] Feb 2010

in form of an interview where the interviewee was asked to report about her/his work during the previous week using casual speech. The analyzed channel for BAStat is the headset microphone (channel c) of the interviewee.
Recordings across speakers differ considerably in length, speech rate, accent and content.

2.1.2. Verbmobil

The Verbmobil corpora (VM1 + VM2) contain conversations between two business colleagues who have to schedule a number of appointments (VM1, VM2) and talk about planing a business trip (VM2 only). While in VM1 the dialogue is structured by a Push-to-Talk button, in VM2 the speakers are free to interrupt their dialogue partner. The analyzed channels are the headset microphones (channel c) of both speakers.

2.1.3. SmartKom

The SmartKom corpus (SK) has been recorded in a multimodal WOZ setting where the speaker was asked to test an information kiosk with fully functional speech and gestural interface. The topics here are sight-seeing, restaurants, travel information, TV guide, VCR programming, fax and email, cinema information and reservation. The analyzed channel is the front directional microphone (channel d, if available) or the headset (channel h).

2.2. Orthographic Transcription

All four resources have been transcribed according to the Verbmobil Transliteration Convention (e.g. (Burger et al., 1997)). The resulting transcripts and other annotations have been summarized in BAS Partitur Format (BPF) files as well as ATLAS Format XML files ('Annotation Graphs'). Compatible pronunciation lexica manually coded in German SAM-PA are available for all resources. Linguistic markers such as Part-of-Speech, word classes etc. were not considered in BAStat because not all BAS corpora provide tagging for these. An exception is the marker for *content* and *function word* which has been considered in BAStat. Para-linguistic markers such as articulatory noise, background noise, cross-talk etc. have also not been taken into account for this analysis. The total orthographic transcript results in 16426 word types.

2.3. Phonetic Transcript and Segmentation

The basic durational linguistic entity in BAStat is the phonemic segment as given by the German SAM-PA definition set[2]. All other durational entities such as phone sequences, syllables or words are derived from this basic segmentation. Aside from SAM-PA there exist deviating phoneme definition sets for German, for instance in some cases the affricates /ts/, /tS/ and /pf/ are treated as separate entities: /t/ + /s/, ... , tense vowels also appear as unlengthened (e.g. /o/) or the set of diphthongs in SAM-PA /aU/, /aI/ and /OY/ is extended by diphthongs and triphthongs that combine the vocalized 'r' /6/ with practically all German vowels and diphthongs. Finally, most phoneme sets for German do not allow the coding of non-German words derived from English or French.

[2] *www.bas.uni-muenchen.de/forschung/ Bas/BasSAMPA*

To accommodate these differing needs we decided to perform all analysis in BAStat based on two phoneme sets: one basic set of 52 phoneme symbols, with separated affricates, the three fundamental German diphthongs and extended by four French nasalized vowels, and a second set, called the *6-set*, which additionally contains all diphthong and triphthong combinations with vocalized 'r'. Table 2 list the two sets in detail.

set	SAM-PA phonemes							
basic	OY	aI	aU	E:	y:	2:	a:	u:
	i:	a~	E~	O~	9~	2	6	9
	E	I	N	O	Q	S	U	Y
	b	d	e	f	g	h	i	j
	m	n	o	p	r	s	t	u
	y	z	o:	e:	@	C	Z	a
	k	l	v	x				
6-set	OY6	aI6	aU6	E:6	y:6	2:6	a:6	u:6
	o:6	e:6	i:6	E6	I6	O6	U6	Y6
	26	a6	e6	i6	o6	u6	y6	96

Table 2: The basic phoneme set and extension to vocalized 'r' (SAM-PA /6/) used in BAStat

The phonetic transcript and segmentation are produced automatically by the Munich AUtomatic Segmentation system (MAUS, e.g. (Schiel, 1999)).
MAUS calculates a string of canonical phonemic segments from the orthographic transcription and then derives hypothetical pronunciation deviations from this canonical pronunciation by means of data-driven substitution rules. In parallel the a-priori expected probabilities of these deviations are calculated and integrated into the hypothesis graph. The result is a directed acyclic graph representing all possible combinations of pronunciation predicted for this utterance together with their combined probability. Finally this graph is passed to a Viterbi decoder (Young, 1995) finding and time-aligning the most likely pronunciation variant through the hypothesis graph given the acoustics of the speech signal. The result is a phonetic transcript and segmentation close to the actual sequence of phones.
MAUS allows for a variety of reduction and assimilation phenomena, detects silence intervals and has a transcription accuracy of about 96% of the inter-labeler agreement in spontaneous speech. Comparing the segmental boundaries with those of a reference segmentation by phoneticians we find a Gaussian-like distribution of about 25msec standard deviation from the reference boundaries.

2.4. Syllabification

The annotation of the analyzed BAS corpora contains no information regarding the syllabification of the phonetic transcript, that is the number and boundaries of syllables per word token is unclear. Furthermore, since our segmentation of the speech signal is non-canonical we are often faced with the problem that syllable nuclei are deleted and new consonantal clusters emerge which are difficult to handle in terms of syllable structure.
To obtain statistical information about empirically found syllables we applied a simple syllabification algorithm de-

orthography	wir	haben	dazu	keinen	eigenen	Beitrag	<"ah>	leisten	müssen
canonical	v'i:6	h'a:b@n	dats'u:	k'aIn@n	Q'aIg@n@n	b'aItr"a:k	Q'E:	l'aIst@n	m'Ys@n
phonetic	vi:6	ha:m	datsu:	kaIn	aIg@nn	baItra:k	QE:	laIstn	mYsn
syllabification	vi:6	ha:m	dat+su:	kaIn	aI+g@n+n	baI+tra:k	QE:	laIs+tn	mYs+n

Table 3: Example for syllabification randomly selected from RVG1; syllable boundaries are marked by a '+'

veloped by U. Reichel in 2009 in our lab. Based on the string of phonetic segments delivered by MAUS this algorithm searches for minima of sonority between syllable nuclei and applies some other heuristic rules to account for special cases where the primary strategy fails. Phonetic segments are then merged according to this syllabification to form syllable segments for the BAStat analysis.
Since the phonetic transcript does not necessarily follow phonotactical rules, this sometimes results in very unusual 'syllables' (e.g. syllables consisting of one or more consonants only), especially in cases where whole words are reduced to single phones.
Table 3 shows a typical utterance recorded in the RVG1 corpus together with its canonical pronunciation, the phonetic segmentation found by MAUS and the automatically derived syllabification (the '<"ah>' represents a filled pause; phonetic coding in SAM-PA). As can be seen in this example the verb 'haben' (to have) is being reduced to a single syllable /ha:m/ in the following syllabification. In the three-syllable word 'eigenen' (own) the glottal stop (/Q/) and the final Schwa (/@/) are deleted by MAUS and a syllabic /n/ is segmented instead: /aIg@nn/. Since the final /n/ represents a syllable, the derived syllabification contains a syllable without a nucleus but a single syllabic /n/: /aI+g@n+n/. The same effect occurs in the words 'leisten' and 'müssen'. As a result the syllabification contains two 'syllable' types /n/ and /tn/ without any vocalic nucleus.

3. The BAStat Resource

BAStat consists of three main parts: phone, syllable and word statistics.
All resources described in this contribution are stored in tables of 7-bit ASCII characters to allow a maximum of interoperability between different operation systems. Numbers are either represented as integers or floating point numbers in the typical scientific notation, e.g. 1.54632e-03 (= 0.00154632). Special characters not included in the 7-bit ASCII table are represented in LaTeX coding (e.g. German Umlaut *ä = "a*). Columns of tables and matrices are always separated by a TAB sign (ASCII octal 011).
Each resource file starts with a four-lines header giving the name of the analyzed data set, the number of entity types[3], the number of entity tokens and a table header describing the following table.
In the following we will describe the details of each resource type.

3.1. BAStat Phone Statistics and Durations

The BAStat phone statistics provides information about duration and probabilities of single phones (monogram), phone bigrams and arbitrary phone sequences.

3.1.1. Phone Monograms

All phone segments belonging to a filled pause (hesitation) are excluded from the analysis, because these phones behave in a different way than phones embedded in spoken words. (For instance they may be exceedingly long.) Also, phone segments of duration more than 1sec are filtered because they most likely stem from errors in the automatic segmentation process.
Phone monograms are provided for both phoneme sets (basic and 6-set) and separately for each BAS corpus as well as for all corpora pooled together (denoted as 'TOTAL'). This allows the comparison of different domains as represented in the different speech resources.
Each phone monogram consists of a 31-column table with the following entries:

- phone label
- absolute count
- probability (this column adds up to 1)
- conditional probability that given the phone the phone is word-initial
- conditional probability that given the phone the phone is word-final
- conditional probability that given the phone the phone is word-internal,
- mean, standard deviation, 25%/50%/75% quantile of duration
- mean, standard deviation, 25%/50%/75% quantile of duration in word-initial/internal/final position
- mean, standard deviation, 25%/50%/75% quantile of duration of single phone words

The total number of analyzed phones is 2126634 derived from 557561 word tokens (development and test sets of VM corpora (VMSets, 2009) not included).

3.1.2. Phone Bigrams

Second order statistics or conditional probabilities $P(phon2|phon1)$ or diphone statistics or bigrams (all synonyms for the same thing[4]) are calculated in form of a squared matrix containing all un-smoothed conditional probabilities $P(phon2|phon1)$ where $phon1$ is the predecessor and $phon2$ is the successor. If n = number of entities, then the matrix is $(n + 3)$ columns times $(n + 2)$ rows since the first column contains an index to the entities and the !ENTER and !EXIT pseudo entities are added to the data to model entry and exit bigram probabilities of utterances.

[3] In case of phoneme statistics this rather trivial entry is replaced by the number of words.

[4] Please carefully distinguish these from the *joined* probability $P(phon1, phon2)$.

The rows define the predecessor phone *phon1* (indexed in the first column), while the columns define the successor phone *phon2* (not indexed but in the same order as the rows), one single element contains the linear conditional probability $P(col = phon2|row = phon1)$. Consequently, the elements of each row sum up to 1.
For technical reasons the last line indexed by $phon1 =$ $!EXIT$ also contains equally distributed probabilities summing up to 1 although !EXIT has no successor. Since the 2nd colum contains the probabilities for $phon2 =$ $!ENTER$ but !ENTER has no predecessor, all values in this column are set to zero, to avoid a distortion of the remaining values in the rows.
If other entries than the first column are zero, this means that the bigram combination was not seen in the input corpus. You may apply standard discounting techniques to obtain non-zero probabilities for these cases. Identical values following each other are indicated by an optional counter glued by a '*' symbol to the probability value[5].
As with the phone monogram statistics in the previous section BAStat provides bigram tables for all BAS corpora individually as well as for the joined data set, and for both phoneme sets.

3.1.3. Phone Sequences

Classical questions often asked are:
'What is the probability for the word-final phone /x/ to occur after the phone /y/?'
or:
'What is the probability for phone /x/ to occur after the phone /y/ and before the word-final phone /z/?'
or:
'What is the estimated probability for a sequence of four phones /xyzw/, even if we never observed such a phone sequence in the corpus?'
To answer these questions we can estimate the probability of the co-occurrence of an ordered pair of phones (y, x) or an ordered triplet (y, x, z) optionally combined with conditional probabilities of position within the word (see list of monogram table entries in section 3.1.1.). The probability of co-occurrence can be estimated by multiplying the first and second order statistics and ignoring higher order statistics

$$\begin{aligned} P(x,y) &= P(y|x)P(x) && (1)\\ P(x,y,z) &\approx P(z|y)P(y|x)P(x) && (2)\\ P(x,y,z,w) &\approx P(w|z)P(z|y)P(y|x)P(x) && (3)\\ &\dots && (4) \end{aligned}$$

where for instance $P(x, y, z)$ denotes the probability of the time-ordered occurrence of three entities x, y and z (in that order).
Examples:
*What is the probability estimate of /n/ following /E/ (e.g. 'M***en***sch')?*

$$P(n|E)P(E) = 0.00238476$$

*What is the probability estimate of a word-final syllable /vOYs/ (e.g. 'Kon***vois***')*

$$P(OY|v)P(s|OY)P(v)P(\text{word-final}|s) \approx 1.806e-09$$

Since these are merely rough estimates, caution should be taken to take these for absolute values. For instance it is probably not correct to state:
"The probability for the word final syllable /g@n/ is 7.627e-5!"
but we can say with some confidence that
"The probability for the syllable /g@n/ is higher in word-final (7.627e-5) than in word-internal position (2.026e-05)."
Likewise durational statistics can be estimated by sums of individual durational measures.

3.2. BAStat Syllable Statistics

This part of BAStat provides a collection of syllable data (raw data), duration and probabilities for single syllables (monogram) and syllable bigrams.

3.2.1. Collection of Syllable Segments

Since syllable analysis can be tricky, we provide the raw data as well as the statistics derived from it (see following sections). The raw data collection consists of a 7-column table describing one syllable in each line:

- the German SAM-PA coding of the syllable with a leading ' if the syllable was marked as lexically accented and a trailing '+' if the syllable is part of a function word
- the duration in secs
- the orthographic word
- the canonical pronunciation of the word coded in German SAM-PA
- the syllable position within the word in the form (Pos,Max), e.g. (2,5) is the second syllable in a 5 syllable word
- a file identifier of our internal database that allows us to find the corresponding recording
- the word position within the recording from which the syllable was taken (words counted starting with 0)
- the word duration in secs

The order of syllables is preserved in this list, that is the context can be derived from the preceeding and following lines. The tagging of lexical accent was taken from the (predicted) lexical accentuation in the lexical pronunciation form (4th column). There are two possible problems with that:
1. words with arbitrary lexical accentuation (e.g. 'umfahren'), which are fortunately very rare in German
2. the mapping from the canonical pronunciation form to the actual pronunciation fails, because syllables are deleted from a word with more than 2 syllables.
In unclear cases no syllable of the respective word is tagged as accented.
The tagging as a syllable stemming from a function word is based on the tagging in the lexicon as well. Since the definition for 'function word' is far from clear, we expect

[5] E.g. '1.234*2' equals '1.234 1.234 1.234'.

syllable	duration	word	pronunciation	position	file ID	word nr	word duration
'da:+	8.993750e-02	da	d'a:+	(1,1)	001/sp100001	19	8.993750e-02
'gIN	2.099375e-01	ging	g'IN	(1,1)	001/sp100001	20	2.099375e-01
's+	7.993750e-02	es	Q'Es+	(1,1)	001/sp100001	21	7.993750e-02
'al+	1.099375e-01	also	Q'alzo:+	(1,2)	001/sp100001	22	1.998750e-01
zo:+	8.993750e-02	also	Q'alzo:+	(2,2)	001/sp100001	22	1.998750e-01
'Um+	5.993750e-02	um	Q'Um+	(1,1)	001/sp100001	23	5.993750e-02
'das+	1.499375e-01	das	d'as+	(1,1)	001/sp100001	24	1.499375e-01
'taIl	1.899375e-01	Teilprojekt	t'aIl#proj"Ekt	(1,3)	001/sp100001	25	4.998125e-01
pro	1.699375e-01	Teilprojekt	t'aIl#proj"Ekt	(2,3)	001/sp100001	25	4.998125e-01
'jEkt	1.399375e-01	Teilprojekt	t'aIl#proj"Ekt	(3,3)	001/sp100001	25	4.998125e-01
ak	2.199375e-01	Akustik	Qak'UstIk	(1,3)	001/sp100001	26	6.098125e-01
'Us	1.799375e-01	Akustik	Qak'UstIk	(2,3)	001/sp100001	26	6.098125e-01
tIk	2.099375e-01	Akustik	Qak'UstIk	(3,3)	001/sp100001	26	6.098125e-01
'E:m	4.199375e-01	<"ahm>	Q'E:m	(1,1)	001/sp100001	27	4.199375e-01
'vi:6+	1.799375e-01	wir	v'i:6+	(1,1)	001/sp100001	28	1.799375e-01
'ha:m+	1.799375e-01	haben	h'a:b@n+	(1,1)	001/sp100001	29	1.799375e-01

Table 4: 15 syllables taken randomly from the BAStat raw syllables list.

a number of inconsistencies in cases where the semantical and syntactical usage of a word allow different interpretations. For instance the word 'da' (there) can be used in a functional way but also as a word carrying important content information. We observed that the annotators of the lexical sources tended to tag such arbitrary cases as a function word rather than a content word.

The BAStat raw syllable collection contains 1030588 syllable tokens representing 9210 syllable types[6] (derived from 689966 word tokens).

Table 4 shows an example of 15 syllables randomly taken from this list.

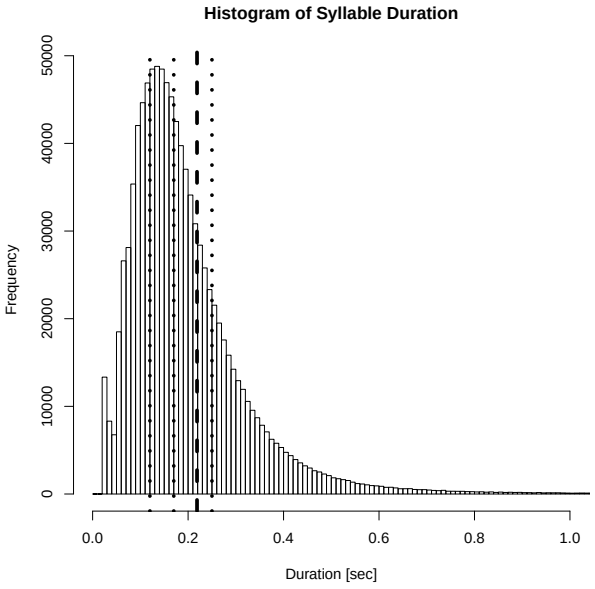

Figure 1: Syllable duration: histogram

[6]Lexically accented and non-accented syllables are counted separately; therefore this number is higher than the number of syllable types (6397) in the syllable monogram and bigram.

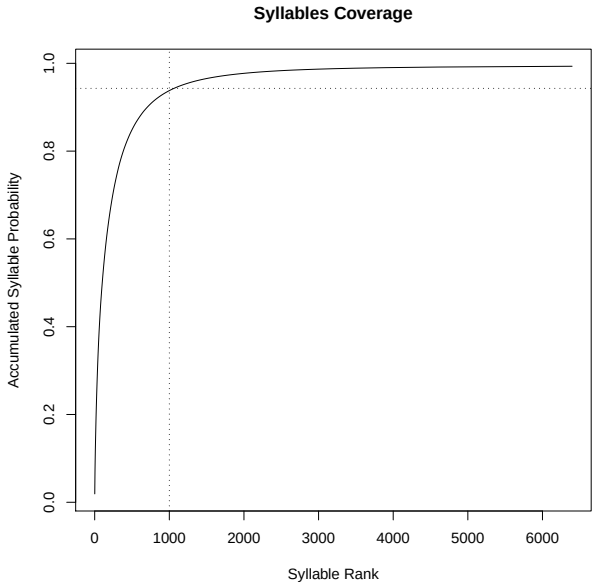

Figure 2: Syllable coverage: 94,4% of the analyzed speech corpora are covered by the top 1000 most probable syllables.

3.2.2. Syllable Durations and Monograms

Figure 1 shows the histogram of syllable durations derived from all collected syllable segments. The dashed line represents the mean of duration at 0,21sec while the three dotted lines mark the 25/50/75% quantiles at 0,12/0,17/0,25sec. The data resemble closely the notion that the average syllable length is in the region of 0,2sec for most languages of the world (e.g. reported for American English and Japanese in (Arai & Greenberg, 1997)). The histogram converges to zero around a length of 1sec. Beyond that duration outliers are found that represent either unnatural sound lengthening, as in extremely lengthened filled pauses, or segmentation errors caused by the MAUS system.

Rank	Syl	Count	P(Syl)	P(Fun|Syl)	P(LA|Syl)	P(WI|Syl)	P(WF|Syl)	P(WM|Syl)	Mean(Dur)
1	ja:	19306	1.910e-02	0.000e+00	3.309e-02	8.960e-03	9.841e-04	2.346e-02	2.746e-01
2	IC	18267	1.807e-02	9.637e-01	6.021e-04	1.423e-03	3.366e-02	1.587e-03	1.254e-01
3	das	16420	1.624e-02	9.982e-01	0.000e+00	1.948e-03	6.090e-05	0.000e+00	1.881e-01
4	n	16191	1.602e-02	4.343e-01	6.176e-05	6.176e-05	7.180e-01	2.983e-02	6.618e-02
5	dan	11181	1.106e-02	9.764e-01	1.788e-04	3.246e-02	1.788e-04	0.000e+00	2.165e-01
6	g@	11087	1.097e-02	1.229e-01	0.000e+00	5.465e-01	2.592e-01	1.914e-01	1.161e-01
7	tn	10200	1.009e-02	5.000e-03	0.000e+00	0.000e+00	9.831e-01	1.686e-02	1.478e-01
8	@	10156	1.004e-02	1.258e-01	2.461e-03	1.900e-02	8.856e-01	9.531e-02	9.218e-02
9	da:	9648	9.546e-03	9.770e-01	1.627e-02	3.917e-02	0.000e+00	1.865e-03	1.654e-01
10	di:	8465	8.375e-03	9.868e-01	1.110e-02	1.813e-01	2.362e-04	6.379e-03	1.387e-01

Table 5: Top 10 ranking syllables from the BAStat raw syllables list.

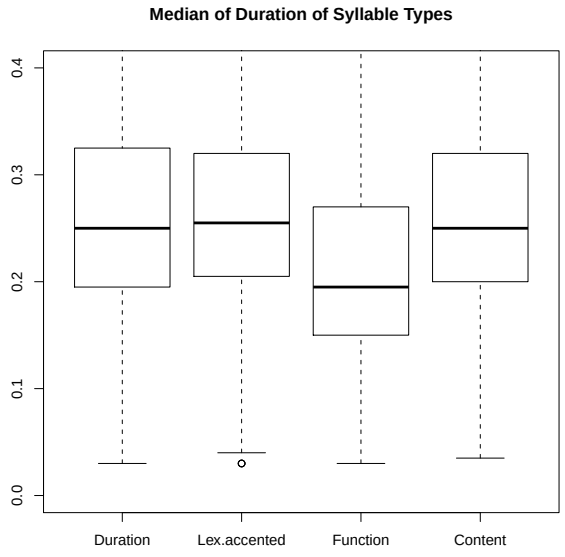

Figure 3: Syllable duration: distribution of duration medians of different syllable types

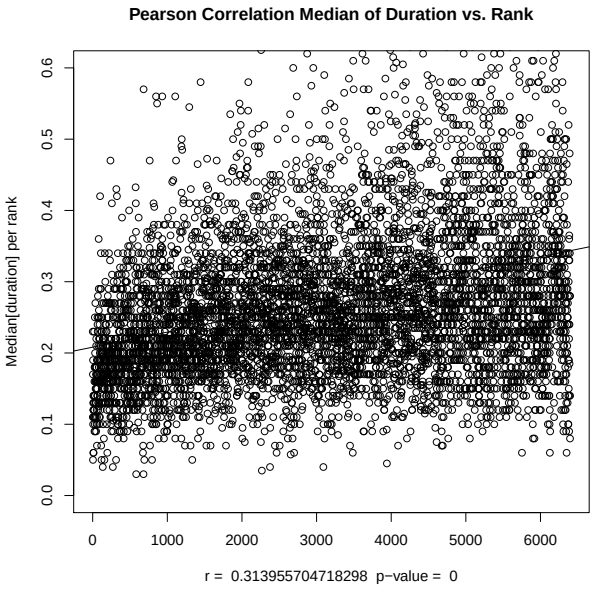

Figure 4: Syllable duration: median of duration across syllable type rank

To eliminate as many segmentation errors as possible the raw syllable list was filtered for syllables that have a duration longer than 1 sec[7] (0.657% of all syllables). Then we filtered the lexical accentuation and function word markers, so that accented and un-accented syllables as well as syllables stemming from a function word or a content word are treated the same. From the remaining syllable corpus we calculate a 50-column table containing 6397 syllable types together with the following information:

- syllable rank
- syllable coding in German SAM-PA (*syl*)
- total count
- probability P(*syl*)
- conditional probability for a content word P(Con|*syl*)
- Conditional probability for a function word P(Fun|*syl*)
- Conditional probability for lexical accentuation P(LA|*syl*)
- Conditional probability being word-initial P(WI|*syl*)
- Conditional probability being word-final P(WF|*syl*)
- Conditional prob. being word-internal P(WM|*syl*)
- duration (mean,SD,25/50/75-quantiles)
- duration in content/function words
- duration in lexically accented position
- duration in word initial/internal/final position
- duration in single syllable words

Table 5 lists the first 10 columns of the 10 top ranking syllables in BAStat. It is interesting to note that the German syllable /ja:/ (the German affirmative 'ja') is the most frequent syllable in conversational speech. It is followed by /IC/ (1st person singular pronoun 'ich' pronounced without a glottal stop). So it seems that Germans mostly talk affirmative about themselves.

[7]This might also eliminate some of the syllables representing filled pauses; insofar the syllable monogram and bigram statistics should not be used in the context of studies about filled pauses.

word	pron.	count	P(*pron*|*word*)
Abend	Qa:b@nt	20	6.688e-02
Abend	Qa:b@nh	1	3.344e-03
Abend	a:mn	2	6.688e-03
Abend	a:bm	31	1.036e-01
Abend	a:bn	1	3.344e-03
Abend	Qa:bmt	2	6.688e-03
Abend	Qa:b@n	9	3.010e-02
Abend	Qa:bm	3	1.003e-02
Abend	a:b@nt	51	1.705e-01
Abend	a:b@nh	2	6.688e-03
Abend	a:mt	114	3.812e-01
Abend	a:b@n	34	1.137e-01
Abend	Qa:mt	27	9.030e-02

Table 6: Examples from the BAStat pronunciation statistics: the word 'Abend' (evening). /Q/ is the glottal stop.

Figure 2 plots the accumulated probability across the ranking of syllable types. The first 1000 top ranked syllables cover over 94,4% of the analyzed corpus speech (dotted lines). 25,6% of the top 1000 ranking syllables are stemming from function words, while only 13,6% of all syllable types are from function words. This concentration in the high-frequent range is also the reason that 40,6% of all syllable tokens are uttered in function words.
High-frequent syllables are expected to be produced faster than low-frequent syllables. On the other hand syllables carrying a lexical accent are expected to be pronounced longer than non-accented syllables.
Figure 3 shows four box-plots for the distribution of the medians of the duration of each syllable type. That is, each syllable type is represented by one data point in this distribution and the probability of the syllable type is not considered here. Contrary to our expectation the distribution of lexically accented syllables in words with more than one syllable ('Lex.accented') does not deviate from the distribution over all syllable types ('Duration'). However, as expected the distribution of syllables derived from function words shows significant smaller durations ('Function') than that of syllables derived from content words ('Content').
In Figure 4 the median duration of syllables types is plotted against the rank (the probability) of the syllable type. There is slight positive linear correlation, but the Pearson correlation is only 0,31.

3.2.3. Syllable Bigrams

The syllable bigram statistics is provided for the same filtered set of syllables as in the monogram statistics. Format and method follow the same schema as used in the phoneme bigram statistic (see above).

3.3. BAStat Word Statistics

The BAStat word statistics is structured into duration and probabilities of single word types (monogram), word bigrams and conditional probabilities of word pronunciations. Since the number of word tokens (689966) is rather small in relation to the number of word types (16426), the word statistics of BAStat cannot be considered as being representative for spoken German. We hope to expand this section in the future by acquiring larger corpora of transcribed conversational German.
Word statistics are given for all word types and the following non-words: *silence interval, articulatory noise* (e.g. cough), *background noise, laughing, breathing*, seven types of *filled pauses, spellings* and a garbage model for *non-intelligible speech parts*.

3.3.1. Word Monograms

The monogram for words provides the following information per word type:

- the orthographic word form and canonical pronunciation in SAM-PA including a marker for content/function word
- count and probability
- the mean duration
- the (canonical) number of syllables

3.3.2. Word Bigrams

The word bigram consists of a simple matrix with unsmoothed conditional probabilities for word tuples. Format and method follow the same schema as used in the phoneme monograms (see above).

3.3.3. Word Pronunciation Statistics

Based on the phonetic segmentation we can derive 28754 different pronunciation forms for the 16431 word types in BAStat. The BAStat word pronunciation statistics lists these pronunciation forms coded in SAM-PA together with their orthographic form, count and conditional probability. Similar resources have been successfully used in automatic speech recognition in form of probabilistic pronunciation lexica (e.g. in (Schiel, 1998)). As an example we list some of the entries for the word 'Abend' (evening) in Table 6. For instance the canonical pronunciation /Qa:b@nt/ is with 20 tokens much less frequent than the reduced mono-syllabic forms /a:mt/ and /Qa:mt/ (141 tokens).

	CELEX	BAStat
word tokens	5002442	689966
word types	84173	16426
syllable tokens	9062607	1030588
syllable types	7030	9210 (6397)

Table 8: Word and syllable counts in CELEX and BAStat

4. Comparison with CELEX

Since BAStat is rather unique in being based solely on empiric speech recordings of conversational speech, it is interesting to compare the statistical data of BAStat to existing resources based on textual data, namely the CELEX lexical database (Baayen et al., 1995).
"CELEX is the Dutch Centre for Lexical Information. It was developed as a joint enterprise of the University of Nijmegen, the Institute for Dutch Lexicology in Leiden, the Max Planck Institute for Psycholinguistics in Nijmegen, and the Institute for Perception Research in Eindhoven. ... CELEX is now part of the Max Planck Institute

CELEX	di:	de:r	**g@**	**t@**	QUnt	QIn	b@	t@n	tsu:	**das**	QaI	fEr	g@n	n@	d@n	de:n
BAStat	ja:	IC	**das**	n	dan	**g@**	tn	@	da:	di:	**t@**	s	d6	vi:6	vi:	zi:

Table 7: Top ranking syllables in CELEX and BAStat.

for Psycholinguistics." (*quoted from the CELEX CD-ROM, README*)

The German part of CELEX contains no empirically based phonetic information about phones and syllables. However, it contains phonological data for phonemes and syllables based on large collections of German texts (derived from the archives of the 'Institut der Deutschen Sprache', Mannheim, Germany).

Table 8 compares CELEX and BAStat with regard to word and syllable types and tokens. The ratio of words types against word tokens is lower in CELEX (1,7%) than in BAStat (2,4%); this is probably caused by the insufficient number of word tokens in BAStat: while the number of word types in CELEX is probably nearly converged, in BAStat the number of word types will probably still grow with increasing corpus size.

Because of the smaller amount of word types in BAStat we would expect a proportional smaller number of syllable types, but this is not the case: the number of syllable types in BAStat exceeds the number in CELEX. The reason is probably that the phonetic variation of syllables produces more syllable forms than in the phonological paradigm of CELEX, where each word token is always assigned to the same (lexical) syllables.

The statistic of syllable types also differs considerably: in Table 7 we compare the top 15 highest ranking syllables from CELEX and BAStat in descending ranking order[8]. The few overlaps in both ranking sets are printed in bold face. If we look at the 1000 top ranked syllables in both resources, we find an overlap of merely 47,5%.

This comparison is not entirely justified since in the case of CELEX the syllabification was done phonologically while in BAStat it is based on the phonetic transcript. For instance the syllabic nasal /n/ is very high in the ranking of BAStat but does not even appear in the CELEX syllable type list. Nevertheless, the comparison shows that phone or syllable statistics from a lexically based resource differ considerably from conversational speech and might not be suitable for experimental setups dealing with spoken language.

5. Conclusion

We presented a new type of language resource BAStat, namely statistical data derived from large primary resources of spoken German. These data are useful for linguists as well as language engineering dealing with statistical models of speech production or speech perception. All LRs described here are available for free from the BAS web site *www.bas.uni-muenchen.de/Bas*. Finally we would like to encourage LR providers of other languages than German to provide similar data for the scientific community.

[8]The CELEX phonologic coding was mapped to German SAM-PA here and word initial glottal stops were inserted.

6. Acknowledgments

This work was partly made possible by the funding of the primary resources *Verbmobil* and *SmartKom* by the *Bundesminister für Bildung und Forschung*, and the speech corpus *RVG1* by the *Bell Laboratories*. The author thanks all colleagues involved in these projects.

7. References

Arai, T. & Greenberg, St. (1997). The Temporal Properties of Spoken Japanese are similar to those of English. *Proc. of the Eurospeech. Rhodes, Sept 1997. pp. 1011-1014.*

Baayen, R.H. & Piepenbrock, R. & Gulikers, L. (1995). The CELEX Lexical Database (CD-ROM). *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA*.

BAS - Bavarian Archive for Speech Signals, (2010). *http://www.bas.uni-muenchen.de/Bas*. cited Feb 2010.

Burger, S. & Weilhammer, K. & Schiel, F. & Tillmann, H.G. (2000). Verbmobil Data Collection and Annotation. In: W. Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin, Heidelberg.

Draxler, Chr. & Jänsch, K. (2004). SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software. *Proc. of the IV. International Conference on Language Resources and Evaluation, Lisbon, Portugal*.

Levelt, W.J.M. (1989). Speaking - from intention to articulation. *MIT Press*.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. *J. Bybee and P. Hopper (eds.): Frequency effects and the emergence of lexical structure. John Benjamins, Amsterdam, pp. 137-157*.

Schiel, F. (1999). Automatic Phonetic Transcription of Non-Prompted Speech. *Proc. of the ICPhS. San Francisco, August 1999. pp. 607-610*.

Schiel, F. & Kipp, A. & Tillmann, H.G. (1998). Statistical Modeling of Pronunciation: It's not the Model, it's the data. *Proc. of the ESCA Tutorial and Research Workshop on 'Modeling Pronunciation Variation for Automatic Speech Recognition, pp. 31-36*.

VMSets - Verbmobil Training, Development and Test Set Definition (2009). *ftp://ftp.bas.uni-muenchen.de/pub/BAS/VM/SETS*. cited Feb 2010.

Young, St. (1995). The HTK Book. Revised 1999, University of Cambridge.