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.

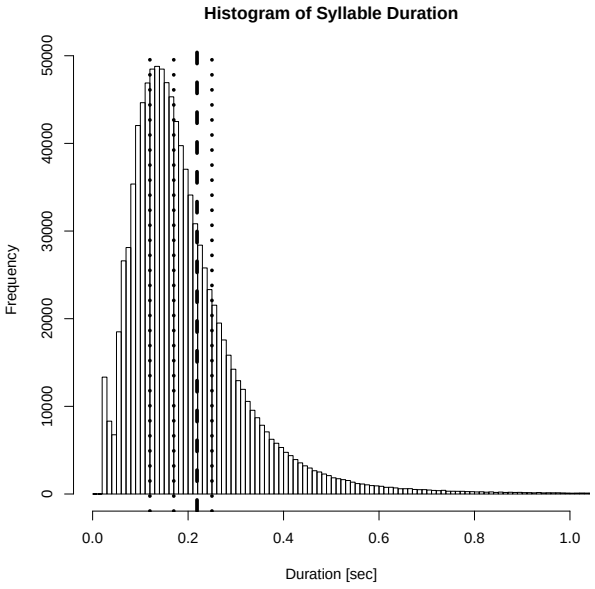

Figure 1: Syllable duration: histogram

[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.

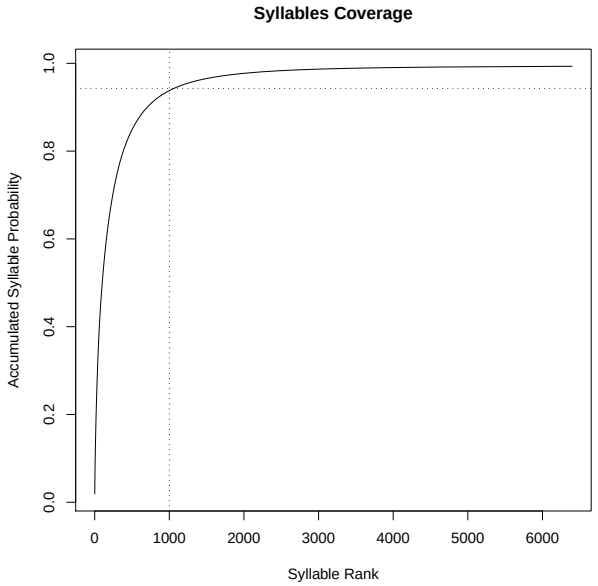

Figure 2: Syllable coverage: 94,4% of the analyzed speech corpora are covered by the top 1000 most probable syllables.

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.

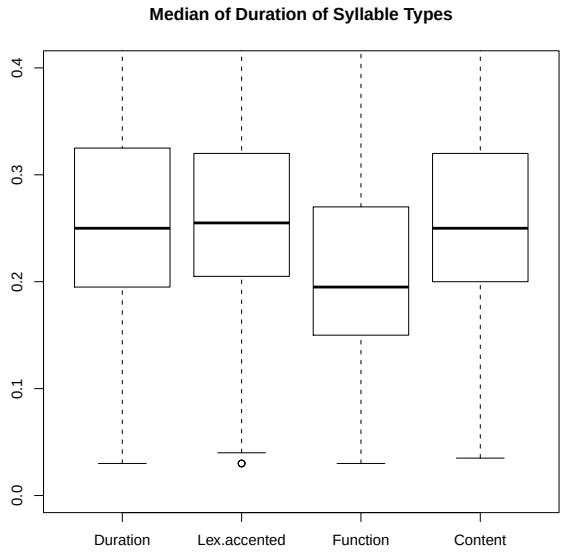

Figure 3: Syllable duration: distribution of duration medians of different syllable types

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.