Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Spoken dialogue data collected in the Waxholm project

Bertenstam, J. and Blomberg, M. and Carlson, R. and Elenius, K. O. E. and Granström, B. and Gustafson, J. and Hunnicutt, S. and Högberg, J. and Lindell, R. and Neovius, L. and Nord, L. and de Serpa-Leitao, A. and Ström, N.

KTH Computer Science and Communication

journal: STL-QPSR
volume: 36
number: 1
year: 1995
pages: 049-074

http://www.speech.kth.se/qpsr

Spoken dialogue data collected in the WAXHOLM project

*Johan Bertenstam, Mats Blomberg, Rolf Carlson, Kjell Elenius, Björn Granström, Joakim Gustafson, Sheri Hunnicutt, Jesper Högberg, Roger Lindell, Lennart Neovius, Lennart Nord, Antonio de Serpa-Leitao, and Nikko Ström**

Abstract

This paper describes the data collected in Wizard of Oz experiments in a spoken dialogue system, WAXHOLM, that provides information on boat traffic in the Stockholm archipelago. The data consist of utterance-length speech files, their corresponding transcriptions, and log files of the dialogue sessions. Apart from the spontaneous dialogue speech, the speech material also comprise recordings of phonetically balanced reference sentences uttered by all 66 subjects. The recording procedure is described as well as some characteristics of the speech data and the dialogue.

Introduction

Our research group at KTH is currently building a generic system in which speech synthesis and speech recognition can be studied and developed in a man-machine dialogue framework. The system is designed to facilitate the collection of speech and text data that are required for development (Blomberg et al., 1993; Carlson, 1994).

In this paper, we describe the data that have been collected in Wizard of Oz experiments with the demonstrator application, WAXHOLM, a spoken dialogue system providing information on the boat traffic in the archipelago of Stockholm. The wizard has only been serving to replace the speech recognition module in the system, a fact which has been known to the subjects.

The collected data consist of utterance-sized speech files that are stored together with the text entered by the wizard and the corresponding phonetic labels. A complete log of the dialogue session is also stored. The acoustic-phonetic database also consists of phonetically rich reference sentences uttered by all subjects. Thus, the data collected provide a good basis for studies of spontaneous speech and related phenomena, as well as read speech material suitable for inter-speaker comparisons, speaker adaptation experiments, etc.

In the first section, we give a brief description of the demonstrator application WAXHOLM. Further, we describe the Wizard of Oz experiments and the recording conditions. Some characteristics of the subjects, word statistics, the labelling of the speech files, and the phonetic composition of the data, are presented in the three following sections. Special care has been taken to label extralinguistic sounds in the recorded speech material and some of these results are reported. Finally, the last section provides an analysis of the dialogue. Translations of the reference sentences and the dialogue scenarios are enclosed in appendices.

---

* Names in alphabetic order

System description

The demonstrator application

The demonstrator application, which we call WAXHOLM, gives information on boat traffic in the Stockholm archipelago. It references timetables for a fleet of some twenty boats from the Waxholm company, which connects about two hundred ports.

Besides the speech recognition and synthesis components, the system contains modules that handle graphic information, such as pictures, maps, charts, and timetables (Fig. 1). This information can be presented to the user as a result of the user initiated dialogue. The application has great similarities to the ATIS domain within the ARPA community, the Voyager system from MIT (Glass et al., 1994), and similar tasks in Europe, for example SUNDIAL (Peckham, 1993), the system for train timetables information developed by Philips (Aust et al., 1994; Oerder & Aust, 1994), and the Danish Dialogue Project (Dalsgaard & Baekgaard, 1994).

The possibility of expanding the task in many directions is an advantage for our future research on spoken dialogue systems. In addition to boat timetables, the database also contains information about port locations, hotels, camping grounds, and restaurants in the Stockholm archipelago. This information is accessed by SQL, the standardized query language (Gustafson, 1992). An initial version of the system based on text input has been running since September 1992.

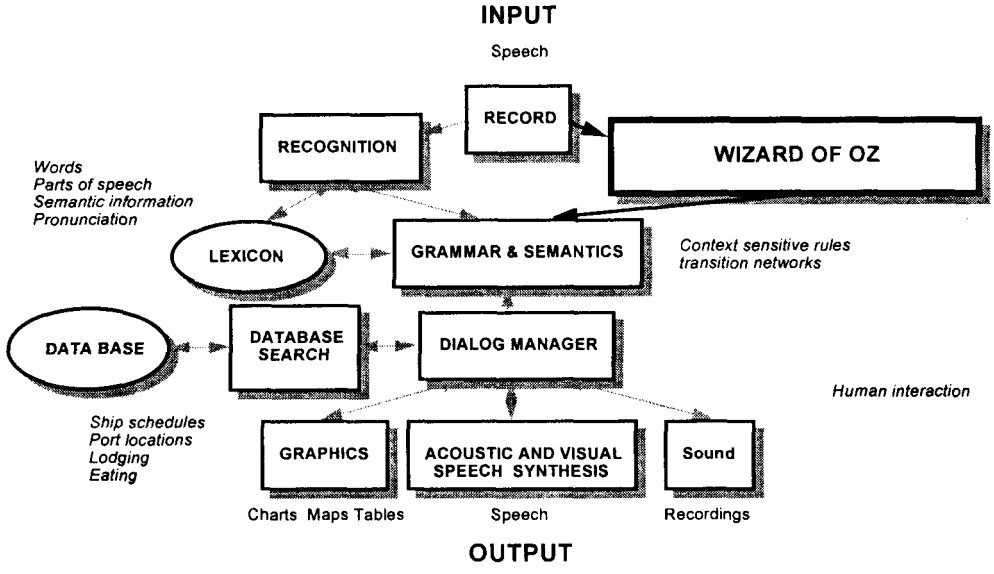

Fig. 1. The modules of the WAXHOLM spoken dialogue system. The function of the Wizard of Oz used for the data collection is indicated.

Speech recognition and lexical search

The speech recognition component, which has been integrated in the system, handles continuous speech with a vocabulary of about 1000 words. The work on recognition has been carried out along two main lines: artificial neural networks (Elenius & Takács, 1990; Elenius & Blomberg, 1992; Elenius & Tråvén, 1993), and a speech production oriented approach (Blomberg, 1991). Since neural nets are general classification tools, it is quite feasible to combine the two approaches.

The frame-based outputs from the neural network form the input to the lexical search. There is one output for each of the 40 Swedish phonemes used in our lexicon. Each word in the lexicon is described on the phonetic level and may include alternative pronunciations of each word. The outputs are seen as the a posteriori probabilities of the respective phonemes in each frame. An A*, N-best search has been implemented using a simple bigram language model (Ström, 1994).

Synthesis

For the speech output component, we have chosen our multi-lingual text-to-speech system (Carlson et al., 1991). The system is modified for this application. The application vocabulary has been checked for correctness, especially considering the general problem of name pronunciation (Gustafson, 1994).

Since the recognition and synthesis modules have similar need of semantic, syntactic and pragmatic information, the lexical information is, to a great extent, shared. In dialogue applications, such as the WAXHOLM, we have a better base for prosodic modeling compared to ordinary text-to-speech, since, in such an environment, we will have access to much more information than if we used an unknown text as input to the speech synthesiser (Bruce et al., 1994).

The speech synthesis has recently been complemented with a face-synthesis module. Both the visual and the acoustic speech synthesis are controlled by the same synthesis software (Beskow, 1995).

The natural language component

Our work on the natural language component is focused on a sublanguage grammar, a grammar limited to the particular subject domain -- that of requesting information from a travel database. Our parser, STINA, is knowledge based and is designed as a probabilistic language model. It contains a context-free grammar, which is compiled into an augmented transition network (ATN). Probabilities are assigned to each arc after training. Characteristics of STINA are a stack-decoding search strategy and a feature-passing mechanism to implement unification.

Dialogue management based on grammar rules and lexical semantic features is implemented in STINA. The notation to describe the syntactic rules has been expanded to cover some of our special needs to model the dialogue. The STINA parser is running with two different time scales corresponding to the words in each utterance and to the turns in the dialogue. Topic selection is accomplished based on probabilities calculated from user initiatives.

We have found it very profitable to handle both the regular grammar analysis and the dialogue control with the STINA parser. The same notation, semantic feature system and developing tools can be shared. The rule-based probabilistic approach has made it reasonably easy to implement an experimental dialogue management module. STINA is described in more detail in Carlson & Hunnicutt (1995).

Wizard of Oz experiments

The data has been collected using the complete system. That is, all system modules, except the speech recognition, which has been replaced by a Wizard of Oz, have been in use since the start of the data collection phase. This section describes the experimental set-up and procedure.

The subjects are seated in an anechoic room in front of a display with a uni-directional cardioid electret condenser microphone mounted on top of it. The wizard is seated in an adjacent room facing two screens, one displaying what is shown to the subject and the other displaying system information. The subjects were all aware of the fact that the wizard replaced the speech recognition. Each utterance is digitally recorded at 16 kHz and stored together with its respective text and, later, label file. The system provides the wizard with a playback function to facilitate correct text entry. All system information is logged during the data collection sessions making it possible to replay the dialogue.

An experimental session starts with a system introduction presented in text on the screen. The text is also read by speech synthesis, thus permitting the subject to adapt to the synthetic voice. The subject practices the push-to-talk procedure reading a sound calibration sentence and a few test sentences followed by eight phonetically rich reference sentences (Appendix A). The reference sentences are designed to cover both common and uncommon phonemes. However, /ʏ/ and retroflex /l/ are not represented. Each subject is provided with three information retrieval scenarios. Fourteen different scenarios are used altogether. The first scenario, which is the same for all subjects, is presented below.

> **Scenario 1**
> It's a beautiful summer day and you are in Stockholm. You decide that you'd like to go to Vaxholm[1].
> Your task is to find out when the boats leave for Vaxholm this evening.

For a complete list of translated scenarios, see appendix B. The scenarios are presented to the subject both as text on the screen and with speech synthesis. The subject is instructed that the scenarios are starting points for further exploration of the system capabilities. It is not emphasised that the subject should regard the scenario as a task to be completed as fast as possible, encouraging the subject to use the system beyond the scope of the scenario. Utterances resulting in extremely low or high sound

[1] "Vaxholm" is a town in the archipelago. The spelling "Waxholm" is used in the name of a boating company and for some boats in their fleet.

levels are automatically detected by the system, which prompts the subject for a repetition.

After the experimental session, the subject fills in a questionnaire with questions about weight, height, age, profession, dialect, speaking habits, native tongue, comments about the experiment, etc.

So far, some 1900 dialogue utterances have been recorded containing 9200 words. The total recording time amounts to 2 hours and 16 minutes, one third of which is labelled as pause. One fourth of the recording time pertains to the calibration and reference sentences.

Speaker characteristics

Initially, 66 different subjects, of which 17 are female, have participated in the first phase of the data collection. The majority of the subjects, 43, were 20-29 years old while 4 were 30-39, 10 were 40-49 and 9 were more than 50 years old. A few voices are trained.

Most subjects are department staff or undergraduate students from the school of electrical engineering and computer science (Fig. 2.)

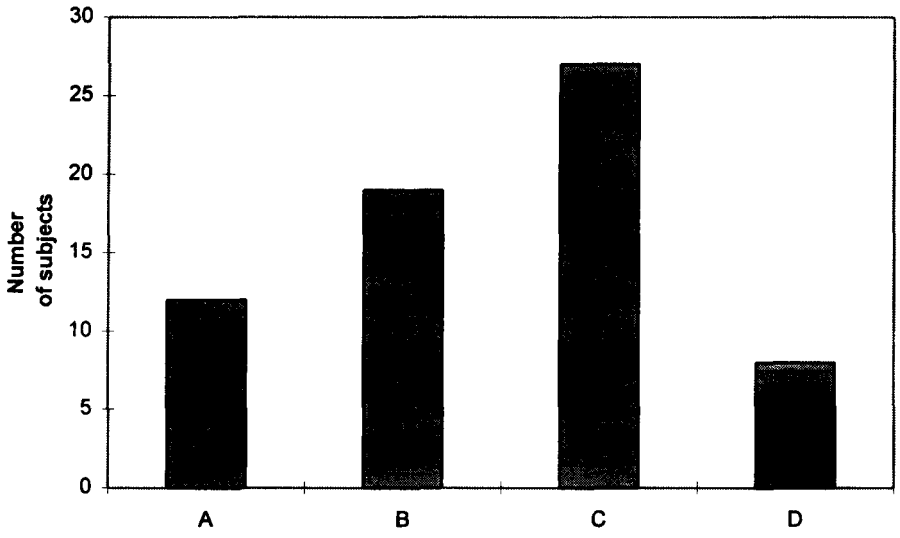

Fig. 2. Professional affiliation of subjects. A: project staff, B: other department staff, C: undergraduate students from KTH, D: others.

The subjects have given a written description of how they sound in terms of dialects or accents. Many speakers do not seem to appreciate the full value of their deviations from standard Swedish. That is, although some subjects have a discernible non-standard Swedish dialect they did not indicate it in the questionnaire.

At least 50% of the speakers of standard Swedish have a Stockholm area dialect. Table 1 displays the number of speakers whose dialect is considered as being non-

standard Swedish by an expert phonetician and the corresponding city, county or country affiliation.

A few foreign accents are represented: Portuguese, Finnish, Norwegian, and American English. There are three speakers from the city of Gothenburg and one from Eskilstuna, Uppsala, and Norrköping, respectively. The dialects of the Norrland and Dalarna regions are represented, as well as a less well defined "countryside" accent.

Table 1. Speakers that don't speak standard Swedish. Columns two and three contain the number of male and female speakers from the city, region or country designated in column one.

Geographical affiliation of accent	Males	Females
Gothenburg	1	2
Eskilstuna	2	-
Dalarna	1	-
Norrköping	-	1
Västmanland	1	-
Värmland	1	-
Norrland	1	-
Finland	1	-
Norway	1	-
Portugal	1	-
USA	-	2

Speaking styles

There is definitely some spread in the speaking styles. The introductory test and calibration sentences sound very different from the following dialogue for some speakers. Thus, many subjects alter speaking style when switching from read to spontaneous speech. Although the dialogue speech has a spontaneous quality it is also characterised by the task the subjects are performing, e.g., effecting the speaking rate.

Figure 3 shows a scatter diagram of mean phoneme durations for the read sentences and the dialogue corpora. Since some phonemes are very rare or even non-existing in the read speech, only phonemes that have more than 50 occurrences in the read speech are included. Most phonemes have a longer duration in the dialogue speech, as is indicated by the regression line which has a slope of 0.85 ($r = 0.91$). All phonemes below the dotted diagonal are longer in the dialogue speech. The phonemically long vowels seem to be especially prolonged in the spontaneous speech. The main reason for this may be that the subjects were slowed down since they were solving a task. Another sign of this is that the mean length of the sentence internal pauses were 216 ms for the dialogue speech compared to 131 ms for the read part. A further reason may be that some speakers, although they knew they were talking to a Wizard, rather performed as if talking to a computer, speaking slowly and elaborately, perhaps

because the system had not understood sentences that seemed very simple to the subject. Still another factor for the articulate speech might be the rather unnatural setup, with the subjects alone in a soundproof booth with a microphone and a terminal.

In particular, one male and one female speaker produce very peculiar speech. They overarticulate in an unnatural way, word-by-word. In one of these cases there was a problem with the sound level and the system repeatedly urged the subject to speak louder during the experimental dialogue session. However, in the other case, the style clearly reflects the subject's attitude to man-machine communication.

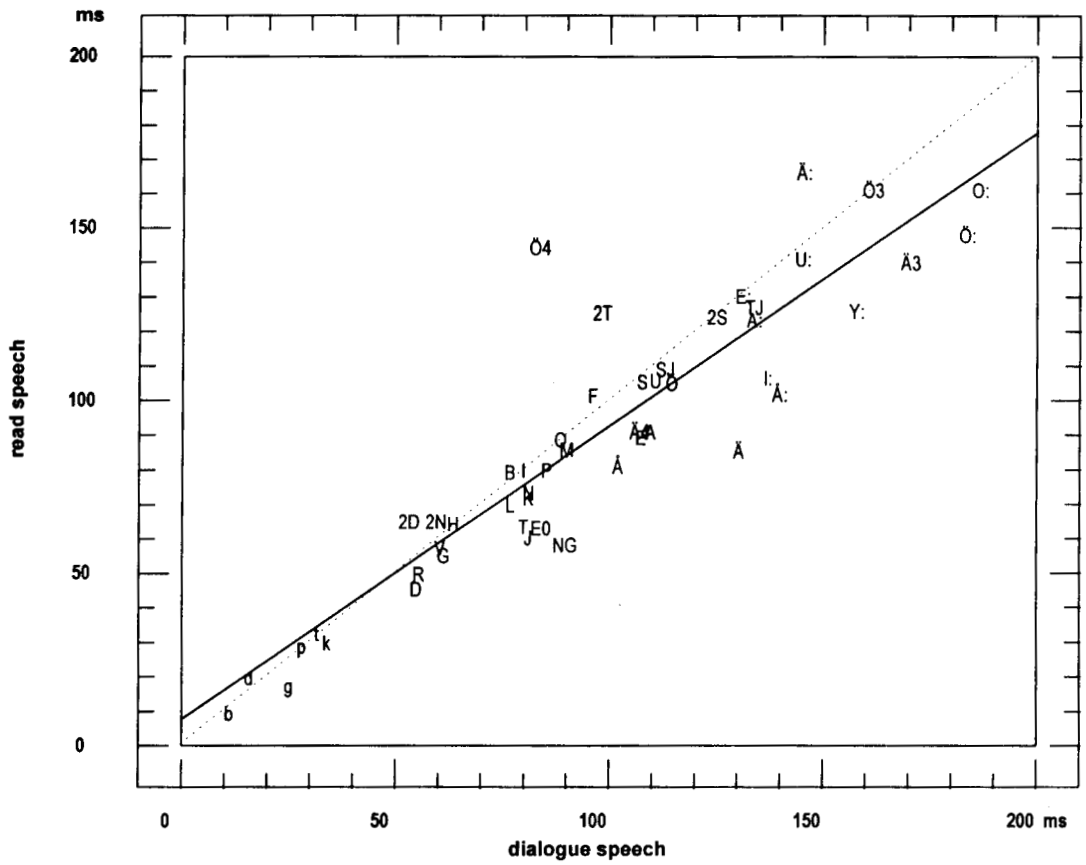

Fig. 3. Scatter plot of the mean phoneme length in ms for the dialogue and read part of the WAXHOLM speech data base. The phonetic symbols are explained in Table 5.

Moreover, intra-subject variations in tempo and loudness are sometimes quite considerable. Extralinguistic phenomena, such as hesitations, pauses, coughs, and sounds of breathing, also contribute to the speaking style. These sounds are discussed separately in a following section.

The dialogue contains spontaneous reactions reflected in the subjects choice of words, the occurrence of extralinguistic sounds and emotional cues in the voice character. As many as five of the subjects have what could be considered "laughter in their voice." Explicit and muffled laughter occur in a few cases. Anger can be heard in

some recordings and also more subtle indications of irritation or aggressiveness, even though the subject does not always express his or her emotional state in words.

Voice characteristics

A voice evaluation was made of the 66 speakers. It turned out that as many as thirty-three speakers have a creaky voice character, especially in phrase endings, while eleven also have some grating. The voicing onset is hard in twenty-seven of the voices.

In five of the voices, the voice register is unstable and a voice break occurs in two. The general term "hoarse" is used for ten of the voices. Twelve of the voices are considered to be carrying very little energy, while five are considered to be strained. These aspects, listed in Table 2, are not pathological deviances, but could influence the performance of a speech recognition system.

One of the female speakers had a special tremor character in her voice. Combined with a guarded speaking style, this gives the voice a very uncertain character. The unstable quality gives the voice a grating character. Two of the women have a girlish tone of voice and one male speaker has a distinct gloomy character to his voice.

Some of the young male speakers use a very low fundamental frequency at the bottom of their registers. Examples of rough, tense, weak, and screaming voice characters are also present in the speech material. Only one speaker sounded as if he had a cold. On the whole, most voices are very good, and the tone is pleasant in most cases.

Table 2. Some common voice characteristics in the speech material. The right column indicates the number of subjects to which the voice characteristic applies.

Voice characteristics	Number of subjects
grating	11
creak	33
hard glottal attacks	27
voice breaks	2
unstable register	5
hoarse	10
laughter in the voice	5
poor timbre	12
strained voice	5
girlish voice	2

Word statistics

Word frequencies

The word frequency ranking of the WAXHOLM corpus is naturally different from that of large corpora collected from written text. In Table 3, it is compared to a written corpus of 150 million Swedish words - the KTH corpus - that has been collected mostly for the purpose of language modelling in the context of speech recognition. It

Table 3. The most frequent words in the WAXHOLM speech database plus some additional ones. An approximate English translation is given as well as the frequency rank in the KTH text corpus of 150 million words.

WAXHOLM rank	*WAXHOLM frequency*	*KTH rank*	*KTH frequency*	*Word*	*English translation*
1	601	18	1248597	jag	I
2	438	13	1537177	till	to
3	395	639	18502	åka	~go
4	373	74	199430	vill	want
5	355	7	2394192	på	on
6	306	4	2902952	det	it
7	295	82	181863	går	~leave
8	260	27	573911	från	from
9	250	8220	1327	Vaxholm	Vaxholm
10	208	2	4261937	i	in
11	186	72	207407	finns	is there (exists)
12	165	148	89468	Stockholm	Stockholm
13	150	5	2543912	en	a
14	137	643	18461	tack	thank you
15	136	31	504248	kan	can
16	135	32	500352	när	when
17	134	20	1066758	var	where
18	115	1471	8552	hotell	hotel
19	113	486	25427	klockan	clock (time)
20	94	10077	1031	ikväll	tonight
21	94	3537	3392	båtar	boats
22	91	39	326109	efter	after
23	84	7502	1468	båtarna	the boats
24	83	28	556489	man	one (person)
25	79	369	34469	vilka	what
26	72	2301	5382	båt	boat
27	72	207	64214	ligger	is located
28	69	6852	1630	skärgården	the archipelago
29	69	1620	7754	fredag	Friday
30	65	71	217583	hur	how
...	...	...	...	...	...
35	54	1	4428106	och	and
...	...	...	...	...	...
-	2	0	0	vartifrån	where from
-	1	0	0	tältningsplats	place for tenting
-	1	0	0	avgångsplatser	places of departure
-	1	0	0	båttid	boat time

consists mostly of newspaper text, but also of text from novels, educational books and almost 5 million words from speeches in the Swedish parliament. The total number of unique words is 1.88 million compared to the about 600 different words in the WAXHOLM dialogue corpus.

The top ranking WAXHOLM words almost make up the sentence "jag vill åka till Vaxholm" (I want to go to Vaxholm), which indicates the influence of the domain and the scenarios given to the subjects. Though many of the most frequent words in the large corpus have a high rank in the dialogue corpus, the most frequent one, *och* (and), only has rank 35. It is also interesting to note that although there are only 600 unique words in the dialogues, 9 of them cannot be found in the large KTH corpus. Five of these are names of small ports, and the other four are listed at the bottom of the table. Of these, *vartifrån,* is a spoken language variant of *VARIFRÅN* (where from) while the others may be seen as very typical for the Vaxholm domain.

As shown in Figure 4, the 10 most frequent words cover 35% of all words in the dialogues and the 200 most frequent words cover 92%.

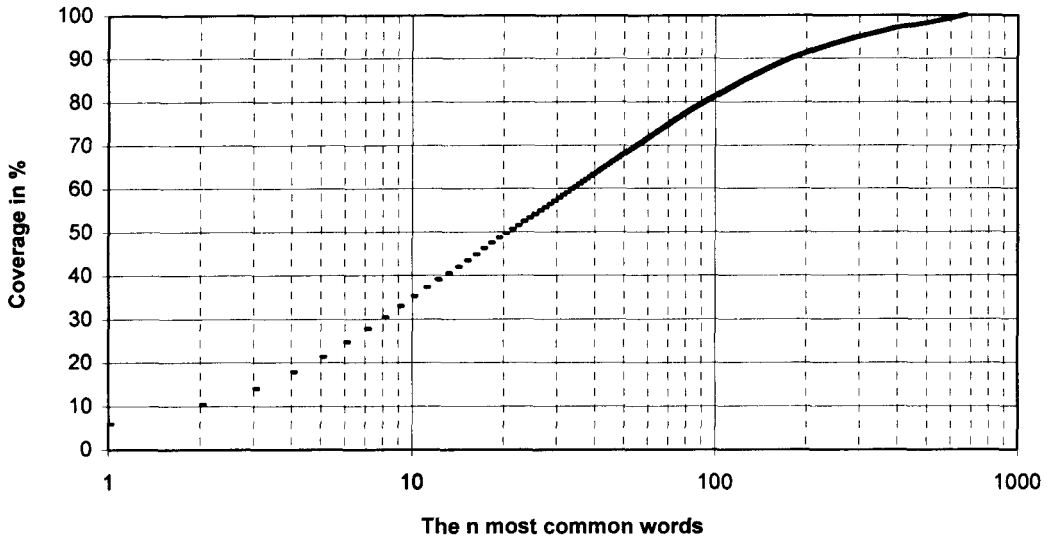

Fig. 4. Word coverage as a function of frequency rank, i.e., the words are ordered by frequency rank along the abscissa.

Labelling and phonetic transcriptions

Automatic alignment followed by manual corrections

The recorded utterances are labelled on the word, phoneme, and phone levels with links between the levels. In this way, it is easy to extract the phonetic realisation of the words, as well as the word affiliation of individual phones or phonemes. We use an automatic labelling and alignment procedure, described by Blomberg & Carlson (1993). In that system, a lexicon and a set of rules designed for text-to-speech applications (Carlson et al., 1982) are used for the generation of a base form phoneme

transcription of an utterance. Optional word pronunciations are added and optional phonological rules are applied. The rules have proven to be especially important at word boundaries. The estimated phonetic transcription of a particular utterance is obtained as a result of the alignment procedure.

In an attempt to perform pitch synchronous labelling, a post-processing procedure adjusts all label positions to an appropriate zero crossing in the speech signal.

The output of the automatic alignment procedure is manually corrected using an interactive program for label editing, which features audio playback together with speech wave and spectrogram displays.

Phoneme inventory

The phoneme alphabet is essentially identical to that used in the KTH text-to-speech system. The correspondence with the IPA symbols is displayed as part of Table 5. Some extensions, not shown in the table, have been added to the text-to-speech inventory in order to account for extralinguistic and non-speech sounds. Another difference is that plosives have been split into an occlusive and a release segment, e.g., *Kk* denotes an occlusion followed by a release. The labels used for the release segments are the lower case corresponding characters to the occlusive segment. Vowel insertions are labelled by *v*, e.g., *H'AMvN* indicates a vowel insertion. Compound morph boundaries are marked at the phoneme level by the '#' sign. Word accents are positioned in front of the vowel in the transcription. Word accent I is marked by a ' before the stressed vowel, e.g., *Dd'E:Tt*. Word accent II is marked by " before the primary stressed vowel. Compounded words also have secondary stress marked by ` before the vowel, e.g., *SJ"Ä3R#G`Å:2DE0N*. Accent II words that are not compounded have no secondary stress, e.g., *SKk"ULE0* (Bruce & Granström, 1993).

Word pronunciations

Certain words have been pronounced in many different ways: Varying tonal accents, consonant deletions, relaxed pronunciation forms, word boundary coarticulations, etc., give rise to a large possible number of unique transcriptions of a word. Also non-canonical pronunciation has sometimes been used, especially for names. A common word in this application, *'skärgården' (the archipelago)*, has been transcribed in 25 different ways. Function words are often quite reduced. Vowels, consonants, and occasionally even syllables, can be deleted in these words. Table 4 shows examples of different observed pronunciation forms of two words. The initial and final phonemes of the words are often modified due to coarticulation with the previous and the following words.

Table 4. Observed pronunciation of two words in the corpus. The word 'skärgården' (the archipelago) has been pronounced as a compound or as a single-morph word. These are distinguished by the stress marks and the existence of a morph delimiter '#' in the compound realisation form.

Word	freq.	Word	freq.
skärgården	*69*	*skulle*	*33*
SJ'Ä3RGÅ:2N	15	SKk"ULE0	26
SJ'Ä3RGgÅ:2N	6	SKk"U	3
SJ'Ä3GgÅ:2N	5	SKk"UL	2
SJ"Ä3R#Gg`Å:2DdE0N	5	SKkLE0	1
SJ'Ä3RGgÅ:2DdE0N	4	SKk"UE0	1
SJ'Ä3RGgÅ:2DE0N	4		
SJ"Ä3R#Gg`Å:2DE0N	4		
SJ'Ä3GgÅ:2DE02N	3		
2S'Ä3RGgÅ:2N	3		
SJ'Ä3RGgÅ:N	2		
SJ"Ä3R#Gg`Å:2DdE02N	2		
SJ"Ä3R#G`Å:2DE02N	2		
SJ"Ä3#Gg`Å:2DdE0N	2		
SJ'Ä3RÅ:NG	1		
SJ'Ä3RGÅ:N	1		
SJ'Ä3RGÅ:2DdE02N	1		
SJ'Ä3RGÅ:2DE0N	1		
SJ'Ä3RGÅ:2DE02N	1		
SJ'Ä3GgÅ:2DdE0Nv	1		
SJ'Ä3GgÅ:2DE0N	1		
SJ'Ä3GÅ:2DE0N	1		
SJ"Ä3R#Gg`Å:RDdE0N	1		
SJ"Ä3R#G`Å:2DE0N	1		
SJ"Ä3#Gg`Å:2DdE0Nv	1		
2S"Ä3R#G`Å:2DdE0N	1		

Phoneme frequencies

The word frequency statistics bias the corresponding phoneme occurrence frequencies somewhat. Table 5 shows a phoneme frequency list of the WAXHOLM corpus, divided into the dialogue and the reference sentence parts. The list is based on the dialogue material. The corresponding rank position in studies by Hedelin et al. (1988) on newspaper text and by Fant (1967) on Swedish telephone conversation, is also displayed. Some phonemes, e.g., [l], [ɔ:], [k] and [ɪ], are more common in the WAXHOLM dialogue corpus than in the other two studies. This is judged to be caused by relatively higher occurrence frequencies of some application words. Figure 5 shows a histogram of the phoneme frequencies in the dialogue part.

Table 5. A frequency sorted list of the phonemes in the corpus, compared with studies by Hedelin et al. (1988), and Fant (1967).

Rank	TTS symbols	IPA symbols	Observations in dialogue part	Observations in reference text	Rank in Hedelin (1988)	Rank in Fant (1967)
1	A	a	3203	1064	2	3
2	N	n	2648	944	3	1
3	R	r	2489	1231	1	2
4	T	t	2376	867	4	5
5	L	l	2267	939	7	12
6	Å:	o:	2099	203	22	15
7	K	k	2050	528	9	16
8	I	ɪ	1707	532	10	17
9	S	s	1602	935	5	9
10	V	v	1453	402	14	19
11	Å	ɔ	1347	592	12	10
12	M	m	1300	463	11	14
13	A:	ɑ	1255	400	15	6
14	E	ɛ	1141	646	17	8
15	D	d	1078	785	8	4
16	G	g	970	196	20	24
17	F	f	946	337	16	20
18	J	j	889	337	25	7
19	H	h	775	175	24	18
20	E0	ə	663	718	6	-
21	B	b	623	67	23	27
22	E:	e:	504	202	13	13
23	P	p	492	404	21	25
24	I:	i:	373	532	19	23
25	Ä3	æ:	321	333	-	-
26	Ö:	ø:	285	134	29	26
27	Ä	ɛ	260	132	18	11
28	O	ʊ	247	132	31	30
29	U:	ʉ:	232	197	32	22
30	U	ɵ	220	67	28	28
31	SJ	ɧ	175	224	35	31
32	2N	ɳ	168	264	36	37
33	2D	ɖ	162	133	39	38
34	Ö	ø	141	134	33	33
35	NG	ŋ	135	129	27	29
36	Y	ʏ	128	1	38	34
37	O:	u:	105	199	30	32
38	Ö3	œ:	104	198	-	-
39	2T	ʈ	100	130	37	35
40	2S	ʂ	77	110	34	-
41	Ä4	æ	71	201	-	-
42	Y:	y:	58	269	40	39
43	2L	ɭ	57	0	42	40
44	Ä:	ɛ:	53	136	26	21
45	TJ	ç	44	133	41	36
46	Ö4	œ	26	68	-	-

The reference sentences are, as previously mentioned, designed to contain uncommon phonemes. The use of these sentences during training of a phoneme library for recognition will therefore raise the coverage of low-frequency phonemes. One remaining problem is the very limited phonemic context of these phonemes due to the low number of different sentences. This will change the acoustic characteristics from neutral positions towards positions given by the surrounding phonemes. Further, the fact that these phonemes pertain to read rather than spontaneous speech has implications for their spectral properties.

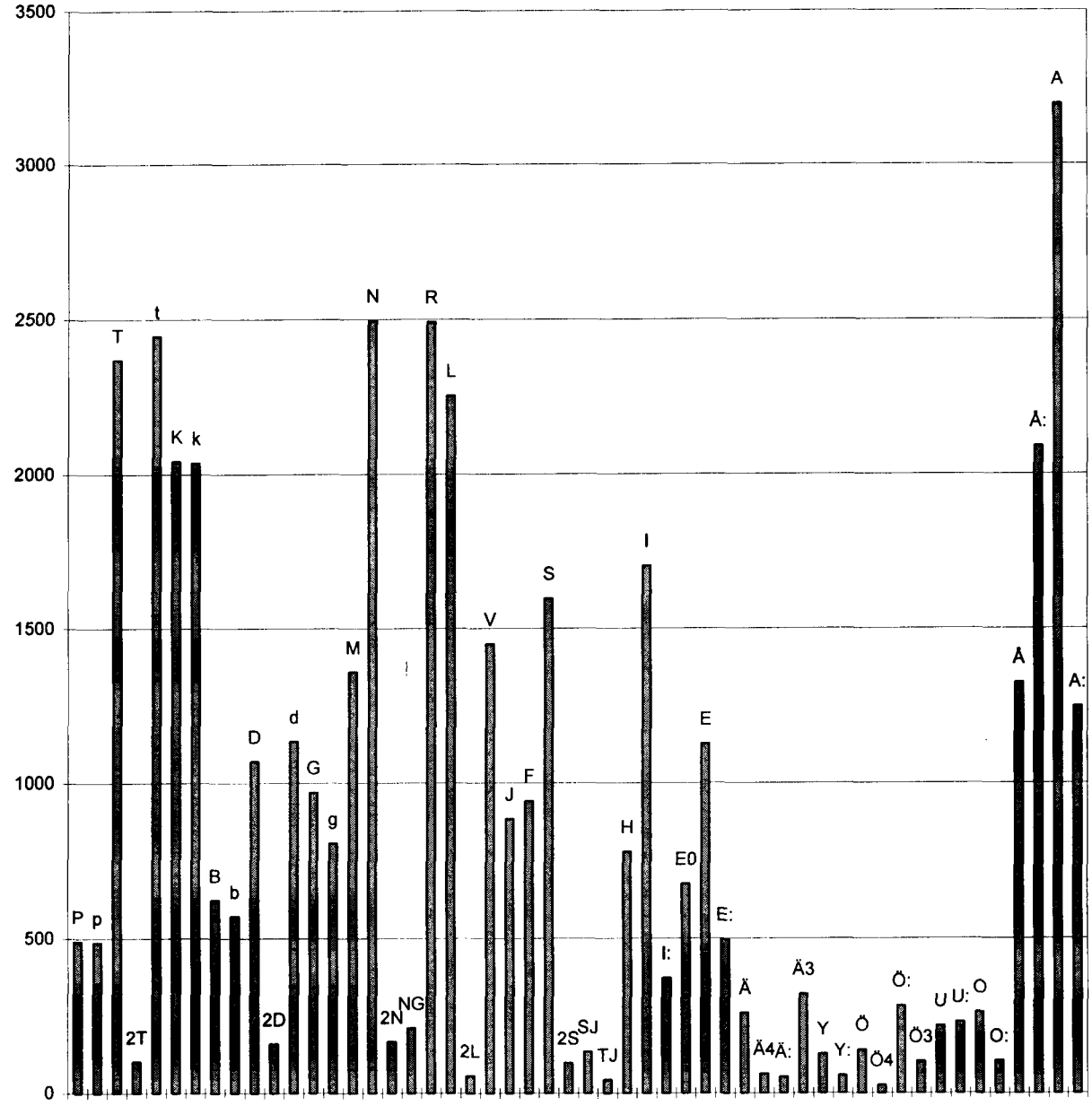

Fig. 5. Phoneme frequencies in the dialogue part of the WAXHOLM corpus. For explanation of phonemes, see Table 5.

Extralinguistic sounds

Extralinguistic sounds are entered manually during the post-processing of the data. The extralinguistic categories that are considered are interrupted words, inhalations, exhalations, clicks, laughter, lip smacks, hesitations, and hawkings.

Inhalations, which often occur in combination with smacks, are the most common extralinguistic events. Figure 6 shows the number of extralinguistic sounds, category by category. All but a few inhalations are utterance initial. There are also aspirated non-speech segments, labelled as parts of pauses, which are generally less prominent than the initial inhalations. Exhalations occur in utterance final positions and, to some extent, in mid-utterance positions. Most utterances end with a relaxation gesture, that is, with centralised and often aspirated segments. When the final aspiration is strong, it is labelled as an exhalation.

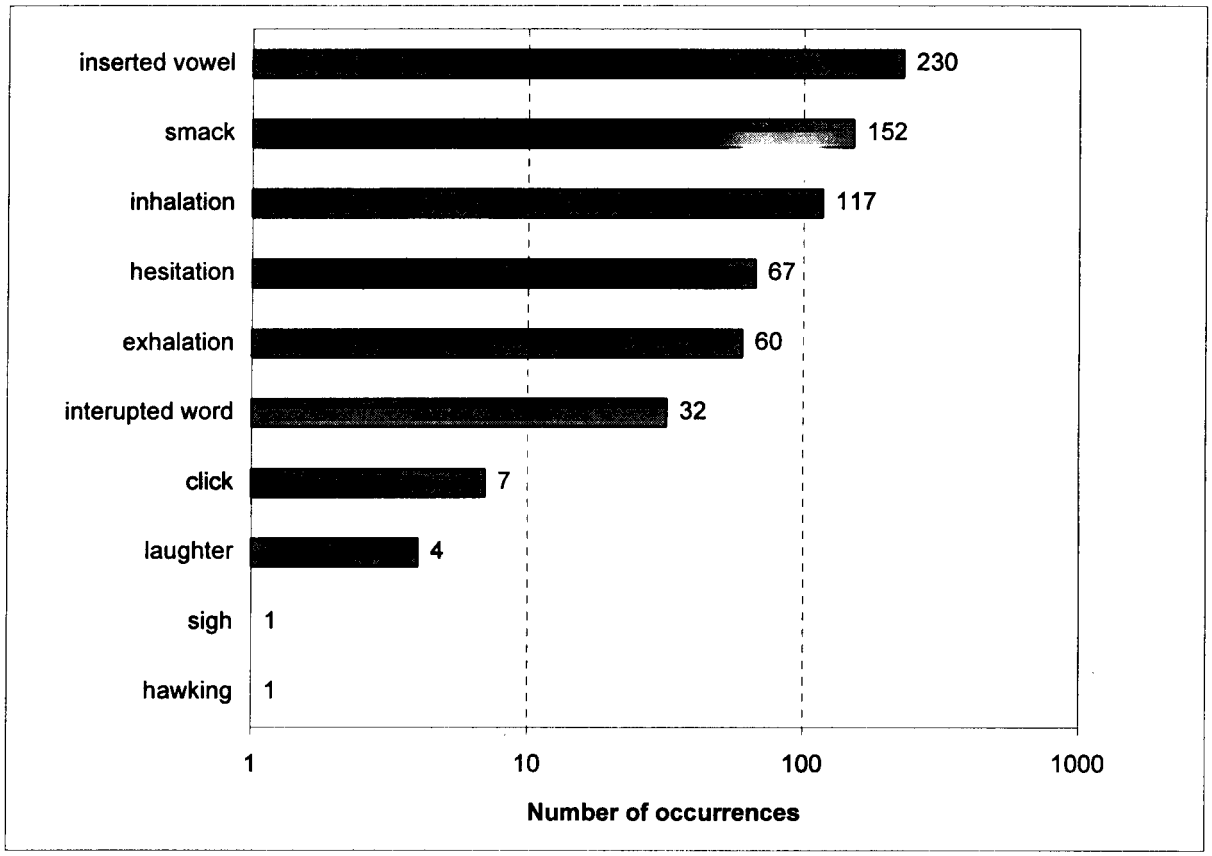

Fig. 6. The number of extralinguistic sounds labelled in the speech material.

Inserted vowel sounds are also labelled. This kind of sound occurs when a consonant constriction is released. More than 80% of the inserted vowel segments occur in word-final position, about 10% can be found in compounds at the intra-word boundary and quite often they are found in utterance final position. The phonetic context of an inserted vowel segment is often /r, l, m, n/. The inserted vowel segments are unevenly distributed over the subjects, as few as 10 speakers making up for more than half of all vowel insertions: 138 insertions out of 238. About one fourth of the

speakers have no inserted vowels at all. Thus, vowel insertion is a speaker-specific feature occurring in clear speech.

Hesitations are commonly found in utterance- or sentence-initial positions. The major part of the remaining cases are found in conjunction with place names as, for example, in the utterance "I would like to go to *uhm* Vaxholm." This could either be a common dialogue phenomenon or an artefact due to the fact that the dialogue is not spontaneous.

Dialogue analysis

The database was collected using preliminary versions of each module in the WAXHOLM system. This procedure has advantages and disadvantages for the contents of the database. System limitations will already from the beginning put constraints on the dialogue, making it representative for a human-machine interaction. However, since the system was under development during the data collection, it was influenced by the system status at each recording time. After about half of the recording sessions, the system was reasonably stable, and the number of system "misunderstandings" had been reduced. In this section, we will discuss subject performance and system performance during the collection of the WAXHOLM database. As research on dialogue systems develops, it becomes more important to develop new methods to evaluate human-machine interaction (Hirschman & Pao, 1993).

Subject performance

The subjects were initially recruited from the staff of the department. After the first stage, students and friends were asked to act as subjects. A total of 66 subjects participated in the experiment. Each subject was presented with 3 scenarios. A total of 198 scenarios were recorded and analysed. Each scenario required that the user solve one to four subtasks. A subtask could be that the subject had to request a timetable, a map or a list of facilities. Each subtask, in turn, required specification of several distinct constraints, such as departure port, destination port and departure day, before the subtask could be solved. The subjects had to provide the system with up to ten such constraints, with a mean of 4.3, in order to solve a complete scenario.

The database contains 265 subtasks, about 84% of which were solved by the subjects. Figure 7 displays the number of utterances needed to solve one subtask. It can be seen that in 75 percent of the cases, 199 out of 265, the subjects had completed a subtask after one to five utterances.

The subjects needed about 7 utterances to solve one scenario. After the task was completed, several subjects continued to ask questions in order to test the system. About 3 additional utterances per scenario were collected this way. This means that we have about ten utterances collected for each scenario.

In 42 cases, a scenario could not be completely solved by a subject, corresponding to a failure rate of 21%. In half of these, 21 scenarios, some of the subtasks were solved by the subjects. Eight times the scenario could not be completed due to

technical problems. In our analysis, we have divided the scenarios into two groups. The first group, consisting of ten scenarios, had a task completion of 75-90%. In this group, there were 173 included. The second group, four scenarios, turned out to be more difficult. The 25 sessions in this group had a task completion of less than 50%. There are several reasons for the low completion score for this last group:

- a difficult scenario, with many subtasks (four recording sessions)
- technical problems (one recording session)
- out of vocabulary words (eight recording sessions)
- grammatical structures not covered by the grammar (two recording sessions)

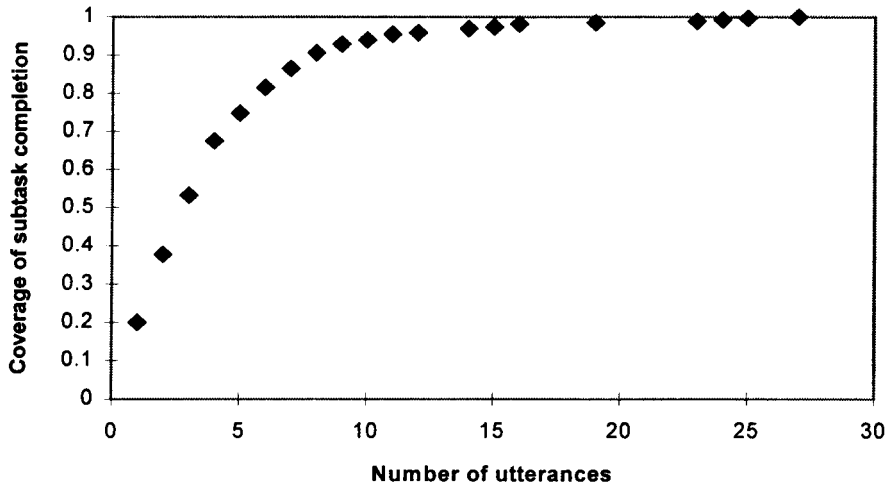

Fig. 7. Cumulative distribution of the number utterances needed by the subjects to complete one subtask.

The average utterance length was 5.6 words. The average length of the first sentence in each scenario was 8.8 words. These numbers are slightly lower than those reported in the SUNDIAL project, 6.2 and 9.6, respectively (Giachin & McGlashan, 1994). This can partly be explained by language-specific differences, e.g., Swedish tends to group words into compounds (Carlson et al., 1986). The utterance length distribution shows a weak maximum at two words and one more pronounced at five words (Fig. 8.) One reason for this distribution is that many utterances were answers to system questions. As an example, one type of system question was: "Which port would you like to go to/from?" A typical answer to this question was: "To/From Stockholm" or "I want to go to/from Stockholm." (The infinitive mark is left out in Swedish.) The question "Which day do you want to travel?" had a similar length distribution with maxima at one and five words. Most subjects answered "Friday" or another weekday. The responses to the question: "What time do you want to go?" had peaks at an utterance length of two and six words.

Even though the system was changed and improved, there is no clear tendency for shorter or longer dialogues in the beginning and the end of the data collection. In our

sample, age or profession did not have any influence on utterance length or number of utterances for each scenario.

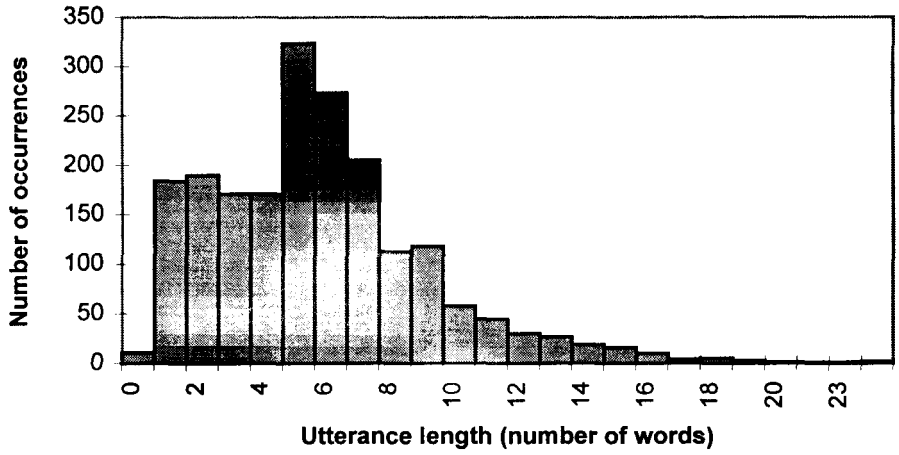

Fig. 8. Utterance length distribution.

We can find a few examples of restarts in the database due to hesitations or mistakes on the semantic, grammatical, or phonetic level. However, less than 3% of the utterances contain such disfluencies, which should be compared to 5% reported in the SUNDIAL project, (Giachin & McGlashan, 1994). Some of the restarts are exact repetitions of a word or a phrase. In some cases a preposition, a question word, or a content word, is changed. We also find repetitions of incorrectly pronounced words. About one fourth of the restarts occur in interrupted words, that is, in words that are not phonetically completed.

System response analysis

The WAXHOLM database contains approximately 1900 dialogue turns. After the first 37 sessions (35 subjects) the system went through a major revision. The first phase included approximately 1000 subject utterances. The system responses "I do not understand" and "You have to reformulate" occurred in 35.8 % of the system responses. In the second phase, the dialogue manager was updated as well as the scenarios. In this phase, 31 subjects produced 900 utterances. The improved system failed to understand 20.9% of the time, an improvement of 15%. It should be noted that this system response in some cases also is the correct one. The responses from the system displayed in Table 6 are divided into six groups (I-VI). In the following, we will discuss each type separately. All the data in the table are pooled over the whole collection period.

Table 6. Distribution of system responses.

Type of system response	Number of system responses	Percent of system responses
I Questions from the system	548	29
II Information from system	639	34
III 'I do not understand'	348	18
IV 'You have to reformulate'	221	12
V 'Please repeat'	76	4
VI 'Thank you'	74	4
Total	1906	100

I. Questions from the system

Most of the questions from the system occurred when the system predicted that the subject wanted a timetable displayed. In these cases, the distinct constraints were evaluated, and if some information was missing, the system took the initiative to ask for this information. System questions and subjects' responses for the timetable topic are shown in Table 7. The subjects answered the system questions in 95.4% of the cases. Thus, the subjects were quite co-operative and rarely, one percent, used the possibility to change the topic during the system-controlled dialogue part. In a more realistic environment, using speech recognition as input, the system might mis-understand the user's goal, and topic changes by the subject will become more frequent.

A special study of the 175 subject responses to the system question "What day would you like to travel?" shows that 57 percent are covered by the four most common syntactic structures ("X-day", "on X-day", "I want to go on X-day" and "I want to go today"). However, we also have an example of a subject asking for hotels.

Table 7. Timetable topic system questions and subject answers.

Question from system	Subject responses	Subject responded to the question	Subject changed the topic	Subject asked a new question	Subject didn't know	Subject ended the scen.
Where do you want to depart from?	207	194	2	10	1	
What day do you want to travel?	175	169	1	2	2	1
What time do you want to go?	125	121	2	1		1
Where do you want to go?	41	39	1	1		

II Presentation by the system of a result

An oral feedback was given by the system with the help of speech synthesis when information requested by the subject was displayed. As an example, the system informed the user "This is a list of hotels in Vaxholm." This presentation was mostly used to carry the dialogue forward, but gave the subject helpful information about the constraints used in the database search.

III System response: 'I do not understand' and IV System response: 'You have to reformulate'

The most serious problems occurred when the system failed to 'understand' an utterance. The first system response was a simple "I do not understand" utterance. If the failure to understand occurred once more, the system elaborated more on the problem. First, the subject was informed where the system failed to understand, if it was a linguistic problem. Second, the system asked the user to use a complete sentence next time. The following utterance from the subject was used to evaluate whether the system-predicted topic actually agreed with this new utterance or whether the topic should be changed.

The system responded 'I don't understand' 575 times corresponding to 268 occasions if consecutive repetitions are counted as one occasion. The number of utterances from the subjects needed for the system to understand is shown in Figure 9. In 50% of the cases the system recovered after one additional utterance.

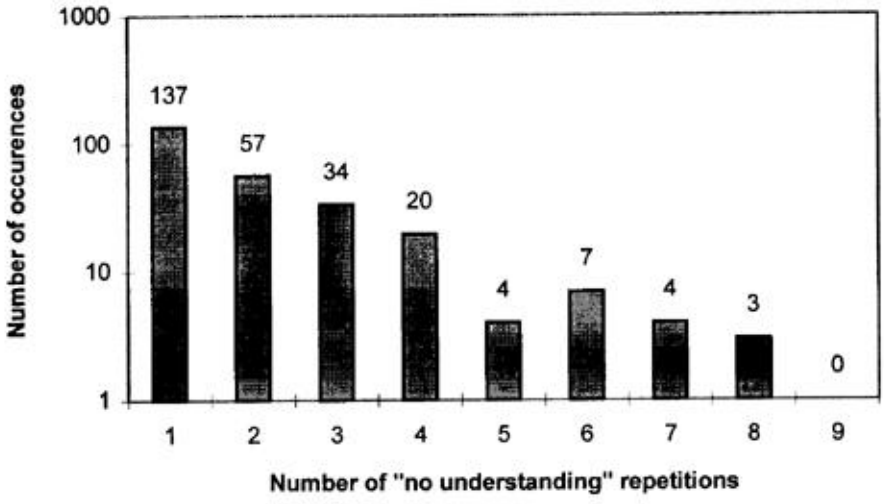

Fig. 9. Distribution of how many utterances the system needed to recover from a "no understanding" state.

V System response: 'Please repeat'

A few of the system responses occurred when the recording level was too high or low. At this point, the system simply asked the user to repeat the last utterance together with a request to the subject to make an appropriate level adjustment.

VI System response: 'Thank you'

One of the topics in the dialogue is the closing topic in which the system thanked the user and ended the scenario.

Final remarks

We are currently updating the system to prepare for a new data collection phase. The users will encounter a more realistic system setup featuring a graphical interface visualising the domain of the application and the system capabilities. The analysis of the dialogue material will be valuable in the development of intelligent system response generation. The spontaneous speech data collected provide good training material for the speech recognition module, which will be tested and evaluated within the framework of the WAXHOLM application.

The intended use of the test sentences is to perform experiments with speaker characterisation and fast speaker adaptation, in which phonetically rich speech data is desired to obtain high performance. The test sentences have been used in a study of human speaker recognition (Carlson & Granström, 1994). Moreover, the speech data is used for acoustic-phonetic studies of context and position-dependent spectral phoneme variations.

Acknowledgement

This work has been supported by grants from The Swedish National Language Technology Program.

References

Aust H, Oerder M, Seide F & Steinbiss V (1994). Experience with the Philips Automatic Train Timetable Information System. In: *Proc IEEE workshop on Interactive Voice Technology for Telecommunications Applications* (IVTTA94), 67-72.

Beskow J. Rule-based visual speech synthesis. (Accepted for *Eurospeech '95,* Madrid.)

Blomberg M (1991). Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references. In: *Speech Communication* 10: 453-462.

Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Lindell R & Neovius L (1993). An experimental dialogue system: WAXHOLM. In: *Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology*, Berlin, 1867-1870.

Blomberg M & Carlson R (1993). Labelling of speech given its text representation. In: *Proc Eurospeech'93, 3rd European Conference on Speech Communication and Technology*, Berlin, 1993, 1775-1778.

Bruce G & Granström B (1993). Prosodic modelling in Swedish speech synthesis. In: *Speech Communication*, 13: 63-76.

Bruce G, Granström B, Gustafson K, House D & Toati, P (1994). Modelling Swedish prosody in a dialogue framework. In: *International Conference on Spoken Language Processing*, Japan, 1099-1102.

Carlson R, Granström B & Hunnicut S (1982). A multilingual text-to-speech module, In: *Proc. ICASSP '82,* Paris, 1604-1607.

Carlson R, Elenius K, Granström B & Hunnicutt S (1986). Phonetic properties of the basic vocabulary of five European languages: Implications for speech recognition. In: *Proc ICASSP 86,* Tokyo, 4: 2763-2766.

Carlson R, Granström B & Hunnicutt S (1991). Multilingual text-to-speech development and applications. In: Ainsworth AW, ed, *Advances in speech, hearing and language processing*, London: JAI Press, UK.

Carlson R & Granström B (1994). An interactive technique for matching speaker identity. In: *Papers from the Eighth Swedish Phonetics Conference, Working Papers 43*, Dept. of Linguistics, University of Lund, 41-45.

Carlson R (1994). Recent developments in the experimental "WAXHOLM" dialog system. In: *ARPA Human Language Technology Workshop*, Princetown, New Jersey, 207-212.

Carlson R & Hunnicutt S (1995). The natural language component - STINA. In: *STL-QPSR, KTH,* 1: 29-48.

Dalsgaard P & Baekgaard A (1994). Spoken language dialogue systems. In: *Proc in Artificial Intelligence, Infix.* Presented at the CRIM/FORWISS workshop on 'Progress and Prospects of Speech Research and Technology, Munich.

Elenius K & Takács G (1990). Acoustic-phonetic recognition of continuous speech by artificial neural networks. In: *STL-QPSR, KTH,* 2-3: 1-44.

Elenius K & Tråvén H (1993). Multi-layer perceptrons and probabilistic neural networks for phoneme recognition. In: *Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology*, Berlin, 1237-1240.

Elenius K & Blomberg M (1992). Experiments with artificial neural networks for phoneme and word recognition. In: *Proc of ICSLP 92,* Banff, 2: 1279-1282.

Fant G (1967). *Kompendium i talöverföring.* Dept. of Speech Communication, KTH. (In Swedish.)

Giachin, E & McGlashan S (1994). Spoken language dialogue systems. In: *Course Notes Volume 2, European Summer School on Language and Speech Communication*, Utrecht University, Holland.

Glass J, Flammia G, Goodine D, Phillips M, Polifroni J, Sakai S, Seneff S & Zue V. Multilingual spoken-language understanding in the MIT Voyager System. In: *Speech Communication.* (To be published.)

Gustafson J (1992). Databashantering som del av ett talförståelsesystem. M Sc thesis, Dept of Speech Communication and Music Acoustics, KTH. (In Swedish)

Gustafson J (1994). ONOMASTICA - Creating a multi-lingual dictionary of European names. In: *Papers from the 8th Swedish PhoneticsConference*, Lund, Sweden, 66-69.

Hedelin P, Huber D, Leijon A (1988). Probability distributions of allophones, diphones and triphones in phonetic transcription of Swedish newspaper text. In: *Technical report no 8,* Dept. of Information Theory, School of Electrical and Computer Engineering, Chalmers University of Technology, Göteborg, Sweden.

Hirschman L & Pao C (1993). The cost of errors in a spoken language system. In: *Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology,* Berlin, 1419-1422.

Oerder M & Aust H (1994). A realtime prototype of an automatic inquiry system. In: *Proc ICSLP International Conference on Spoken Language Processing,* Yokohama, 703-706.

Peckham J (1993). A new generation of spoken dialog systems: results and lessons from the SUNDIAL project. In: *Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology,* Berlin, 33-40.

Ström N (1994). Optimising the lexical representation to improve A* lexical search. In: *STL-QPSR, KTH,* 2-3: 113-124.

Appendix A: Reference sentences

1. Vaxholm ligger i Stockholms skärgård.
   (Vaxholm is located in the Stockholm archipelago.)

2. Lila stolar bärs in i salen.
   (Purple chairs are being carried into the hall.)

3. Fortkörning är värre än mord, sa konstapel Thörnhjort.
   (Speeding is worse than murder, said Constable Thörnhjort.)

4. Det var kyligt i luften och stjärnorna skimrade.
   (The air was cool and the stars shimmered.)

5. Lediga och utvilade tittade dom på föreställningen i en timme.
   (Off work and rested, they looked at the performance for an hour.)

6. Sprakande fyrverkeripjäser exploderade över oss
   (Sparkling fireworks exploded above us.)

7. Där kommer nya röda hus att skjuta i höjden.
   (New red buildings will shoot up into the air over there.)

8. Öppna dörren innan jag fryser ihjäl.
   (Open the door before I freeze to death.)

Appendix B: Scenarios

Scenario 1
It's a beautiful summer day and you are in Stockholm. You decide that you'd like to go to Vaxholm.

Your task is to find out when boats leave for Vaxholm this evening.

Scenario 2
You would like to take a trip out into the archipelago during the weekend. Because you'd like to live comfortably, you want to stay at a hotel. You quit work at 3:00 p.m. on Friday and start work again at 10:00 on Monday.

Find out where you can stay and when you can travel.

Scenario 3
You've found out that Gällnö is a very beautiful island, so you've decided to set up your tent there, if possible.

You should therefore find out where Gällnö is located and if it is possible to camp there.

Scenario 4
Your wife's (husband's) birthday is tomorrow, Saturday, and you've considered surprising her (him) with a romantic dinner at a restaurant out in the archipelago. You can't afford a hotel, so you have to stay at a hostel.

Your task, then, is to find a port where both a restaurant and a hostel can be found, and then to find out which boats you can take in order to arrive there on Saturday and go back on Sunday. You'll be leaving from Strömkajen.

Scenario 5
You are now in Vaxholm and want to travel to Vegabryggan. You want to go after lunch today.

Your task is to find out which boats you can take.

Scenario 6
The group that you work with is planning to have a conference out in the archipelago and you have been assigned the task of booking hotel rooms.

You should, then, find out which ports have hotels.

Scenario 7
You want to stay at a hostel out in the archipelago over the weekend.

Find out where hostels are located and how you can get there.

Scenario 8
You are now in Vaxholm and have just missed the last boat home, so you have to stay overnight. You're also starting to be rather hungry.

Find out if you can eat somewhere in Vaxholm, and if there is a place to stay overnight there.

Scenario 9
You're planning to camp out in the archipelago over the weekend.

Your task is, then, to find out where you can camp and which boats go there.

Scenario 10
You're planning to travel to Kymmendö.

Find out if it is possible to stay overnight there or at some nearby port.

Scenario 11
Think of a possible destination and find out if it is possible to get there by boat.

Also, find out where it is located.

Scenario 12
You're planning to visit a friend at Ramsöberg, and want to stay as late as possible.

Find out when the boat leaves this evening.

Scenario 13
You're planning to visit your friends who have a summer cottage on Lådna.

Find out where Lådna is located and how you can get there.

Scenario 14
You're planning to travel to Staveström today, and then continue to Finnhamn.

Find out how you can do this.