

DELIVERABLE IDENTIFICATION

Identification number	LE2-4001—design.doc
Type	Technical Report
Title	Language Specific Document
Status	Draft
Deliverable	
Work Package	WP 1
Task	Task 1.1
Period covered	T0 - T6
Date	31 May, 1996
Version	1.0
Number of pages	17
Author(s)	Jari Yli-Hietanen, DMI
Work package (WP) / Task (T) responsible	WP1 - Richard Winsky Task 1.1 - Franco Senia
Project contact point	Harald Höge, Siemens AG, ZFE T SN 5, D-81730 München Phone: +49 89 636 53374, Fax: +49 89 636 49802 E-mail: Harald.Hoegel@zfe.siemens.de
CEC project officer	José Soler
Status	Public
Actual distribution	Consortium and CEC
Supplementary notes	
Key words	
Abstract	
Status of abstract	

Received on	
Recipients catalogue number	

DOCUMENT EVOLUTION

Version	Date	Status	Notes
1.0	96-05-31	Draft	

--	--	--	--

1. INTRODUCTORY REMARKS	1
2. DATABASE CONTENTS	1
2.1 <i>Isolated digits</i>	2
2.1.1 Isolated digit	2
2.1.2 Isolated digit sequence	2
2.2 <i>Digit/number strings</i>	2
2.2.1 Prompt sheet number	2
2.2.2 Telephone number	3
2.2.3 Credit card number	3
2.2.4 Pin code	3
2.3 <i>Natural number</i>	3
2.4 <i>Money amounts</i>	3
2.5 <i>Times</i>	4
2.6 <i>Dates</i>	4
2.7 <i>Spelled words</i>	5
2.8 <i>Application keywords / keyphrases</i>	5
2.9 <i>Yes/no and spontaneous questions</i>	5
2.10 <i>Phonetically rich sentences</i>	6
2.11 <i>Phonetically rich words</i>	6
2.12 <i>Directory assistance names</i>	6
3. LINKS TO SDB	6
4. PROMPT SHEETS / INSTRUCTION SHEETS	6
4.1 <i>Creation</i>	6
APPENDIX	7

FINNISH-SWEDISH LANGUAGE-SPECIFIC DETAILS OF THE SPEECHDATFDB DATABASE

1. INTRODUCTORY REMARKS

The Fixed Network database of the Finnish-Swedish language will comprise a total of 1000 speakers. This is a new database and none of the speech was collected during the SpeechDat(M) project.

2. DATABASE CONTENTS

Each call within the database will consist of a total of 43 items. All off these 43 items are mandatory within SpeechDat.

Number of items	Utterance description
2	Isolated digit items
1	- single isolated digit
1	- sequence of 10 isolated digits in one utterance
4	Digit/number strings
1	- prompt sheet number
1	- telephone number
1	- credit card number
1	- 6-digit PIN code
1	Natural number
1	Money amount
2	Yes/no questions
1	- predominantly yes
1	- predominantly no
3	Dates
1	- birthdate
1	- prompted date phrase
1	- relative and general date expression
2	Times
1	- time of day
1	- prompted time phrase
6	Application keywords
5	Directory assistance names
1	- city of birth / growing up
1	- most freq. cities
1	- most freq. companies/agencies
1	- proper name
1	- spontaneous proper name

3	Spellings
1	- real/artificial words
1	- spelling e.g. of directory assistance city name
1	- spelling of proper name
4	Phonetically rich words
9	Phonetically rich sentences
43	TOTAL utterances

In the following subsections the specific contents of each item will be commented on. These comments are only showing indicative deviations from the SD1.1.1. specifications. If no deviations exists, the corresponding paragraph has been left blank.

2.1 Isolated digits

2.1.1 Isolated digit

Finnish-Swedish digits and their variants are shown in the table below. The variants of one and nine are rather common in spoken Finnish-Swedish while the variants of the other digits are rarely used.

English	Swedish	variants
zero	noll	nolla
one	ett	en
		etta
two	två	tvåå
three	tre	trea
four	fyra	fyr
five	fem	femma
six	sex	sexa
seven	sju	sjuå
eight	åtta	
nine	nio	nie
		nia

2.1.2 Isolated digit sequence

Optional star and hash will not be used. The sequence will be different on all prompt sheets in order not to get any bias from the context the numbers are in. The different sequences will be generated in a random way. No attempts will be made to control the variants of different digits that might be uttered by the speakers (see table above).

2.2 Digit/number strings

2.2.1 Prompt sheet number

In order to be able to identify the prompt sheet used a five digit prompt sheet number will be used.

2.2.2 Telephone number

The format of the telephone number will be standard Finnish number including area code: 2-3 digits area code plus 6-8 digits phone number. GSM- and NMT-numbers will also be used: the digits 049, 040 or 050 followed by a 7 digit phone number.

2.2.3 Credit card number

The 150 credit card numbers will be taken from the list supplied by the SDB definitions.

2.2.4 Pin code

The 150 PIN code numbers will be taken from the list supplied by the SDB definitions.

2.3 Natural number

For formats of natural numbers see the tables in “2.4 Money amounts” below.

2.4 Money amounts

Each prompt sheet contains either a small or a big money amount. The small money amounts are chosen among 150 different quantities randomly distributed in the range of 1-5.000 Finnish marks. These amounts contain decimal numbers.

The large money amounts are chosen among 150 different quantities randomly distributed in the range of 100.000-10.000.000 Finnish marks. These do not contain any decimals.

Digits, numbers and amounts are formed by the words included in the table below. They are constructed in a way similar to English.

Million	thousand	hundred
miljon	tusen	hundra
miljoner		

30..90	variants	20..29	variants
tretti	trettio	tjugo	tju
fyrti	fyrtio		tjugi
fenti	fentio	tjugoett	
sexti	sextio	tjugotvå	
sjutti	sjuttio	tjugotre	
åtti	åttio	tjugofyra	
nitti	nittio	tjugofem	
		tjugosex	
		tjugosju	
		tjugoåtta	
		tjugonio	

10..19	variants	units	variants
tie	tio	noll	nolla
elva		ett	en
tolv			etta
tretton		två	tvåa
fjorton		tre	trea
femton		fyra	fyr
sexton		fem	femma
sjutton		sex	sexå
arton		sju	sjuå
nitton		åtta	

		nio	nie
			nia

2.5 Times

Swedish words used in SpeechDat for stating analogue times are:

på, morgonen, kvällen, natten, förmiddagen, eftermiddagen, kvart, över, i, midnatt, halv, timme, timmar, minuter, minut, klockan, exakt, cirka, ungefär, innan, före, efter, strax.

From these words 150 different analogue time prompts are generated.

2.6 Dates

The date words used are:

januari, februari, mars, april, maj, juni, juli, augusti, september, oktober, november, december
måndag, tisdag, onsdag, torsdag, fredag, lördag, söndag

150 different dates will be generated in the period from January the first, 1920, to December 31:st, 2020

Swedish dates are accompanied by the numeral words in the table below.

Dates have the standard form of: day of the week “-en den”, day of the month, month, year.

e.g. Måndagen den tolfte augusti, 1996 = Monday, 12/08/96.

1..9	alternatives	10..19	alternatives
första		tionde	tio
andra		elfte	
tredje		tolfte	
fjärde		trettonde	
femte		fjortonde	
sjätte		femtonde	
sjunde		sextonde	
åttonde		sjuttonde	
nionde		artonde	
		nittonde	

20..31	alternatives
tjugonde	
tjugoförsta	tjuförsta
tjugoandra	tjuandra
tjugotredje	tjutredje
tjugofjärde	tjuffjärde
tjugofemte	tjufemte
tjugosjätte	tjusjätte
tjugosjunde	tjusjunde
tjugoåttonde	tjuåttonde
tjugonionde	tjunionde
trettionde	
trettioförsta	trettiförsta

Word used for relative dates are:

i morgon, igår, i övermorgon, i förrgår, nästa, innan, före, vecka, månad, dag, idag, dan före, dan efter, julafton, juldagen, trettondagen, påskdagen, långfredagen, skärtorsdagen, annandag

påsk, pingst, kristi himmelfärdsdag, nyårsafton, nyårsdagen, midsommarafton, midsommardagen, på, måndag, tisdag, onsdag, torsdag, fredag, lördag, söndag.

2.7 Spelled words

The spelling vocabulary used in Finnish-Swedish is:

A-Z, Å, Ä, Ö, stort, stora, litet, lilla.

2.8 Application keywords / keyphrases

The application words that will be recorded for Finnish-Swedish are listed below. For a description of the function behind each word see the SD1.1.1 document.

No.	Keyword	Translation
1	cancel	avbryt
2	stop	stopp
3	help	hjälp
4	repeat	repetera
5	menu/options	meny
6	dial	slå nummer
7	operator	telefonist
8	call	ring
9	next	nästa
10	language	svenska
11	delete	radera
12	terminate	avsluta
13	continue	fortsätt
14	end	slutet
15	change	ändra
16	play	spela upp
17	send	skicka
18	previous	föregående
19	directory	katalog
20	list	lista
21	add	lägg till
22	save	spara
23	record	spela in
24	redial	ring igen
25	program	programmera
26	Synonym for list	räkna upp
27	Synonym for end	sista
28	Synonym for previous	förra
29	Synonym for change	byt
30	Synonym for delete	ta bort

2.9 Yes/no and spontaneous questions

(Indicate which question(s) should elicit 'yes' responses and which 'no' responses)

Expected yes answer:

“Is Copenhagen capital of Denmark ?”

“Är Köpenhamn huvudstad i Danmark ?”

Expected no answer:

“Is Oslo capital in Finland ?”

“Är Oslo huvudstad i Finland ?”

2.10 Phonetically rich sentences

Sentences are derived from newspapers and books

2.11 Phonetically rich words

Phonetically rich words are extracted from the same material as the phonetically rich sentences.

2.12 Directory assistance names

The list of 500 Finnish most frequent cities, the list of 500 of the most frequent companies/agencies and the list of 150 Finnish proper names are used. The lists are not complete yet..

3. LINKS TO SDB

Forename and Surname will not be drawn from the SDB design. Sentences and isolated words are not correlated with the SDB design. The reason is that no SDB is planned for Finnish-Swedish. However, all other items are used in accordance with section 4.1 of SD1.1.1.

4. PROMPT SHEETS / INSTRUCTION SHEETS

4.1 Creation

The variable speech material for each speaker will be randomly selected according to the SpeechDat specifications. The prompt sheets will be generated using shell scripts in UNIX and LATEX. Every prompt sheet will have a unique identification number.

APPENDIX

Finnish area codes in the fixed network

02	02	03	04	05	06	07	08	09	013	014	015
016	017	018	019								