



UPPSALA  
UNIVERSITET

# Draft Questionnaire for the Swedish BLARK

GSLT retreat workshop  
Gullmarsstrand 2007-01-28

Eva Forsbom  
Beáta Megyesi



# Outline

1. Goals for the Swedish BLARK
2. Definition and previous BLARKs
3. Method and plans for the Swedish BLARK
4. BLARK questionnaires



# Goals for the Swedish BLARK

- Define a minimal set of language resources (LRs) to be made available for Swedish
- Skip minority languages (for now): Finnish, Meänkieli, Sami, Romani chib, Jiddisch
- Gather information from universities, organizations and industry
- Base on previous BLARK initiatives: Dutch/Flemish and Arabic.



# Basic Language Resource Kit

- Basic language resources:
  - written/spoken mono-/multilingual corpora
  - mono- and multilingual dictionaries
  - terminology collections
  - grammars
- Benchmarks for evaluation
- Basic tools:
  - modules (e.g. taggers, parsers, grapheme-to-phoneme converters)
  - annotation standards and tools
  - corpus exploration and exploitation tools



# Previous experiments and guidelines

- ELRA/ELDA provides four matrices for LRs online where potential BLARK definers and instantiators can
  - grade resources for various applications or modules to get an overview of what LRs are needed, and
  - set up a priority list for developing the LRs that are missing (with the Arabic experience as example).

<http://www.elda.org/blark/>

	: Important
	: Very important
	: Essential
	: Not applicable
	: For all languages
	: Language specific



# Written: application vs. module

close window	ASR/Dictation	Classification	Dialog Systems	Document Production	IE	Indexing	IR/Filtering	MAT	MT	Summarization	TTS
Alignment								++			
Diacritizer											+++
Grapheme Recognition (for Handwritten Ocr)											
Grapheme Recognition (for Typewritten Ocr)											
Morphological Comp.	++	+++	+++	+++	+++	+++	++	+++	+++	+++	+++
Named Entity Recognition		+++	+	+++	+++	+++	+++	+++	+++	+++	+
Pos Disambiguator/tagger	++	++	+++	+++	+++	+++	++	+++	+++	+++	+++
Semantic Analysis		++	+++		+++	+	++	++	++	++	+
Sentence Boundary Detection			++	+++	+++		++	+++	+++	+++	++
Sentence Synthesis And Generation			+++	++				++	+++	+++	
Shallow Parsing		+	+++	++	++		+	+++	+++	++	++
Syntactic Analysis Compounded			+++	++	++			++	+++		+
Term Extraction		+++	++	+	+++	+++	+++	+++	+++	+++	
Transfer Tool (software)									+++		
Word Sense Disambiguation		+++	+++		+++	++	++	++	+++	+++	+++



# Written: resource vs. module

<b>close window</b>	Annotated Corpora	Monolingual Lexicon	Multi/Bilingual Lexicon	Multimodal Corpora for (hand) OCR	Multimodal Corpora for (typed) OCR	Parallel Multiling Corpora	Proper Names	Thesauri, Ontologies, Wordnets	Unannotated Corpora
Alignment		+++	+++			+			
Diacritizer		+++					++	++	
Grapheme Recognition (for Handwritten Ocr)		++		+++					+++
Grapheme Recognition (for Typewritten Ocr)		++			+++				+++
Morphological Comp.	++	+++							
Named Entity Recognition	+	+++					+++		
Pos Disambiguator/tagger		+++					++		
Semantic Analysis		+++						+++	
Sentence Boundary Detection	++	+++							
Sentence Synthesis And Generation	++	+++						++	+
Shallow Parsing		+++							
Syntactic Analysis Compounded	+	+++							
Term Extraction		+++							+++
Transfer Tool (software)			+++						
Word Sense Disambiguation	++	+++							++



# Missing matrix? preprocessing vs. module

	Grapheme recognition (handwritten OCR)	Grapheme recognition (typed OCR)	Format conversion	Encoding conversion	Tokenisation	Normalisation	Sentence segmentation
Alignment							
Morphological comp.							
Named Entity recognition							
PoS disambiguator /tagger							
Semantic analysis							



# BLARK inventory for Swedish

Goal:

to make an inventory of

- available resources
- industrial needs

from

- organizations, industry
- researchers at universities

in

- Sweden
- Nordic countries
- world-wide



# Method

Necessary components to gather information about SWEBLARK:

- drafting the questionnaires (sv, en : 1+1, 1)
- letter (motivations and instructions)
- testing the questionnaires
- revising the questionnaires
- distribution of questionnaires (e-mail, web)
- collection of answers
- defining the matrices
- compilation and publication of results



# 1. Exploratory questionnaires

- Purpose:
  - To get an overview & collect contacts
  - To define the BLARK matrices
- How: Extract information about
  - actors
  - LRs (needs, types, use, tools, validation, distribution, participation)
  - the market
- 2 versions:
  - For organizations, universities & experts
  - For industry (needs)
  - Based on the NEMLAR project surveys



## 2. Follow-up questionnaire

- Purpose: To complete information on LR quantity/quality (metadata):
  - availability (public domain, freeware, shareware, legal aspects, freedom to manipulate the source, costs)
  - programming code (language, makefile, stand-alone, or part of a larger module?)
  - platform
  - documentation
  - compatibility with standards
  - re-usability, adaptability, extendibility



UPPSALA  
UNIVERSITET

# Contact

## Sweden:

- GSLT list
- [sprakteknologi.se](http://sprakteknologi.se)
- DISC at KFI (Eva Strangert)

## Nordic:

- Nodali list
- NGSLT list

## World wide

- Corpora list

Other suggestions?



# Discussion points

- Informants (whom and how to contact)
- Number of questionnaires (2 or more)
- Division of labour speech/written resources
- Content of questionnaires



# References 1

- ELDA (Evaluations and Language resources Distribution Agency). URL <http://www.elda.org/blark/>.
- E. D'Halleweyn, J. Odijk, L. Teunissen and C. Cucchiarini. The Dutch-Flemish HLT Programme STEVIN: Essential Speech and Language Technology Resources. In Proceedings of LREC, Genoa, Italy. 2006. URL <http://taalunieversum.org/taal/technologie/stevin/>.
- S. Krauwer. ELSNET and ELRA: A common past and a common future. ELRA Newsletter, 3(2), 1998. URL <http://www.elda.org/blark/fichiers/elsnet&elra.doc>.
- S. Krauwer, B. Maegaard, K. Choukri, and L. Damsgaard Jørgensen. Report on BLARK for Arabic, 2006. URL <http://www.nemlar.org/Publications/index.htm>.
- V. Mapelli and K. Choukri. Report on a (minimal) set of LRs to be made available for as many languages as possible, and map of the actual gaps. ENABLER Deliverable 5.1, 2003. URL <http://www.elda.org/blark/fichiers/report.doc>.
- NEMLAR (Network for Euro-Mediterranean Language Resources). URL <http://www.nemlar.org/>.



# References 2

- M. Nikkhou and K. Choukri. Report on Survey on Arabic Language Resources and Tools in the Mediterranean Countries. 2005. URL <http://www.nemlar.org/Survey-questionnaires/index.htm>.
- M. Nikkhou and K. Choukri. Report on Survey on Industrial needs for Language Resources. 2004. URL <http://www.nemlar.org/Survey-questionnaires/index.htm>.
- STEVIN (Spraaak- en Taaltechnologische Essentiële Voorzieningen In het Nederlands). URL: <http://taalunieversum.org/taal/technologie/stevin/>.
- H. Strik, W. Daelemans, D. Binnenpoorte, J. Sturm, F. de Vriend, and C. Cucchiarini. Dutch HLT resources: From BLARK to priority lists. In Proceedings of ICSLP, Denver, pp. 1549 1552. 2002. URL <http://lands.let.kun.nl/literature/strik.2002.2.pdf>.



# Spoken: application vs. module

<a href="#">close window</a>	Customization to Different	Dialect/Language	Dictation	Embedded Speech	Emotion Identification	Emotion/Prosody Output	Generation Lips Movement	Lips Movement Reading	Speaker 2 Speaker Mapping	Speaker Adaptation
Acoustic Models	+++	+++	+++	+++	+++	+++	+++	+++	++	+++
Dialect/language Identification		+	+	+	+			+		+
Emotion Identification		+	+	+		++		+	++	+
Language Models		++	+++	++		++				
Lexicon Adaptation			+	+					++	
Lips Movement Reading		++						+++		
Phoneme Alignment			+	+					++	
Pronunciation Lexicon			+++	+++					++	
Prosody Prediction						+++				
Prosody Recognition		+	+	+	+++				++	+
Segmenter Speech/silence		++	++	++	++	+		+		+
Sentence Boundary Detection		+	+	+	++	++		+		+
Speaker Adaptation		+	++	++	+			+	++	+
Speaker Recognition/identification		+	+	+	+			+	++	+
Speech Units Selection						+++				
Speech/non-speech Music Detection		+	+	+	++			+		+
Word Boundary Identification		+	+	+	+	++		+		+



# Spoken: resource vs. module

<a href="#">close window</a>	Annotated Written Corpus	Audio Data with Prosodic Markers and other	BNSC	Desktop/Microphone & High Quality	Non Vowelised Corpus	Onomastica (proper names)	Phonetic Lexicon	Telephony	Unannotated Written Corpora	Visual Data (faces, lips, etc.)	Vowelised Corpus
Acoustic Models		+++	+++	+++				+++			
Dialect/language Identification		+	++	++		+	+	++			
Emotion Identification		+	+	+		+	+	+			
Language Models	++				++				+++		++
Lexicon Adaptation	+				+	+++	+++		+		++
Lips Movement Reading										+++	
Phoneme Alignment	++	++	++	++		+++	+++	++			+
Pronunciation Lexicon	+					+++	+++				++
Prosody Prediction	++	++				++	++				++
Prosody Recognition	++	+++		+		++	++	+			+
Segmenter Speech/silence		++	++	++				++			
Sentence Boundary Detection		++	++	++		+	+	++			
Speaker Adaptation		+	++	++				++			
Speaker Recognition/identification		+	+	+				+			
Speech Units Selection	++	+++		+		+	+	+			
Speech/non-speech Music Detection		++	++	+				+			
Word Boundary Identification		+	+	+		+	+	+			



# Attributes for Dutch/Flemish

- Availability
  - public domain, freeware, shareware, legal aspects
- Programming code
  - language, makefile, stand-alone or part of a larger module?
- Platform
- Documentation
- Compatibility with standards/standard packages
- Reusability/adaptability/extendibility



# Attributes for Arabic

- Availability
  - freedom to use
  - cost
  - freedom to manipulate source
- Quality
  - standard compliance
  - soundness
  - task relevance
  - environment relevance
- Quantity
- Standards