



UPPSALA
UNIVERSITET

A Swedish BLARK

GSLT retreat workshop
Gullmarsstrand 2006-01-29

Anna Sågvall Hein
Eva Forsbom



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



Basic LAnguage Resource Kit cont'd

- Define a minimal set of language resources (LRs) to be made available for as many languages as possible (Swedish in our case).
- Use for education and research, and pre-competitive industry needs.
- BLARK initiatives for at least two languages are on the way: Dutch/Flemish and Arabic.



Discussion points

- Definition
- Grading and priorities
- Inventory of LRs and tools
- Standards
- Quality assessment
- Development of missing LRs and tools
- Availability and open source aspects
- Organisation
- Funding



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		+++	+++		+++	++	++	++	+++	+++	+++



Spoken: application vs. module

close window	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		+	+	+	+	++		+		+



Written: resource vs. module

close window	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	++	+++							++



Spoken: resource vs. module

close window	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**



Discussion points

- Definition
- Grading and priorities
- Inventory of LRs and tools
- Standards
- Quality assessment
- Development of missing LRs and tools
- Availability and open source aspects
- Organisation
- Funding



References

- OLAC (Open Language Archives Community). URL <http://www.language-archives.org/>.
- ELDA (Evaluations and Language resources Distribution Agency). URL <http://www.elda.org/blark/>.
- E. Forsbom. Återanvändbarhet för språkvara, 2004. URL <http://stp.lingfil.uu.se/~evafo/gslt/java/paper.ps.gz>. Literature review for GSLT Java course.
- 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, 2004. 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.enabler-network.org/reports.htm>.
- 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, USA, pp. 1549-1552, pages 1549 1552, Denver, USA, 2002. URL <http://lands.let.kun.nl/literature/strik.2002.2.pdf>.