# Selection Criteria for Low Resource Language Programs

**Christopher Cieri[◦], Mike Maxwell[□], Stephanie Strassel[◦], Jennifer Tracey[◦]**
[◦] Linguistic Data Consortium. University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA. 19104 USA
email AT ldc.upenn.edu

[□] University of Maryland
College Park, MD 20742, USA
email AT umd.edu

{ccieri, mmaxwell, strassel, garjen}

**Abstract**

This paper documents and describes the criteria used to select languages for study within programs that include low resource languages whether given that label or another similar one. It focuses on five US common task, Human Language Technology research and development programs in which the authors have provided information or consulting related to the choice of language. The paper does not describe the actual selection process which is the responsibility of program management and highly specific to a program's individual goals and context. Instead it concentrates on the data and criteria that have been considered relevant previously with the thought that future program managers and their consultants may adapt these and apply them with different prioritization to future programs.

**Keywords:** language resources, low resource languages, common task programs

## 1. Introduction

The past 10 years have seen significant growth in work on resource-poor languages within the Human Language Technology (HLT) research community. Whether one sees this growth as the natural outcome of successful HLT development in well-resourced languages or as an opportunity to test the generality of HLT, the shift in focus is undeniable. Within the United States alone, the TIDES, REFLEX LCTL, Babel and LORELEI programs have all focused on developing language resources and technologies for low resource languages. However, differences in the terminology, available information and goals of low resource language efforts lead to variability and some obscurity in the language selection process.

Moving beyond the four US programs named above, there is an even greater range of motives for studying low resources languages. For example, while the LORELEI program, described in greater detail below, seeks technologies to facilitate situational awareness in the event of a disaster, the EU funded METANET (2010) program asserts that "*The majority of European languages are severely under-resourced*" and proposes that a "*coordinated, large-scale effort has to be made in Europe to create the missing technologies and transfer this technology to the languages faced with digital extinction*". The motivations for the proposed effort include quality of life, information access and the ability to collaborate across multilingual Europe. The US National Science Foundation's Documenting Endangered Languages program (2014) gives very different motivations: "*Most of what is known about human communication and cognition is based on less than 10 percent of the world's 7,000 languages. We must do our best to document living endangered languages and their associated cultural and scientific information before they disappear.*" Such differences in motivation clearly lead to very different languages studied and differences in the languages studied affect opportunities for collaboration across programs.

In addition to the programs' commitments of time and finances, the new language resources (LRs) they create are critical for bringing HLTs to new languages, a matter of great importance for their speakers. Given the size of the program investment and the potential to impact speakers' lives, we believe that selection criteria constitute a topic worthy of study. This paper, in an attempt to begin a dialog about how the community decides which languages to study, surveys the selection criteria used and available for use by low resource language research. The discussion herein focuses on several US programs for which the authors have provided information about the characteristics deemed relevant to the choice of languages. Importantly, our intent is not to sketch the actual decision making process which was the responsibility of program management and, we believe, highly specific to the programs' needs and contexts. Instead we will detail the kinds of information requested by program management and suggested by consultants as relevant to the decision making process expecting that future decision makers will assign different priorities to the same kinds of data.

## 2. Definitions of Low Resource Language and Related Terms

Before describing low resourced language selection criteria, it will be useful to try to define terms and

understand the relations among them. In the past decade of work on creating language resources for languages that lack them we have seen terms such as *low density*, *less commonly taught*, *under-resourced*, *less resourced* and *low resource*. Related fields speak of *critical* and *endangered* languages. Distinguishing these will help justify the specific selection criteria used.

*Endangered* refers to languages that are at risk of losing their native speakers through a combination of death and shift to other languages. The term typically fits within classifications of languages according to risk of intergenerational disruption that may distinguish safe from multiple levels of endangerment and moribundity and, rarely, revitalization (Krauss 1992). In addition to reduction in speakers, both decline in domains of use and structural changes characterize endangered languages (Dorian 1980). Notably, the absence of a writing system increases risk of language death (Fishman 1991). We will have little else to say about endangered languages in this paper if only because they have not been the principal focus of the HLT projects we surveyed. To illustrate this point, Figure 21 charts the number of languages selected by each of programs surveyed according to the languages' Extended Graded Intergenerational Disruption Scale (EGIDS) rating. EGIDS scores are a measure of endangerment ranging from 1 to 13 with higher numbers indicating greater threat (Lewis and Simons 2010). As we see, nearly all of the languages have EGIDS scores of 1 or 2, which refer, respectively, to official national and provincial languages. An EGIDS score of 3 marks a language of broader communication lacking official status while 4 and 5 indicate languages in vigorous use with standardization, literatures and, in the case of 4, the support of educational institutions. None are described as threatened (EGIDS=6b).

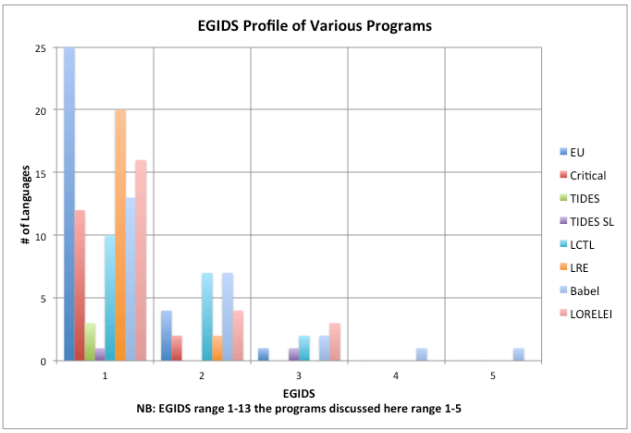


**Figure 2: EGIDS scores of languages selected by the programs surveyed**

In contrast, *critical* has typically referred to languages that suffer an undesirable ratio of supply to demand, typically of teachers and translators. In the US, one sees the term used in government programs that sponsor language and cultural immersion. It is more difficult to find explicit definitions than enumerations of the critical languages. The US State Department funded Critical Language Scholarship Program for 2015 listed: Arabic, Azerbaijani, Bangla, Chinese, Hindi, Indonesian, Japanese, Korean, Persian, Punjabi, Russian, Swahili, Turkish and Urdu. If we assume these labels refer to standard languages spoken in their homelands then none are endangered. In fact Ethnologue (Lewis, Simons, Fennig 2015) lists all as *statutory national languages* except Punjabi and Swahili which it lists as a *statutory provincial language* and *de facto national language*, respectively. However almost half have appeared in one or more of the HLT programs mentioned above. One might also note that the action of these programs and others have greatly increased the resources available for Standard Arabic and Mandarin Chinese though they remain critical languages.

Within the HLT literature, *low density* refers to languages "*for which few online resources exist*" (Megerdoomian, Parvaz 2008) or "*for which few computational data resources exist*" (Hogan 1999). The terms under-resourced or low resource seem to have similar semantics. However, as Hammarström (2009) explains, it is unclear whether this is measured in absolute terms or relative to some other language. If the former, then simply creating resources in the language could cause its classification to change while in the latter changes to the resources available for other languages could affect the classification. Hammarström also introduces the alternative *low-affluence* defined via the metric of *Gross Language Product* (GLP) which is the product of the number of native speakers of the language in any country and the country's per capita Gross National Product. We add to this discussion the possibility that 'low' is, like 'critical', relative to some expectation based on the importance of the language. Hammarström's low-affluence has the advantage of a clear definition; however, its correlation with resource availability is imperfect as Figure 12 shows.

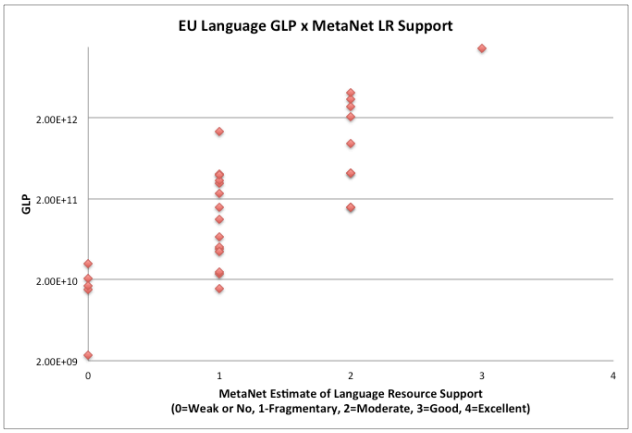


**Figure 1: Gross Language Product correlates moderately well with MetaNet's Estimate of Language Resource Support for European Languages**

The MetaNet White paper series classifies European languages into five categories of language resource availability: Excellent, Good, Moderate, Fragmentary, Weak/No. Correlating the two measures we find several

problems: for example, Portuguese has a higher GLP than Dutch, Swedish, Polish, Czech and Hungarian but fewer LRs and Lithuanian has a higher GLP but fewer resources than Serbian, Basque, and Estonian.

The term *less commonly taught* seems to have been borrowed in the HLT community from the second language teaching community where it refers to instruction within a specific target market, be it the United States, the Western Hemisphere or perhaps outside of the region where the language is official. The US National Council of Less Commonly Taught Languages (LCTL) described its focus as languages "*critically important to our national interest in the 21st century*" but not the "*French, German, Italian, or Spanish*" studied by 91% of US collegiates. The discussion names Arabic, Chinese, Japanese, Yoruba, Russian, Swahili as specific LCTLs. That reference to teaching in a specific market operates within US HLT programs where Hindi and Bengali were selected though they are the native language and/or the language of instructions for millions in India.

We also see the term *surprise language* used in relation to low resourced languages within several DARPA and IARPA HLT programs. Here 'surprise' does not refer to any inherent characteristic of the language though surprise languages have usually been low resource. Instead the term refers to a specific HLT research management technique to determine the extent to which systems are portable, and to estimate the time required to port to the language from a standing start as might be necessary in the event, for example, of a natural disaster.

For the remainder of this paper we will use the term low resource languages (LRL) by default to refer to those that have fewer technologies and especially data sets relative to some measure of their international importance.

## 3. Programs

The programs surveyed for this paper differed in goals and thus the languages studied. In this section, as background to the discussion of selection criteria, we sketch each.

### 3.1. TIDES

DARPA TIDES (Translingual Information Detection, Extraction and Summarization) was originally conceived and presented as an intensively multilingual program with multiple technology development goals. The program manager's brief from 1999 envisions at least query translation for 30 relevant languages but also machine translation, information retrieval and extraction and summarization for a subset. After some turn-over in project management, TIDES focused the bulk of its attention on English, (Mandarin) Chinese and (Modern Standard) Arabic but planned for Surprise Language Exercises. Of course, it is critical to remember that Chinese and Arabic were terribly under-resourced at the turn of the millennium and that it was the attention of programs like TIDES that increased the number, size and quality of available resources with the end result that they are now among the more richly resourced. The TIDES Surprise Language exercises were intended to evaluate the HLT community's ability to rapidly develop technologies for a low resource language with no prior warning as would be necessary as a response to a natural disaster. Performers, including LDC as the data provider, were given one month from the date the surprise language – Hindi as it happened – was announced to create best-of-breed TIDES technologies. In preparation for the exercise, LDC supplied program management with a table of language characteristics as described below and managed a dry run of the data collection activities focused on another low resource language, Cebuano. Again it is important to note that, like Mandarin Chinese and Modern Standard Arabic, the number of resources for Hindi has grown over the intervening years.

### 3.2. REFLEX LCTL

The US government sponsored the REFLEX (Research on English and Foreign Language Exploitation) LCTL program, which sought to create basic technologies in a number of low resource languages. Simpson et al. (2008) characterize the selected languages: "*Some of the languages (Thai, Urdu) were chosen to exercise a resource collection paradigm in which raw text is available digitally in sufficient quantity; others (Amazigh, Guarani, Maguindanao) were chosen to force the program to deal with cases in which it certainly is not. The cluster of Indic languages (Bengali, Punjabi, Urdu) was chosen to give researchers the opportunity to experiment with bootstrapping systems from material in related languages. Amazigh, Hungarian, Pashto, Tamil, and Yoruba were chosen to take advantage of existing collaborations in order to reduce costs. Finally there was a general desire to select languages that are quite different from each other and from well-resourced languages in order to maximize the generality of our methods. As a group, the LCTL languages are linguistically and geographically diverse …*"

### 3.3. NIST LRE

The US National Institute of Standards and Technologies (NIST) has organized Language Recognition [1] (LRE) technology evaluations since 1996 for which LDC has often provided data. LRE does not explicitly seek to work on low resource languages. However, since LRE's goal is to develop robust technologies that perform well even as the number of linguistic varieties increases, and since the number of well-resourced varieties is relatively small, it is inevitable that LRE would include low resource varieties. We use the term linguistic varieties because LRE requires performers to also distinguish confusable varieties including closely related languages and mutually intelligible dialects. The LRE selection process begins with a set of candidate varieties proposed by the US government sponsor from which the data provider selects

[1] http://www.nist.gov/itl/iad/mig/lre.cfm

a subset based on two types of criteria: confusability and a series of factors related to the probability of success in data collection. Data are typically segments of broadcast and telephone conversations audited for the linguistic variety spoken, speaker number and sex, and sound quality. Thus, the 'success' criteria deal with the availability, in the variety of interest, of the desired data types and native speakers capable of the annotation. The 2011 campaign included the following potentially confusable sets: Iraqi, Levantine, Maghreb and Modern Standard Arabic; American and Indian English; Czech, Polish, Russian, Slovak and Ukrainian; Dari and Persian; Bengali, Hindi, Punjabi and Urdu; and Thai and Lao plus Mandarin, Pashto, Spanish, Tamil, Turkish. The 2015 evaluation used the same selection procedures but added Egyptian to the Arabic cluster, British to the English cluster and created three new clusters: Chinese (Mandarin, Cantonese, Min Nan, Wu); Spanish-Portuguese (Brazilian Portuguese, Caribbean Spanish, European Spanish, Latin American Spanish) and French (Haitian Creole, West African French). It also reduced the Slavic cluster to Russian and Polish.

### 3.4. IARPA Babel

IARPA Babel [2] sought to escape what the program described as an English bias present in existing speech recognition technologies. Babel systems should be capable of building a keyword search system for audio in essentially any language in a very short timeline. Program challenges included multilingual speech recognition and keyword search under difficult conditions, including resource scarcity and noisy environments, with the capability to rapidly adapt to new languages and environments. Selection criteria included estimates of language importance or interest, linguistic factors including diversity and factors related to the ability to collect data within the designated schedule.

### 3.5. DARPA LORLEI

The LORELEI[3] program seeks to advance the state of the art in human language technologies to allow rapid porting to low resource languages for purposes of information awareness in the event of a disaster. To accomplish those goals LORELEI technologies include speech recognition, machine translation and the extraction of information including topics, entities and their relations to each other, events and sentiment. The program is creating language resources for 23 representative and 12 incident languages, the latter to be used for estimating system performance in the event of a disaster. For each of these, the program will create language packs, the composition of which differs for representative and incident languages and also depending on whether the latter have been chosen for evaluation. The range of language resources in a pack could include monolingual and parallel text; found dictionaries, grammars, gazetteers and primers; entity, morphological, syntactic and semantic annotations; morpheme level alignment of source and translation; text processing tools and entity taggers; lexicons and grammatical sketches; and test data including parallel text with entity and topic annotation for a portion of the documents.

## 4. Selection Criteria

Because US LRL programs generally work from a presumption that resource availability should be in proportion to some measure of language importance, many of the selection criteria deal with demographic factors and the current resource supply. As Simpson and colleagues (2008) reported for the REFLEX LCTL program: "*All meet the basic criteria of being significant in terms of the number of native speakers but poorly represented in terms of available language resources.*" Another major concern for these projects is the probability of success that is reflected partially in the former criteria types but also in the language typology and the availability of raw data in digital form.

### 4.1. Demographic

The population of native speakers as a raw number, rank or class (e.g. >1 million), either in the homeland or worldwide may stand as a proxy for the language's influence though the correlation is imperfect. English for example is 3rd behind Mandarin and Spanish in native speakers but probably greater in influence. Hammarström's GLP tries to correct for languages with many native speakers having less economic power. In 2009, he listed English as having the highest GLP with Spanish third and Mandarin seventh. GLP could be included among selection criteria of future project, with the caveats given in Section 2; however none of the programs surveyed used GLP explicitly. If we consider, retrospectively, the GLPs of languages included in the US HLT programs we see the expected variation in profile. Figure 3 charts the number of languages studied in several US HLT in categories according to their GLP. For purposes of comparison, official EU languages and 2015 critical languages are similar plotted. Categories of GLP are the log-scaled x-axis and the number of languages in that category on the vertical. The leftmost column shows the number of languages for which Hammaström's list of the 140 most affluent provides no GLP. As we see only one language, Welsh, has a GLP greater than 1 but less than 10 billion. One other, English, has a GLP greater than 10 trillion. Among the remaining GLP categories the critical languages list is evenly distributed, as are the LRE languages. TIDES focused most of its attention on the very affluent English, Mandarin and Modern Standard Arabic while the TIDES Surprise Languages were significantly less so. LCTL, Babel and LORELEI, like the languages of the EU, all tend toward the less affluent end of the scale, at least for languages whose GLP we know.

[2] http://www.iarpa.gov/index.php/research-programs/babel
[3] http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents

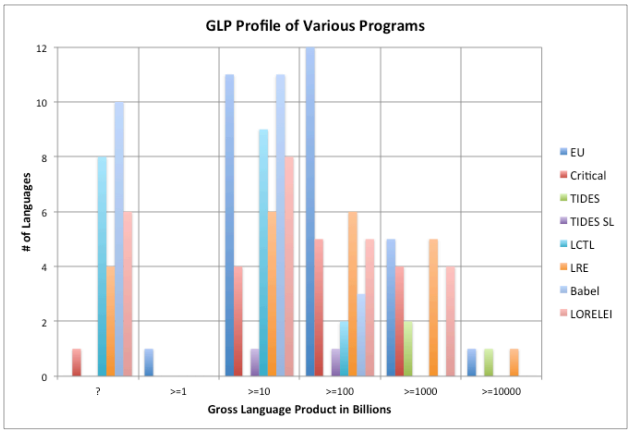


**Figure 3: Languages Selected by US HLT programs by GLP**

Notwithstanding the number of native speakers, if most of those also speak another language with an even greater population or prominence, that could reduce the language's importance to projects whose goal is to develop news understanding technologies. In preparation for the TIDES Surprise Language project we discussed this characteristic with program management and provided a table showing, for each language, whether a significant portion of its native speakers also spoke a language with a greater population of speakers. The rationale was that news transcription, translation and summarization technologies would do the most good when processing the languages in which the world's news is likely to appear. On the Italian peninsula, Napoletano-Calabrese, Sicilian, Piemontese, Venetian, Emiliano-Romagnolo and Ligurian are among Hammarström's 60 most affluent languages, scoring higher than Urdu, Vietnamese, Indonesian, multiple varieties of Arabic, Tagalog, Afrikaans, Yoruba and Latvian but lower than Italian by an order of magnitude. The very large percent of native speakers who also speak Italian and the probability that events of international importance taking place in Italy would like appear in the Italian language press seems to have contributed to the slow rate of technology development for the other languages of Italy.

In some cases, the most telling determinant of a language's importance is almost certainly the population of second language speakers or the total number of speakers. For example, Swahili is spoken by far more second language speakers than by first language speakers, and its importance as a regional language therefore outweighs its importance as a first language.

Finally, if a significant number of its speakers are currently involved in some international event, such as a natural disaster, that naturally increases the language's priority. The case of Haitian Creole comes immediately to mind.

## 4.2. Linguistic

The Ethnologue classification according to the family tree model provides information about the historical connections between languages that could prove useful in migrating HLTs. For example English and Frisian are both classified as: Indo-European, Germanic, West, sharing a closer relation than either do to, say, Danish which is: Indo-European, Germanic, North, East Scandinavian. Given the time and cost requested to develop HLTs, numerous researchers have focused on the challenges of porting or migrating specific HLTs from one language to another or on developing HLTs that are intended to be general requiring only training data in the target language in order to process that language. For example Vergyri et al. (2005) report "*We found that most of the techniques developed for English or ECA ASR could be ported to the development of a LCA system.*" The abbreviations ECA and LCA refer to Egyptian and Levantine Colloquial Arabic, respectively. Beyerlein et al. (1999) reported on experiments to create a speech recognition system for a low resource language, Czech, by augmenting its acoustic model with resources borrowed from other languages. Although the authors do not emphasize this point, the improvements gained by augmenting with acoustic models from a single language are greater for closely related Russian than for more distantly related English and Spanish and least for Mandarin. Elmahdy and colleagues (2014), working with more closely related varieties, report: "*Due to the limitation of dialectal speech resources, by utilizing MSA data, cross-dialectal phone mapping, data pooling, acoustic model adaptation and system combination methods, has achieved 21.3% and 28.9% relative WER reduction on QA development set and evaluation set respectively.*" The REFLEX LCTL program sought to encourage research into technology development across multiple closely related varieties by including several Indo-Aryan languages in the program: Bengali, Punjabi and Urdu. When data on mutually intelligibility is absent LRE has also used the family tree as a way to locate potentially confusable varieties. Of course, the family tree model by itself does not consider language contact phenomena such as the many borrowings from French into English that can also affect mutual intelligibility and comparability.

A number of other factors can affect the effort required to create certain LRs for the language, for example whether: the language is generally written by native speakers, its orthography is standardized, words and sentences are delimited in writing and the ease with which one may map written words to their pronunciations. Similarly the nature of the morphology affects HLT development, not only whether the language tends toward an analytic or synthetic morphology but also such factors as the number of morphological classes and the degree of irregularity and syncretism present.

Some LRL programs strive to develop general computational methods applicable to a variety of other languages. For such programs, it is therefore necessary to choose languages with typological diversity in phonology, morphology, syntax, etc., so that the computational

methods developed on the chosen languages can later be applied more broadly. Both REFLEX LCTL and IARPA Babel explicitly sought linguistic diversity.

### 4.3. Resource

There is a challenge in low resource language selection that involves actual resource availability. If the language has too few resources, the project could mire in LR creation. On the other hand if the language were too well resourced, the experience might not represent other low resource languages. It is therefore important before embarking on a project to pre-screen potential languages for the desired level of resource availability. All of the programs surveyed considered the range of raw data and existing LRs available: TIDES, LCTL, LRE, Babel and LORELEI.

In addition to the number of available LRs, programs might also look for specific types: news broadcasts from Voice of America which is public domain in the US, translations of religious texts such as the Bible, Qur'an or Book of Mormon, other commonly translated texts such as the Universal Declaration of the Rights of Man or indeed any translations. LRL programs may also pre-screen to determine if there exist newspapers, radio and TV broadcasts in the language, or more recently: webcasts, user contributed videos such as YouTube, informal user-generated content including blogs and discussion forums, micro-blogs as in Twitter or other social media. Beyond the raw resources they may look for dictionaries, gazetteers and grammatical descriptions.

In terms of human resources, the project may try to find a local speaker community, preferably literate, including students and especially an expert. Alternately, the project may seek partners in country or other conditions favorable to a successful remote collection such as pre-requisite infrastructure and incentives appropriate to the native speaker population.

Additional desired LRs might include a standard digital encoding, and supplies of news text, parallel text, translation dictionaries, name taggers, segmenters, and morph analyzers.

Finally LRL programs have different goals so that the criteria used and the weight given to each will vary. Nevertheless, sharing information about the criteria used in those programs will benefit the community.

## 5. Implementation Challenges

The selection criteria we have briefly described form a kind of superset of those we have seen used in US HLT programs focused on LRL. Not all were used in all programs and the criteria have also evolved over time. Even if selection criteria were identified explicitly, other challenges await. Classifications differ in how they determine what constitute a separate language. Languages have multiple names, some ambiguous (e.g. "He is speaking Creole/ Patois/ Dialect.") and some overlapping. The data on demographics, linguistic features and resource availability are difficult to collect and to weight. Furthermore demographics change – sometimes abruptly such as the number of Syrian Arabic speakers in Europe. Resource availability also changes. Fifteen years ago, Quechua had virtually no web presence beyond a few sites with some Bible passages. Today the Quechuan languages have a fairly substantial web presence including audio newscasts on YouTube. During the TIDES Surprise Language program, Quechua was not a viable option but some varieties might be today. Although the spread of Internet access has proven helpful in documenting some languages others have died during the web era and sites that host language data have disappeared.

## 6. Conclusions

We have sketched the criteria: demographic, linguistic and resource related that have been considered in the process of selecting linguistic varieties for study in several US low resource language, common task HLT programs. The inventory of factors to be considered have varied by program as, apparently, has the weighting given to each. Nonetheless we see that programs balance resource availability against some measure of the languages importance or interest. They may also consider linguistic factors especially those that permit the selection or either highly confusable or typologically diverse languages or both. By documenting these criteria we hope to open discussion concerning selection criteria in low resource language programs so that future project may build on the early work surveyed here.

## 7. Acknowledgements

Several low resource language programs: TIDES, REFLEX LCTL, Babel and LORELEI provided the opportunity to develop much of the information summarized here. For LORELEI, this material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0123. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or any sponsoring agency.

## 8. References

Beyerlein, P., W. Byrne, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, W. Wang. 1999. Towards Language Independent Acoustic Modeling. IEEE Workshop on Automatic Speech Recognition and Understanding, December 12 - 15, Keystone, Colorado, U.S.A

Dorian, Nancy C. 1980. Language shift in community and individual: The phenomenon of the laggard semi-speaker. International Journal of the Sociology of Language 25.85-94.

Elmahdy, Mohamed, Mark Hasegawa-Johnson, and Eiman Mustafawi, "Development of a tv broadcasts

speech recognition system for Qatari Arabic," in The 9th edition of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, 2014.
Fishman, Joshua A. 1991. Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages. Clevedon, UK: Multilingual Matters.
Hammarström, H.. A survey of computational morphological resources for low-density languages. Journal of the NEALT, 2009.
Hogan, Christopher. 1999. "OCR for Minority Languages" pp. 235-244 in David Doermann (ed.) "Proceedings SDIUT 1999: The 1999 Symposium on Document Image Understanding Technology." University of Maryland Institute for Advanced Computer Studies, College Park MD.
Krauss, Michael. 1992. The world's languages in crisis. Language 68(1).1-42.
Lewis, M. Paul and Gary F. Simons. 2010. Assessing Endangerment: Expanding Fishman's GIDS. Revue Roumaine de Linguistique 55(2):103–120. http://www.lingv.ro/RRL 2 2010 art01Lewis.pdf. Accessed March 21, 2011.
Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2015. Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.
Maxwell, Mike, Baden Hughes. 2006. Frontiers in Linguistic Annotation for Lower-Density Languages in Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006, Sydney, Australia, Association for Computational Linguistics, pp. 29—37, URL: http://www.aclweb.org/anthology/W/W06/W06-0605
Megerdoomian, Karine, Dan Parvaz, 2008. Low-density language bootstrapping: The case of Tajiki Persian. In Proceedings of LREC 2008. Marrakech, Morocco, May 2008.
METANET. 2010. META-NET White Paper Series: Press Release, http://www.meta-net.eu/whitepapers/press-release-en, accessed March 16, 2016.
National Science Foundation Documenting Endangered Languages Program. 2014. Press Release 14-098: Federal agencies provide new opportunities for dying languages, August 15, 2014, http://www.nsf.gov/news/news_summ.jsp?cntn_id=132370, accessed March 16, 2016.
Rehm, Georg, Hans Uszkoreit, eds. 2012. META-NET White Paper Series: Europe's Languages in the Digital Age, URL: www.meta-net.eu/whitepapers
Simpson, Heather, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, Boyan Onyshkevych. 2008. Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources, paper presented at the SALTMIL Workshop: Free/Open-Source Language Resources for the Machine Translation of Less-Resourced Languages satellite to the 7th International Conference on Language Resources and Evaluation, Marrakesh, May 28-30
Vergyri, D., K. Kirchhoff, R. Gadde, A. Stolcke, J. Zheng. 2005. Development Of A Conversational Telephone Speech Recognizer For Levantine Arabic. Proceedings of Interspeech, Lisboa, Portugal.