Frontiers in Linguistic Annotation for Lower-Density Languages

Mike Maxwell
Center for Advanced Study of Language
University of Maryland
mmaxwell@casl.umd.edu

Baden Hughes
Department of Computer Science
The University of Melbourne
badenh@csse.unimelb.edu.au

Abstract

The languages that are most commonly subject to linguistic annotation on a large scale tend to be those with the largest populations or with recent histories of linguistic scholarship. In this paper we discuss the problems associated with lower-density languages in the context of the development of linguistically annotated resources. We frame our work with three key questions regarding the definition of lower-density languages; increasing available resources and reducing data requirements. A number of steps forward are identified for increasing the number lower-density language corpora with linguistic annotations.

1 Introduction

The process for selecting a target language for research activity in corpus linguistics, natural language processing or computational linguistics is largely arbitrary. To some extent, the motivation for a specific choice is based on one or more of a range of factors: the number of speakers of a given language; the economic and social dominance of the speakers; the extent to which computational and/or lexical resources already exist; the availability of these resources in a manner conducive to research activity; the level of geopolitical support for language-specific activity, or the sensitivity of the language in the political arena; the degree to which the researchers are likely to be appreciated by the speakers of the language simply because of engagement; and the potential scientific returns from working on the language in question (including the likelihood that the language exhibits interesting or unique phenomena). Notably, these factors are also significant in determining whether a language is worked on for documentary and descriptive purposes, although an additional factor in this particular area is also the degree of endangerment (which can perhaps be contrasted with the likelihood of economic returns for computational endeavour).

As a result of these influencing factors, it is clear that languages which exhibit positive effects in one or more of these areas are likely to be the target of computational research. If we consider the availability of computationally tractable language resources, we find, unsuprisingly that major languages such as English, German, French and Japanese are dominant; and research on computational approaches to linguistic analysis tends to be farthest advanced in these languages.

However, renewed interest in the annotation of lower-density languages has arisen for a number of reasons, both theoretical and practical. In this paper we discuss the problems associated with lower-density languages in the context of the development of linguistically annotated resources.

The structure of this paper is as follows. First we define the lower-density languages and linguistically annotated resources, thus defining the scope of our interest. We review some related work in the area of linguistically annotated corpora for lower-density languages. Next we pose three questions which frame the body of this paper: What is the current status of in terms of lower-density languages which have linguistically annotated corpora? How can we more efficiently create this particular type of data for lower-density languages? Can existing analytical methods methods perform reliably with less data? A number of steps are identified for advancing the agenda of linguis-

Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006, pages 29–37,
Sydney, July 2006.

tically annotated resources for lower-density languages, and finally we draw conclusions.

2 Lower-Density Languages

It should be noted from the outset that in this paper we interpret 'density' to refer to the amount of computational resources available, rather than the number of speakers any given language might have.

The fundamental problem for annotation of lower-density languages is that they are lower-density. While on the surface, this is a tautology, it in fact is the problem. For a few languages of the world (such as English, Chinese and Modern Standard Arabic, and a few Western European languages), resources are abundant; these are the high-density Languages. For a few more languages (other European languages, for the most part), resources are, if not exactly abundant, at least existent, and growing; these may be considered medium-density languages. Together, high-density and medium-density languages account for perhaps 20 or 30 languages, although of course the boundaries are arbitrary. For all other languages, resources are scarce and hence they fall into our specific area of interest.

3 Linguistically Annotated Resources

While the scarcity of language resources for lower-density languages is apparent for all resource types (with the possible exception of monolingual text), it is particularly true of linguistically annotated texts. By annotated texts, we include the following sorts of computational linguistic resources:

- Parallel text aligned with another language at the sentence level (and/or at finer levels of parallelism, including morpheme-level glossing)
- Text annotated for named entities at various levels of granularity
- Morphologically analyzed text (for non-isolating languages; at issue here is particularly inflectional morphology, and to a lesser degree of importance for most computational purposes, derivational morphology); also a morphological tag schema appropriate to the particular language
- Text marked for word boundaries (for those scripts which, like Thai, do not mark most word boundaries)
- POS tagged text, and a POS tag schema appropriate to the particular language
- Treebanked (syntactically annotated and parsed) text
- Semantically tagged text (semantic roles) cf. Propbank (Palmer et al., 2005), or frames cf. Framenet[1]
- Electronic dictionaries and other lexical resources, such as Wordnet[2]

There are numerous dimensions for linguistically annotated resources, and a range of research projects have attempted to identify the core properties of interest. While concepts such as the Basic Language Resource Kit (BLARK; (Krauwer, 2003; Mapelli and Choukri, 2003)) have gained considerable currency in higher-density language resource creation projects, it is clear that the baseline requirements of such schemes are significantly more advanced than we can hope for for lower-density languages in the short to medium term. Notably, the concept of a reduced BLARK ('BLARKette') has recently gained some currency in various forums.

4 Key Questions

Given that the vast majority of the more than seven thousand languages documented in the Ethnologue (Gordon, 2005) fall into the class of lower-density languages, what should we do? Equally important, what can we realistically do? We pose three questions by which to frame the remainder of this paper.

1. **Status Indicators**: How do we know where we are? How do we keep track of what languages are high-density or medium-density, and which are lower-density?
2. **Increasing Available Resources**: How (or can) we encourage the movement of languages up the scale from lower-density to medium-density or high-density?

[1] http://framenet.icsi.berkeley.edu/
[2] http://wordnet.princeton.edu

3. **Reducing Data Requirements**: Given that some languages will always be relatively lower-density, can language processing applications be made smarter, so that they don't require largely unattainable resources in order to perform adequately?

5 Status Indicators

We have been deliberately vague up to this point about how many lower-density languages there are, or the simpler question, how my high and medium density languages there are. Of course one reason for this is that the boundary between low density and medium or high density is inherently vague. Another reason is that the situation is constantly changing; many Central and Eastern European languages which were lower-density languages a decade or so ago are now arguably medium density, if not high density. (The standard for high vs. low density changes, too; the bar is considerably higher now than it was ten years ago.)

But the primary reason for being vague about how many – and which – languages are low density today is that no is keeping track of what resources are available for most languages. So we simply have no idea which languages are low density, and more importantly (since we can guess that in the absence of evidence to the contrary, a language is likely to be low density), we don't know which resource types most languages do or do not have.

This lack of knowledge is not for lack of trying, although perhaps we have not been trying hard enough. The following are a few of the catalogs of information about languages and their resources that are available:

- The Ethnologue[3]: This is the standard listing of the living languages of the world, but contains little or no information about what resources exist for each language.

- LDC catalog[4] and ELDA catalog[5]: The Linguistic Data Consortium (LDC) and the European Language Resources Distribution Agency (ELDA) have been among the largest distributors of annotated language data. Their catalogs, naturally, cover only those corpora distributed by each organization, and these include only a small number of languages. Naturally, the economically important languages constitute the majority of the holdings of the LDC and ELDA.

- AILLA (Archive of the Indigenous Languages of Latin America[6]), and numerous other language archiving sites: Such sites maintain archives of linguistic data for languages, often with a specialization, such as indigenous languages of a country or region. The linguistic data ranges from unannotated speech recordings to morphologically analyzed texts glossed at the morpheme level.

- OLAC (Open Archives Language Community[7]): Given that many of the above resources (particularly those of the many language archives) are hard to find, OLAC is an attempt to be a meta-catalog (or aggregator)of such resources. It allows lookup of data by type, language etc. for all data repositories that 'belong to' OLAC. In fact, all the above resources are listed in the OLAC union catalogue.

- Web-based catalogs of additional resources: There is a huge number of additional websites which catalog information about languages, ranging from electronic and print dictionaries (e.g. yourDictionary[8]), to discussion groups about particular languages[9]. Most such sites do little vetting of the resources, and dead links abound. Nevertheless, such sites (or a simple search with an Internet search engine) can often turn up useful information (such as grammatical descriptions of minority languages). Very few of these web sites are cataloged in OLAC, although recent efforts (Hughes et al., 2006a) are slowly addressing the inclusion of web-based low density language resources in such indexes.

None of the above catalogs is in any sense complete, and indeed the very notion of completeness is moot when it comes to cataloging Internet resources. But more to the point of this paper, it

[3] http://www.ethnologue.org
[4] http://www.ldc.upenn.edu/Catalog/
[5] http://www.elda.org/rubrique6.html
[6] http://www.ailla.utexas.org
[7] http://www.language-archives.org
[8] http://www.yourdictionary.com
[9] http://dir.groups.yahoo.com/dir/Cultures_Community/By_Language

is difficult, if not impossible, to get a picture of the state of language resources in general. How many languages have sufficient bitext (and in what genre), for example, that one could put together a statistical machine translation system? What languages have morphological parsers (and for what languages is such a parser more or less irrelevant, because the language is relatively isolating)? Where can one find character encoding converters for the Ge'ez family of fonts for languages written in Ethiopic script?

The answer to such questions is important for several reasons:

1. If there were a crisis that involved an arbitrary language of the world, what resources could be deployed? An example of such a situation might be another tsunami near Indonesia, which could affect dozens, if not hundreds of minority languages. (The December 26, 2004 tsunami was particularly felt in the Aceh province of Indonesia, where one of the main languages is Aceh, spoken by three million people. Aceh is a lower-density language.)

2. Which languages could, with a relatively small amount of effort, move from lower-density status to medium-density or high-density status? For example, where parallel text is harvestable, a relatively small amount of work might suffice to produce many applications, or other resources (e.g. by projecting syntactic annotation across languages). On the other hand, where the writing system of a language is in flux, or the language is politically oppressed, a great deal more effort might be necessary.

3. For which low density languages might related languages provide the leverage needed to build at least first draft resources? For example, one might think of using Turkish (arguably at least a medium-density language) as a sort of pivot language to build lexicons and morphological parsers for such low density Turkic languages as Uzbek or Uyghur.

4. For which low density languages are there extensive communities of speakers living in other countries, who might be better able to build language resources than speakers living in the perhaps less economically developed home countries? (Expatriate communities may also be motivated by a desire to maintain their language among younger speakers, born abroad.)

5. Which languages would require more work (and funding) to build resources, but are still plausible candidates for short term efforts?

To our knowledge, there is no general, on-going effort to collect the sort of data that would make answers to these questions possible. A survey was done at the Linguistic Data Consortium several years ago (Strassel et al., 2003) , for text-based resources for the three hundred or so languages having at least a million speakers (an arbitrary cutoff, to be sure, but necessary for the survey to have had at least some chance of success). It was remarkably successful, considering that it was done by two linguists who did not know the vast majority of the languages surveyed. The survey was funded long enough to 'finish' about 150 languages, but no subsequent update was ever done.

A better model for such a survey might be an edited book: one or more computational linguists would serve as 'editors', responsible for the overall framework, and training of other participants. Section 'editors' would be responsible for a language family, or for the languages of a geographic region or country. Individual language experts would receive a small amount of training to enable them to answer the survey questions for their language, and then paid to do the initial survey, plus periodic updates. The model provided by the Ethnologue (Gordon, 2005) may serve as a starting point, although for the level of detail that would be useful in assessing language resource availability will make wholesale adoption unsuitable.

6 Increasing Available Resources

Given that a language significantly lacks computational linguistic resources (and in the context of this paper and the associated workshop, annotated text resources), so that it falls into the class of lower-density languages (however that might be defined), what then?

Most large-scale collections of computational linguistics resources have been funded by government agencies, either the US government (typically the Department of Defense) or by governments of countries where the languages in question are spoken (primarily European, but also a

few other financially well-off countries). In some cases, governments have sponsored collections for languages which are not indigenous to the country in question (e.g. the EMILLE project[10], see (McEnery et al., 2000)).

In most such projects, production of resources for lower-density languages have been the work of a very small team which oversees the effort, together with paid annotators and translators. More specifically, collection and processing of monolingual text can be done by a linguist who need not know the language (although it helps to have a speaker of the language who can be called on to do language identification, etc.). Dictionary collection from on-line dictionaries can also be done by a linguist; but if it takes much more effort than that – for example, if the dictionary needs to be converted from print format to electronic format – it is again preferable to have a language speaker available.

Annotating text (e.g. for named entities) is different: it can only be done by a speaker of the language (more accurately, a reader: for Punjabi, for instance, it can be difficult to find fluent readers of the Gurmukhi script). Preferably the annotator is familiar enough with current events in the country where the language is spoken that they can interpret cross-references in the text. If two or more annotators are available, the work can be done somewhat more quickly. More importantly, there can be some checking for inter-annotator agreement (and revision taking into account such differences as are found).

Earlier work on corpus collection from the web (e.g. (Resnik and Smith, 2003)) gave some hope that reasonably large quantities of parallel text could be found on the web, so that a bitext collection could be built for interesting language pairs (with one member of the pair usually being English) relatively cheaply. Subsequent experience with lower-density languages has not born that hope out; parallel text on the web seems relatively rare for most languages. It is unclear why this should be. Certainly in countries like India, there are large amounts of news text in English and many of the target languages (such as Hindi). Nevertheless, very little of that text seems to be genuinely parallel, although recent work (Munteanu and Marcu, 2005) indicates that true parallelism may not be required for some tasks, eg machine translation, in order to gain acceptable results.

Because bitext was so difficult to find for lower-density languages, corpus creation efforts rely largely, if not exclusively, on contracting out text for translation. In most cases, source text is harvested from news sites in the target language, and then translated into English by commercial translation agencies, at a rate usually in the neighborhood of US$0.25 per word. In theory, one could reduce this cost by dealing directly with translators, avoiding the middleman agencies. Since many translators are in the Third World, this might result in considerable cost savings. Nevertheless, quality control issues loom large. The more professional agencies do quality control of their translations; even so, one may need to reject translations in some cases (and the agencies themselves may have difficulty in dealing with translators for languages for which there is comparatively little demand). Obviously this overall cost is high; it means that a 100k word quantity of parallel text will cost in the neighborhood of US$25K.

Other sources of parallel text might include government archives (but apart from parliamentary proceedings where these are published bilingually, such as the Hansards, these are usually not open), and the archives of translation companies (but again, these are seldom if ever open, because the agencies must guard the privacy of those who contracted the translations).

Finally, there is the possibility that parallel text – and indeed, other forms of annotation – could be produced in an open source fashion. Wikipedia[11] is perhaps the most obvious instance of this, as there are parallel articles in English and other languages. Unfortunately, the quantity of such parallel text at the Wikipedia is very small for all but a few languages. At present (May 2006), there are over 100,000 articles in German, Spanish, French, Italian, Japanese, Dutch, Polish, Portuguese and Swedish.[12] Languages with over 10,000 articles include Arabic, Bulgarian, Catalan, Czech, Danish, Estonian, Esperanto and Ido (both constructed languages), Persian, Galician, Hebrew, Croatian), Bahasa Indonesian, Korean, Lithuanian, Hungarian, Bahasa Malay, Norwegian

[10] http://bowland-files.lancs.ac.uk/corplang/emille/

[11] http://en.wikipedia.org

[12] Probably some of these articles are non-parallel. Indeed, a random check of Cebuano articles in Wikipedia revealed that many were stubs (a term used in the Wikipedia to refer to "a short article in need of expansion"), or were simply links to Internet blogs, many of which were monolingual in English.

(Bokmál and Nynorsk), Romanian, Russian, Slovak, Slovenian, Serbian, Finnish, Thai, Turkish, Ukrainian, and Chinese. The dominance of European languages in these lists is obvious.

During a TIDES exercise in 2003, researchers at Johns Hopkins University explored an innovative approach to the creation of bitext (parallel English and Hindi text, aligned at the sentence level): they elicited translations into English of Hindi sentences they posted on an Internet web page (Oard, 2003; Yarowsky, 2003). Participants were paid for the best translations in Amazon.com gift certificates, with the quality of a twenty percent subset of the translations automatically evaluated using BLEU scores against highly scored translations of the same sentences from previous rounds. This pool of high-quality translations was initialized to a set of known quality translations. A valuable side effect of the use of previously translated texts for evaluation is that this created a pool of multiply translated texts.

The TIDES translation exercise quickly produced a large body of translated text: 300K words, in five days, at a cost of about two cents per word.

This approach to resource creation is similar to numerous open source projects, in the sense that the work is being done by the public. It differed in that the results of this work were not made publicly available; the use of an explicit quality control method; and of course the payments to (some) participants. While the quality control aspect may be essential to producing useful language resources, hiding those resources not currently being used for evaluation is not essential to the methodology.

Open source resource creation efforts are of course common, with the Wikipedia[13] being the best known. Other such projects include Amazon.com's Mechanical Turk[14], LiTgloss[15], The ESP Game[16], and the Wiktionary[17]. Clearly some forms of annotation will be easier to do using an open source methodology than others will. For example, translation and possibly named entity annotation might be fairly straightforward, while morphological analysis is probably more difficult, particularly for morphologically complex languages.

[13]http://www.wikipedia.org
[14]http://www.mturk.com/mturk/
[15]http://litgloss.buffalo.edu/
[16]http://www.espgame.org/
[17]http://wiktionary.org/

Other researchers have experimented with the automatic creation of corpora using web data (Ghani et al., 2001) Some of these corpora have grown to reasonable sizes; (Scannell, 2003; Scannell, 2006) has corpora derived from web crawling which are measured in tens of millions of words for a variety of lower-density languages. However it should be noted that in these cases, the type of linguistic resource created is often not linguistically annotated, but rather a lexicon or collection of primary texts in a given language.

Finally, we may mention efforts to create certain kinds of resources by computer-directed elicitation. Examples of projects sharing this focus include BOAS (Nirenburg and Raskin, 1998), and the AVENUE project (Probst et al., 2002), (Lavie et al., 2003).

7 Reducing Data Requirements

Creating more annotated resources is the obvious way to approach the problem of the lack of resources for lower-density languages. A complementary approach is to improve the way the information in smaller resources is used, for example by developing machine translation systems that require less parallel text.

How much reduction in the required amount of resources might be enough? An interesting experiment, which to our knowledge has never been tried, would be for a linguist to attempt as a test case what we hope that computers can do. That is, a linguist could take a 'small' quantity of parallel text, and extract as much lexical and grammatical information from that as possible. The linguist might then take a previously unseen text in the target language and translate it into English, or perform some other useful task on target language texts. One might argue over whether this experiment would constitute an upper bound on how much information could be extracted, but it would probably be more information than current computational approaches extract.

Naturally, this approach partially shifts the problem from the research community interested in linguistically annotated corpora to the research community interested in algorithms. Much effort has been invested in scaling algorithmic approaches upwards, that is, leveraging every last available data point in pursuit of small performance improvements. We argue that scaling down (ie using less training data) poses an equally sig-

nificant challenge. The basic question of whether methods which are data-rich can scale down to impoverished data has been the focus of a number of recent papers in areas such as machine translation (Somers, 1997; Somers, 1998), language identification (Hughes et al., 2006b) etc. However, tasks which have lower-density language at their core have yet to become mainstream in shared evaluation tasks which drive much of the algorithmic improvements in computational linguistics and natural language processing.

Another approach to data reduction is to change the type of data required for a given task. For many lower-density languages a significant volume of linguistically annotated data exists, but not in the form of the curated, standardised corpora to which language technologists are accustomed. Neverthless for extremely low density languages, a degree of standardisation is apparent by virtue of documentary linguistic practice. Consider for example, the number of Shoebox lexicons and corresponding interlinear texts which are potentially available from documentary sources: while not being the traditional resource types on which systems are trained, they are reasonably accessible, and cover a larger number of languages. Bible translations are another form of parallel text available in nearly every written language (see (Resnik et al., 1999)). There are of course issues of quality, not to mention vocabulary, that arise from using the Bible as a source of parallel text, but for some purposes – such as morphology learning – Bible translations might be a very good source of data.

Similarly, a different compromise may be found in the ratio of the number of words in a corpus to the richness of linguistic annotation. In many high-density corpora development projects, an arbitrary (and high) target for the number of words is often set in advance, and subsequent linguistic annotation is layered over this base corpus in a progressively more granular fashion. It may be that this corpus development model could be modified for lower-density language resource development: we argue that in many cases, the richness of linguistic annotation over a given set of data is more important than the raw quantity of the data set.

A related issue is different standards for annotating linguistic concepts We already see this in larger languages (consider the difference in morpho-syntactic tagging between the Penn Treebank and other corpora), but has there is a higher diversity of standards in lower-density languages. Solutions may include ontologies for linguistic concepts e.g. General Ontology for Linguistic Description[18] and the ISO Data Category Registry (Ide and Romary, 2004), which allow cross-resource navigation based on common semantics. Of course, cross-language and cross-cultural semantics is a notoriously difficult subject.

Finally, it may be that development of web based corpora can act as the middle ground: there are plenty of documents on the web in lower-density languages, and efforts such as projects by Scannell[19] and Lewis[20] indicate these can be curated reasonably efficiently, even though the outcomes may be slightly different to that which we are accustomed. Is it possible to make use of XML or HTML markup directly in these cases? Someday, the semantic web may help us with this type of approach.

8 Moving Forward

Having considered the status of linguistically-annotated resources for lower-density languages, and two broad strategies for improving this situation (innovative approaches to data creation, and scaling down of resource requirements for existing techniques), we now turn to the question of where to go from here. We believe that there are a number of practical steps which can be taken in order to increase the number of linguistically-annotated lower-density language resources available to the research community:

- Encouraging the publication of electronic corpora of lower-density languages: most economic incentives for corpus creation only exhibit return on investment because of the focus on higher-density languages; new models of funding and commercializing corpora for lower-density languages are required.

- Engaging in research on bootstrapping from higher density language resources to lower-density surrogates: it seems obvious that at least for related languages adopting a derivational approach to the generation of linguistically annotated corpora for lower-density languages by using automated annotation tools trained on higher-density lan-

[18] http://www.linguistics-ontology.org
[19] http://borel.slu.edu/crubadan/stadas.html
[20] http://www.csufresno.edu/odin

guages may at least reduce the human effort required.

- Scaling down (through data requirement reduction) of state of the art algorithms: there has been little work in downscaling state of the art algorithms for tasks such as named entity recognition, POS tagging and syntactic parsing, yet (considerably) reducing the training data requirement seems like one of the few ways that existing analysis technologies can be applied to lower-density languages.

- Shared evaluation tasks which include lower-density languages or smaller amounts of data: most shared evaluation tasks are construed as exercises in cross-linguistic scalability (eg CLEF) or data intensivity (eg TREC) or both (eg NTCIR). Within these constructs there is certainly room for the inclusion of lower-density languages as targets, although notably the overhead here is not in the provision of the language data, but the derivatives (eg query topics) on which these exercises are based.

- Promotion of multilingual corpora which include lower-density languages: as multilingual corpora emerge, there is opportunity to include lower-density languages at minimal opportunity cost e.g. EuroGOV (Sigurbjönsson et al., 2005) or JRC-Acquis (Steinberger et al., 2006), which are based on web data from the EU, includes a number of lower-density languages by virtue of the corpus creation mechanism not being language-specific.

- Language specific strategies: collectively we have done well at developing formal strategies for high density languages e.g. in EU roadmaps, but not so well at strategies for medium-density or lower-density languages. The models for medium to long term strategies of language resource development may be adopted for lower density languages. Recently this has been evidenced through events such as the LREC 2006 workshop on African language resources and the development of a corresponding roadmap.

- Moving towards interoperability between annotation schemes which dominate the higher-density languages (eg Penn Treebank tagging conventions) and the relatively ad-hoc schemes often exhibited by lower-density languages, through means such as markup ontologies like the General Ontology for Linguistic Description or the ISO Data Category Registry.

Many of these steps are not about to be realised in the short term. However, developing a cohesive strategy for addressing the need for linguistically annotated corpora is a first step in ensuring committment from interested researchers to a common roadmap.

9 Conclusion

It is clear that the number of linguistically-annotated resources for any language will inevitably be less than optimal. Regardless of the density of the language under consideration, the cost of producing linguistically annotated corpora of a substantial size is significant, Inevitably, languages which do not have a strong political, economic or social status will be less well resourced.

Certain avenues of investigation e.g. collecting language specific web content, or building approximate bitexts web data are being explored, but other areas (such as rich morphosyntactic annotation) are not particularly evidenced.

However, there is considerable research interest in the development of linguistically annotated resources for languages of lower density. We are encouraged by the steady rate at which academic papers emerge reporting the development of resources for lower-density language targets. We have proposed a number of steps by which the issue of language resources for lower-density languages may be more efficiently created and look forward with anticipation as to how these ideas motivate future work.

References

Rayid Ghani, Rosie Jones, and Dunja Mladenic. 2001. Mining the web to create minority language corpora. In *Proceedings of 2001 ACM International Conference on Knowledge Management (CIKM2001)*, pages 279–286. Association for Computing Machinery.

Raymond G. Gordon. 2005. *Ethnologue: Languages of the World (15th Edition)*. SIL International: Dallas.

Baden Hughes, Timothy Baldwin, and Steven Bird. 2006a. Collecting low-density language data on the web. In *Proceedings of the 12th Australasian Web Conference (AusWeb06)*. Southern Cross University.

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. 2006b. Reconsidering language identification for written language resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*. European Language Resources Association: Paris.

Nancy Ide and Laurent Romary. 2004. A registry of standard data categories for linguistic annotation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004*, pages 135–139. European Language Resources Association: Paris.

Steven Krauwer. 2003. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of 2nd International Conference on Speech and Computer (SPECOM2003)*.

A. Lavie, S. Vogel, L. Levin, E. Peterson, K. Probst, A. Font Llitjos, R. Reynolds, J. Carbonell, and R. Cohen. 2003. Experiments with a hindi-to-english transfer-based mt system under a miserly data scenario. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2).

Valerie Mapelli and Khalid Choukri. 2003. Report on a monimal set of language resources to be made available for as many languages as possible, and a map of the actual gaps. ENABLER internal project report (Deliverable 5.1).

Tony McEnery, Paul Baker, and Lou Burnard. 2000. Corpus resources and minority language engineering. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2002)*. European Language Resources Association: Paris.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Sergei Nirenburg and Victor Raskin. 1998. Universal grammar and lexis for quick ramp-up of mt. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 975–979. Association for Computational Linguistics.

Douglas W. Oard. 2003. The surprise language exercises. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):79–84.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1).

K. Probst, L. Levin, E. Peterson, A. Lavie, and J. Carbonell. 2002. Mt for minority languages using elicitation-based learning of syntactic transfer rules. *Machine Translation*, 17(4):245–270.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a parallel corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33(1-2):129–153.

Kevin Scannell. 2003. Automatic thesaurus generation for minority languages: an irish example. In *Actes des Traitement Automatique des Langues Minoritaires et des Petites Langues*, volume 2, pages 203–212.

Kevin Scannell. 2006. Machine translation for closely related language pairs. In *Proceedings of the LREC2006 Workshop on Strategies for developing machine translation for minority languages*. European Language Resources Association: Paris.

B. Sigurbjönsson, J. Kamps, and M. de Rijke. 2005. Blueprint of a cross-lingual web collection. *Journal of Digital Information Management*, 3(1):9–13.

Harold Somers. 1997. Machine translation and minority languages. *Translating and the Computer*, 19:1–13.

Harold Somers. 1998. Language resources and minority languages. *Language Today*, 5:20–24.

R. Steinberger, B. Pouliquen, A. Widger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*. European Language Resources Association: Paris.

Stephanie Strassel, Mike Maxwell, and Christopher Cieri. 2003. Linguistic resource creation for research and technology development: A recent experiment. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(2):101–117.

David Yarowsky. 2003. Scalable elicitation of training data for machine translation. *Team Tides*, 4.

10 Acknowledgements

The authors are grateful to Kathryn L. Baker for her comments on earlier drafts of this paper.

Portions of the research in this paper were supported by the Australian Research Council Special Research Initiative (E-Research) grant number SR0567353 "An Intelligent Search Infrastructure for Language Resources on the Web."