Berkeley Linguistics Society. 2013. 273-286. http://dx.doi.org/10.3765/bls.v39i1.3886
Published by the Linguistic Society of America

Automatic Extraction of Linguistic Data from Digitized Documents

TERRENCE SZYMANSKI[1]
University of Michigan

Introduction

This paper presents a system for automatically extracting linguistic data from digitized linguistic documents using a combination of existing software packages and custom scripts. The system is designed to leverage existing resources in online digital libraries in order to bootstrap the creation of large, multi-lingual linguistic corpora, which can then be used to conduct data-driven experimental research into cross-linguistic or universal linguistic phenomena. The system identifies instances of foreign-language text accompanied by reference-language translations within the text of printed books that have been scanned into digital format, and extracts these to produce a parallel corpus of example sentences. While the system achieves a high precision on predicting foreign text, its accuracy overall is low, and directions for improvement and future work are identified.

1 Background and Objectives

1.1 Motivation

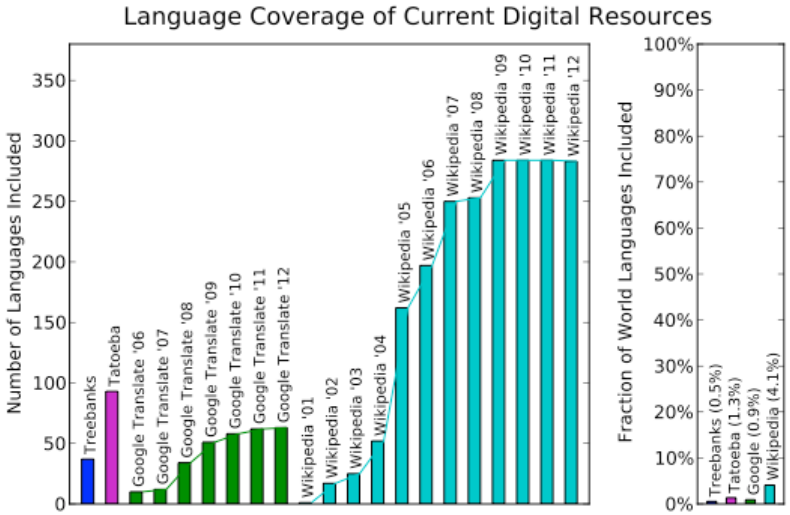
The increasing availability of large amounts of linguistic data in digital form, combined with the development of computational methods for analyzing such data, leads naturally to the question of what can be learned about the nature of language from analyzing large, multi-lingual corpora. John Goldsmith, advocating

[1] I would like to thank and acknowledge Steven Abney for his guidance and involvement in this project. Work on this project was partially funded by a Google Digital Humanities grant.

for the use of formal, mathematical models of grammar in linguistics, motivates his approach with the observation that "the goal of the linguist is to provide the most compact overall description of *all of the linguistic data that exists at present* [emphasis added]" (Goldsmith 2007). Steven Abney, similarly arguing for the use of computational methods to study the fundamental questions that linguists ask, writes "Any experimental foray into universal linguistics will be a data-intensive undertaking. It will require substantial samples of many languages—*ultimately all human languages* [emphasis added]—in a consistent form that supports automated processing across languages." (Abney 2011).

Both of these quotations emphasize that in order for a computational analysis of language to yield truly universal linguistic insights, the analysis must be performed on a data set that represents the full linguistic diversity that exists on this planet. However, the number of the world's languages currently represented in machine-readable corpora readily available online falls well short of the total number of languages currently spoken. Figure (1) below illustrates the comparative numbers of languages available in a variety of corpus types, and also compares these numbers to the total number of languages spoken around the globe.

(1) The current state of language resources available in digital form.



The data sources in (1) represent decreasing levels of annotation from left to right. Treebanks, used to train syntactic parsers, are corpora that have been manually annotated with phrase structure trees. Parallel corpora, of which Tatoeba[2] is one example, pair text from two languages and are essential for

[2] http://tatoeba.org/

training machine translation systems, of which Google Translate[3] is one example. Monolingual corpora, represented here by Wikipedia,[4] are the most abundantly available but also of the least use to linguists because they lack any linguistic annotation or reference outside of the text itself. While the number of languages represented in these resources has grown significantly in recent years, these totals are but a small fraction of the world's total languages, as illustrated in the smaller chart on the right-hand side.

In addition to the data sources included in (1), there is much more data that exists in digital form, but is not in a machine-readable format. This is a crucial distinction to make, because while such resources may be immensely useful to human linguists, they are useless from a computational linguist's perspective, at least until they have been converted in some way into a more processing-friendly format. The objective of this project is to explore the potential for automated methods to extract relevant linguistic data from online digital sources, converting that data into a machine-readable format that can then be used as a data source for computational linguistic research.

1.2 System Overview

The system proposed and described in this paper takes as its input digitized books from online sources, and produces as output a machine-readable corpus of bitexts. The term bitext here refers to paired text and its translation in a second language. The input documents, described in more detail in the following section, are descriptive linguistic books containing text examples of the target language. The figure in (2) below illustrates the goal of this process and the types of bitexts that we would like to produce as output.

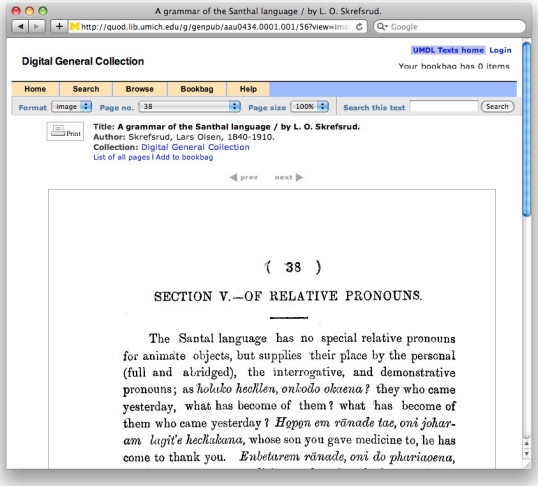
The system processes these documents in two major stages. The first stage identifies instances of foreign text, classifying each word in the document as either belonging to the target (foreign) language or the reference language. (In this project the reference language is always English, although it could be any other language provided that good NLP tools exist for that language.) Then, for each instance of foreign-language text, the second stage identifies an adjacent span of reference-language text that serves as a translation of that text. These two processing stages are described in detail in sections 2 and 3. In practice, this is a challenging process, however, and the actual output of the system contains errors; performance and directions for improvement are discussed in detail in section 4.

1.3 Data Sources

The data sources targeted in this project are descriptive linguistic books, e.g. grammars, dictionaries, and readers, which were originally published in print

[3] http://translate.google.com
[4] http://www.wikipedia.org

(2) The high-level objective of bitext data collection.

Input (scanned digital book)	**Output** (bitext corpus)
	... F-52: *holako hechlen, onkodo okaena?* E-52: they who came yesterday, what has become of them? F-53: *Hopon em ranade tae, oni jo-haram lagit'e hechakana* E-53: whose son you gave medicine to, he has come to thank you F-54: *Enbetarem ranade, oni do phariaoena,* E-54: to whom you gave medicine at that time, he has recovered. ...

form and have since been scanned into digital libraries. These sources were targeted as a potentially large and valuable source of data that is readily available in electronic form, but is not in a machine-readable format. The advantage of extracting the linguistic data from these books is that it could produce data for a large number of languages that previously were unrepresented in digital corpora. The benefit of scale applies mainly to cross-linguistic research; presumably a researcher interested in a specific language could extract the data from a single document by hand relatively easily.

The types of documents targeted are one of the key differences between the present work and ODIN, the Online Database of Interlinear Text (Lewis and Xia 2010). ODIN looks at linguistics articles containing interlinear glossed text (IGT); in such cases the text is relatively easily identified by its distinctive three-line format. The linguistic books targeted in this project may contain IGT, but they also contain instances of foreign text in wordlists, paradigm tables, and inline bitext. Inline bitext occurs when a text and its translation are given sequentially in a running sentence, and cannot be identified simply by looking at the page layout.

A hands-on approach to identifying relevant books was used, manually searching the University of Michigan's Digital General Collection. Queries included searching for the word “language” in the subject field (which matches subject codes like “Thai language dictionaries” or “Czech language Grammar”), and searching terms like “Grammar of” or “Dictionary of” in the title of the book. A list of 110 relevant documents was produced, though not all of these texts were suitable for automated processing: for example, some used non-Roman orthography, which is not recognized by the optical character recognition (OCR) process. Ultimately, a collection of 20 books was chosen for annotation and

additional processing. Basic statistics about this collection are given in (3).

Portions of each document were manually annotated for instances of embedded bitext. Looking at pages that were annotated by more than one person, we calculated an average inter-annotator agreement rate of 0.95 and a kappa value of 0.88. Kappa (Carletta, 1996) is a measure of inter-annotator agreement that takes into account the expected rate of accidental agreement between annotators, and a score of 0.5 or higher is generally considered a good level of agreement. Thus, these results show that there is strong inter-annotator agreement, which is encouraging for the possibility of high-accuracy automated tagging. These annotations were also used for training and evaluating components of the extraction system.

(3) Summary of the scanned linguistics documents used in this project.

Bilingual Texts	11 (Caddoan, Fox, Haida, Kickapoo, Koryak, Kutenai, Maidu, Menomini, Ojibwa, Passamaquoddy, Zuni)
Dictionaries	2 (Burmese, Hungarian)
Grammars	7 (Arapesh, Filipino, Italian, Navaho, Malayan, Pangasinan, Santhal)
Annotated pages	304 (from 9 documents)
Total pages	7,479
Total words	780,000 (estimated)

Most of the results presented in this paper focus on a single representative book, *Grammar of the Santhal Language* (Skrefsrud 1873), which describes Santhali, an Austroasiatic language of about 6 million speakers mostly located in India (Lewis 2009). Several features of this book make it well-suited to this project. Due to its age, this book belongs to the public domain, meaning that the extracted data could be reproduced in a corpus without any concerns of copyright. Also, it is written in English, and the target language is represented in a Latin-based orthography.

2 Language Identification

The first major processing stage is the language identification stage. The objective of language identification is to label each word token in the document as either English or foreign. Linguistics documents are unique in that they are bi- or multi-lingual, combining text from multiple languages in a single document. Outside of texts which are explicitly about language, it is rare to find texts that combine significant amounts of material from multiple languages, and as a result there is fairly little prior research on automatic language identification of individual words within a text. Traditional language ID aims to classify entire documents, not individual words, and does so by comparing the text to samples of known text

from a variety of languages and identifying the sample that best matches the test data. While it is possible to achieve 99% identification accuracy using samples of just a few hundred sentences apiece (Kruengkrai et al. 2006), such approaches still require a sample of text from each language for training.

The creators of the ODIN corpus of interlinear glossed text faced a slightly different variation on the language ID problem (Xia et al. 2009); in their case the IGT instances are already identified within the text, but each IGT needs to be associated with a language. However, this still differs from the present task, in which the documents typically only contain a single target language and the objective is not to identify the language, but to identify the tokens that belong to that language.

Often, target-language text will be distinguished in print by some typographic features, e.g. bold or italic text. While some OCR systems produce output in a markup language (such as HTML or Rich Text) which preserves such typographical information, the OCR used in this project was plain unformatted text. Therefore, the language identification component is tasked with classifying each text token as either an English or a foreign word, based purely on its orthographic form.

2.1 Dictionary and Statistical Methods

One natural approach to the language ID task is a dictionary-based approach, in which tokens are compared to a list of known words in the reference language. One complication is that OCR errors in the English text pose a potential problem since tokens with OCR errors would not be in the dictionary but should be correctly labeled as English. To evaluate the dictionary-based method, we created an English dictionary based off of the *ispell* spell-checking program dictionary, which we augmented with a list of common linguistic terms and abbreviations. This dictionary was used to classify each word in the Santhali grammar: if the token appeared in the dictionary then it was classified as English, otherwise it was classified as foreign.

Another approach is to use a statistical model, for instance one based on n-gram features, to classify word tokens. This approach has the benefit of being tailored to the particular language in question and it is "softer" in the sense that an English word that doesn't appear in training set could still potentially be classified as English. The main drawback of this approach is that it requires a sample of foreign text to train the model on, and in the context of this project we cannot assume that a sample of text from the target language is available beforehand. Still, it is not unreasonable to manually annotate a small number of tokens from the document in order to automatically label the remainder.

To evaluate this approach, we used a 2,620 token subset of the Santhali grammar that had been manually annotated for bitexts. This corresponds to roughly 10 pages of annotated text, and it is a small data set by machine learning standards. Each token was represented as a vector of n-gram features, with *n*

ranging from 1 to 3. The *svmlight* software package (Joachims 1999) was used to train and evaluate a support vector machine (SVM) model. Due to the small data set, we used a hold-one-out methodology for evaluation.

Both the dictionary and SVM models were evaluated on the same data set of manually annotated tokens from the Santhali grammar. The results are shown in (4) below. Here, recall indicates how many true foreign words were correctly predicted as foreign, and accuracy indicates the proportion of predicted foreign words that were true foreign words. Accuracy is the number of tokens (both English and foreign) that were correctly labeled overall.

(4) Comparison of dictionary and statistical language identification results.

	Dictionary	SVM
Precision	66.9	81.7
Recall	76.0	66.0
Accuracy	86.7	88.0

Both systems achieved similar and reasonably high levels of accuracy, with the SVM performing slightly better. However, the two approaches had different characteristic behaviors with respect to precision and recall. The dictionary-based approach predicted more foreign words overall, but with a lower precision: this is likely due at least in part English words that were mis-recognized by OCR.

3 Translation Identification

Once foreign text has been identified in a document, the next step is to identify nearby English text that acts as a translation of the foreign text. In the case of inline bitext, the gloss is either immediately preceding or immediately following the foreign text, but it is unknown which is correct. A statistical translation model could be used to identify the true translation: if the foreign sentence is statistically aligned to two hypothesized translations (one from the preceding text and one from the following text), then the alignment corresponding to the true translation should display a much lower alignment cost than the other alignment.

However, in the absence of a separate corpus of bitext to train the translation model, we are forced to somehow train a translation model without knowing in advance what the bitexts are. A possible solution to this problem is to consider both the preceding and the following text as candidate translations and train a translation model on all of these sentence pairs, even though half of the pairs will be false translations. In order to evaluate the feasibility of this approach, we conducted an experiment on a controlled parallel corpus taken from the Tatoeba database. For this experiment, we collected all of the English-French sentence pairs from the database. To mimic the application setting, each English sentence in the database was also paired with a randomly-chosen French sentence to

produce a false translation for that sentence.[5] The false translations were controlled for length to roughly match the true translations, in order to avoid biasing the results (all else being equal, an alignment with more word tokens will generally have a higher alignment cost).

A statistical translation model was then trained using the combined set of true and false translation pairs. The model was trained using the GIZA++ software package with its default settings (Och and Ney 2003). The alignment scores produced during training were then used to select the better candidate translation for each English sentence. The table in (9) below illustrates the scenario: in each case the *a* translation is the correct one, and accordingly it has a lower cost than the false translation in both instances.

(5) Example alignment costs of true and false translation pairs.

Sentence and Candidate Translations	Cost
'He abused our trust.'	
a) *Il a abusé de notre confiance.*	*18.5*
b) *Il éclata en larmes.*	40.3
'The floor was covered with blood.'	
a) *Le sol était couvert de sang.*	*15.9*
b) *La machine était recouverte de poussière.*	46.7

Because the number of sentence pairs in a single document is generally much less than is usually used to train machine translation systems, we performed this experiment on differently sized subsets of the Tatoeba data set to explore the effect of corpus size (using sets of 500, 5k, and 50k sentence pairs). The translation-selection process was repeated for each corpus under two scenarios: in the first, "gold" scenario, the translation model was trained only on the true translation pairs; in the second, "both" scenario, the translation model was trained on both the true and the false sentence pairs, mimicking the actual case encountered in bitext extraction, where the true translation is not known in advance. Accuracy, defined as the percentage of test sentences for which the true translation received a better alignment score than the false translation, is averaged over five folds of cross-validation, with standard deviation in parentheses. The results are summarized in (6) below:

[5]Note: here French is being treated as the reference language and English the foreign language. This has no significance and the results are expected to hold in the reverse direction as well.

(6) Translation ID accuracy, compared by corpus size and training set.

Corpus size (sentences)	Accuracy (train on gold)	Accuracy (train on both)
500	71.2% (4.7)	72.8% (5.2)
5,000	89.3% (.98)	87.9% (1.3)
53,129	95.4% (.15)	94.4% (.11)

From these results, it is clear that the size of the corpus has a strong effect on the prediction accuracy, which is expected. Also as expected, training on only the true translation pairs yields higher prediction accuracy than training on both the true and false translations. However, this effect is not very large, and for the 500-word corpus any advantage this may have offered is obscured by the noise associated with training on such a small data set.

These experiments show that it is possible to effectively use a translation model that is trained on noisy data to select true glosses from a candidate set containing both true and false glosses. For a small data set, such as might be obtained from a single book, the accuracy rate drops significantly, but is still well above chance. The performance of this technique on a digitized linguistic document is addressed in the following section.

4 Evaluating the System

This section explores the performance of the end-to-end system, taking OCR text from the Santhali grammar as input and producing bitext sentence pairs as output. Word tokens were classified using the same SVM method described in section 2, and each sequence of two or more foreign word tokens (ignoring all non-word tokens, such as punctuation and numbers) was selected as a foreign text. For each foreign text, a preceding and following candidate translation was identified by choosing the appropriate number of tokens to approximately match the length in characters of the foreign text. Finally, these pairs were used as input to the same translation ID system described above in section 3, and the best translation for each foreign text was identified in this way.

This procedure produced 3,503 predicted Santhali bitexts. Nearly none of the predicted bitexts are exactly perfect; even the most accurate are off by a few characters or tokens. Because of this, and because all of the annotated text was used to train the SVM classifier, a random sample of 100 predicted bitexts was chosen for manual inspection. Each of these was assesd on three yes/no questions to determine the quality of the predicted bitext: the questions and results are given in table (7) below.

(7) Santhali bitext extraction evaluation questions.

Question	Yes	No	Pct
Is the predicted foreign text actually foreign text?	99	1	99%
Is this actually an inline bitext?	69	31	69%
If this is an inline bitext, is the prediction approximately correct?	19	50	28%

The first question is meant to assess how well the language ID component performed. 99 out of the 100 bitexts were in fact centered on foreign text, indicating that the precision of the SVM language classifier, when combined with the two-or-more token restriction, is sufficiently high. It is not possible to estimate the recall using this method of evaluation, so 10 pages of the document were randomly selected to inspect. Those pages contained 136 instances of actual bitext, of which 61 were identified by the system, resulting in a recall of 44.9%. The limit of two sequential foreign words for predicting foreign bitexts means that many single-word instances (such as found in inflectional paradigms) were omitted, and this is partially responsible for the low recall.

The second question addresses the fact that not all instances of foreign text have an English translation immediately preceding or following the foreign text. In the sample of 100 predicted bitexts, 69 were in fact inline bitext, meaning that an English translation was present immediately before or after the span of foreign text, and therefore retrievable in principle. In the remaining 31 cases, the present system will always fail to find the translation because it is not immediately adjacent to the foreign text. The third question is a somewhat subjective evaluation of overall correctness. Three examples of the predictions made by this procedure are displayed in (8), along with the responses to the three questions used for evaluation.

Example 1 in (8) shows a three-column table, which are common in the Santhali grammar. This illustrates the need for a method to detect the table structure and deal with it appropriately, since the present system is forced to look only at adjacent text for the translation. Example 2 illustrates a case where the prediction is correct: the full foreign text span was correctly identified as well as the adjacent English translation. Example 3 shows foreign text within a paragraph; the foreign span is cut short (perhaps due the presence of the token “do,” which is a frequent English word), and the translation is misidentified. This may be due to the fact that the actual translation is non-adjacent in this example. These examples are illustrative of the type of texts that are encountered and their associated challenges.

(8) Three examples of predicted bitexts from the Santhali grammar.

1)

had struck him. had struck him. he had struck hitn.
DUAL. DUAL. DUAL.
I D-al-a1,kat'-ti;4-**ta- Dal-akat'-li.-tcth'-** Paset'-e-dat-a~cat'-liti..
lt-1can-a-e, He kan-A-han-e, If tcth~1oan, Perhaps
had struck us he had struck us he had struck us

Foreign? Yes. Inline? No Correct? No

2)

strike.
INCHOATIVE PAST.
Dal-Jko-dagido1l-kan-tahVkan,
Tlhey whom they were about
to strike.
OPTATIVE.

Foreign? Yes. Inline? Yes Correct? Yes

3)

oni hola-m del-led-e, what has become of him whom you
saw yesterday? This is much more elegant and certainly more
correct than to say: **oni hola-m diel-ed-e-a, oni** do okare,
for the latter means literally: you saw him yesterday, what
has become of him?

Foreign? Yes. Inline? No Correct? No

4.1 Directions for Improvement

Clearly, the precision rate of 19 correct out of 100 predicted bitexts leaves something to be desired. Accounting for the fact that in 31 instances it would be impossible to identify the English translation simply by looking at adjacent text, precision increases to 28%, which is still not nearly good enough to be useful for data collection. There are a number of improvements to the system which could not be made in the present study, but which have the potential to yield more favorable results. Some of these are discussed below.

4.1.1 Improving Language ID

If the foreign text spans were detected perfectly, then a simple baseline of always choosing the text to the left or the text to the right would be expected to be correct 50% of the time overall. However, the most common reason for a predicted bitext to be judged incorrect is that the foreign span is too short. If the foreign span is predicted too short, then this will usually throw off the range of the predicted English translation as well. The current language ID system achieves high precision at the cost of low recall; it is essentially too conservative. It is possible that tuning the classifier or training on more data could alleviate this problem.

Another possible solution is to use a sequential model, such as a Hidden Markov Model, to label sequences of foreign words in a soft manner. This should help in cases where an English-looking word appears in the midst of a sequence of foreign words. For example, in Santhali the tokens *an*, *a*, *do* and *than*, among others, could be either English or Santhali, depending on the context, but a token-

based classifier must always label them in the same way regardless of context. (In addition to truly shared words, noisy tokens also pose a challenge.) When such words occur within a Santhali sentence, they incorrectly cause a break in the predicted foreign span. While belonging to an entirely different domain, this is conceptually related to work using HMMs to extract structured information from classified ads (Grenager et al. 2005). Such an approach models a document as being generated from multiple sources, which aligns well with the concept of a bilingual document being generated by two sources (i.e. two languages).

4.1.2 Improving OCR

OCR quality is better today that it ever has been, but OCR errors are a major problem for this type of project. One issue is that the books that we have collected are more prone to OCR errors than typical books. In addition to being old, with faded text and stray marks on the page, the foreign-language text causes problem for OCR that expects English text. Several of the books we originally identified could not be used because they include non-Latin scripts, which are either skipped entirely by the OCR software or produce gibberish output. Even when the foreign-language text uses Latin-based orthography, that text often includes various diacritic marks which lead to errors in the OCR. The figure in (9) below illustrates a typically frustrating example: the grammar presents a paradigm of the Santhali noun *Ṭaṅga* 'axe'.

(9) Comparison of a portion of a scanned page and its OCR output.

Scanned Image	OCR Text
Instr. *Ṭaṅga-te,* by, with, the axe. Dat. *Ṭaṅga-ṭhen,* to the axe. Acc. *Ṭaṅga,* the axe. Abl. *Ṭaṅga-kho̱n, kho̱ĕ̄,* etc., from the axe. Loc. *Ṭaṅga-re,* in, on the axe. Voc. *e̱ Ṭaṅga !* O, axe !	Instr. Tasga-te, by, with, the axe. Dat. Taiga-then, to the axe. Acc. Tagga, the axe. Abl. Tariga-khon, khoci, etc., from the axe. Loc. Tatiga-re, in, on the axe. Voc. e Talga! 0, axe

This example illustrates how the OCR process loses typographic (e.g. italics) and layout information (the spacing and line breaks), but more significantly the letters themselves are misidentified. Although the stem is identical in all six forms of the noun, the OCR software has rendered the same stem in six different ways: Tasga, Taiga, Tagga, Tariga, Tatiga, and Talga. This type of error poses a serious impediment to using the text for further downstream linguistic processing. A morphological analysis based on this data would wrongly posit some strange sort of stem-internal process when in fact there is none.

It is possible that using commercial OCR software could provide improvements. No direct comparisons of quality could be done for this project, but some experiments with a commercial OCR package seemed to improve the

quality of the OCR text. Additionally, commercial OCR software is capable of preserving typographic information and tabular layouts by producing HTML, rather than plain text, output.

4.1.3 Utilizing Typographic and Layout Information

The system described here models the document as a sequence of tokens., which discards much of the typographic and layout information that human readers use to identify foreign text in one of these books. Much of the foreign-language data in linguistic documents is given in a structured format, such as wordlists and paradigms. If this format could be preserved (by using OCR software such as described in the previous section), then it is possible that this information could be used to improve the language ID and translation ID systems. However, the techniques used would need to be modified accordingly to take advantage of this additional markup.

The HMM approach mentioned in section 4.1.1 could incorporate typographic features into its emission probability model. However, while conventions tend to be consistent within a single book, there is not always consistency across books. One author might use italics for foreign text, while another might use it for the reference text. Similarly, in one book, the foreign text might consistently follow the reference text, while in another book the order is reversed. These are parameters that would need to be set (either manually, or inferred automatically) on a per-document basis, but once set should improve the performance of the translation ID system within that document.

5 Conclusion

This paper has presented a system for automatically extracting instances of bitext from scanned linguistic books found online in digital libraries. The performance of the system at present is not sufficient to produce output that could reliably be used to perform linguistic analysis, but there is reason to believe that the performance could be improved with additional work. It is also possible that the output of the current system could be useful in a context where the primary use is to identify interesting instances of bitext which are then manually verified and inspected by the user.

The quality of text produced by OCR is a major issue, even when the remainder of the system works as intended. While the OCR quality may be improved by using different OCR software, it remains unknown whether the quality will reach the level needed to perform reliable linguistic analysis. The OCR issue could be avoided entirely by looking at digitally-composed documents, for instance modern journals and conference proceedings or language-themed web pages.

One alternative to the type of automated process described in this paper is to use a crowd-sourcing approach, using human annotators to identify foreign text

and its translations. If automated processing is not feasible, then this may be a more effective way forward. Ultimately, the types of documents addressed in this project contain a wealth of information of value to researchers in linguistics and computational linguistics, and this value will only be increased if the data can be extracted to a format that facilitates automatic processing.

References

Abney, Steven. 2011. Data-Intensive Experimental Linguistics. *Linguistic Issues in Language Technology* 6.

Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* 22:249–254.

Goldsmith, John. 2007. Towards a new empiricism. *Recherches Linguistiques à Vincennes* 36:9–36.

Grenager, Trond, Dan Klein, and Christopher D. Manning. 2005. Unsupervised Learning of Field Segmentation Models for Information Extraction. In *ACL* 43:371–378.

Joachims T. 1999. Making large-Scale SVM Learning Practical. In B. Schölkopf, C. Burges and A. Smola, eds., *Advances in Kernel Methods - Support Vector Learning*, MIT-Press.

Kruengkrai, Canasai, Prapass Srichaivattana, Virach Sornlertlamvanich, and Hitoshi Isahara. 2006. Language Identification Based on String Kernels. In *Proceedings of the First International Conference on Knowledge, Information and Creativity Support Systems*. Ayutthaya, Thailand.

Lewis, M. Paul, ed. 2009. *Ethnologue: Languages of the World*. Online version: http://www.ethnologue.com/. Dallas, TX: SIL International, Sixteenth Edition.

Lewis, William D. and Fei Xia. 2010. Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages. *Literary and Linguistic Computing* 25(3):303–319.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29:19–51.

Skrefsrud, Lars Olsen. 1873. *Grammar of the Santhal Language*. Benares: Medical Hall Press.

Xia, Fei, William Lewis, and Hoifung Poon. 2009. Language ID in the Context of Harvesting Language Data off the Web. In *EACL* 12:870–878.

Terrence Szymanski

tdszyman@umich.edu