Repurposing Bible Translations for Grammar Sketches*

Paul M. Heider, Adam Hatfield, & Jennifer Wilson
State University of New York at Buffalo
{pmheider,ah63,jlw55}@buffalo.edu

With the number of languages expected to go extinct in the coming century, language documentation as a priority is gaining increasing support. We discuss an experimental method for augmenting the number and scope of available language descriptions. Unlike traditional language descriptions, this work is largely based on translations of Bible verses with the accompanying English text used as a guide to the underlying semantics. Methodologically, our work sits at the intersection of three different approaches to language and linguistics: Classics studies of undeciphered languages, traditional field methods, and corpus linguistics.

We further motivate this methodology and discuss related work in section 1. Section 2 describes some of the language-general challenges posed, both practical and philosophical. Section 3 covers the traditional methodologies from which we extended our work. Section 4 includes short examples from four languages spoken in Papua New Guinea: Folopa, Mufian, Suki, and Urim. In the final section, we sketch several directions for future work.

1. Introduction

With the number of languages expected to go extinct in the coming century, language documentation as a priority has gained increasing support. Unfortunately, the requisite field work to create a thorough language description can be prohibitively expensive (in terms of time and/or money) or temporally impossible (for those languages with no living speakers). We propose a methodology for developing grammatical descriptions based on treating English-language Bible translations[1] as the underlying semantics for a passage in the target language. We do not presume to replace the standard methodologies of careful interviews and elicitation sessions with native

* Many thanks to Matthew Dryer for dreaming up and organizing the project. He has been patient throughout. Other members of the Papuan Language Description Group at SUNY Buffalo provided a wall to bounce ideas off of. Finally, questions asked by various audience members have helped push us towards a clearer understanding of what we do and try.
[1] English verses cited below are from Engelbrite (1999)'s American King James edition.

Studies in the Linguistic Sciences: Illinois Working Papers **2011**: 51-65

speakers. Instead, we hope to augment the tools available to a language documentor.

The work detailed below pulls from our experiences in writing descriptions based on partial Bible translations for four under- or un-documented Papuan languages. Folopa, a Teberan language, has a translation of Genesis (1980) and much of the New Testament (2005). Mufian, an Arapesh language, and Suki, a Gogodala-Suki language, each have a translation of the New Testament (1988; 1981: respectively). Urim, a Torricelli language, has a translation of Matthew (1999) and Paul's Papers (2003).

Before discussing particular lessons learned from working with each language, we will further motivate the methodology as a practical alternative in some circumstances. Second, we will describe some of the language-general challenges posed, both practical and philosophical. Third, we will outline our approach in more detail, including several metaphors used in approaching the data. Finally, we sketch several directions for future work.

1.1. Methodological Motivations

Prolonged personal fieldwork is the undisputed ideal method for collecting data and helping to preserve and document languages. However, many practical issues (e.g., political turmoil, travel expense, work leave) can constrain this option. Relying on ambient linguistic data (i.e., linguistically rich data sources that already exist in an easily accessible format) removes these limitations. Using low-end estimates from Matthew Dryer (personal communication), there are roughly 600 languages with an extended description despite some 5500 total languages in the world. The United Bible Societies lists 2479 language communities as having access to at least some portion of the Bible translated into their native language as of December 2008 (United Bible Society). According to Vision 2025, Wycliffe Bible Translators hopes to have begun a Bible translation for every remaining language community needing one by 2025 (WycliffeVision2025.org). While the number of Bible translation projects is growing, thereby increasing the number of Bible translations available for indigenous languages in the future, not every translation project includes a grammar sketch as a priority. Hopefully, our methodology can help bridge this gap between ambient linguist data (i.e., the Bible translations) and structured linguistic data (i.e., grammatical descriptions).

Two other practical benefits back our methodology: traditional field workers can better prepare for future trips and linguistic analyses are introduced to the research community much sooner than would otherwise be possible. As a precursor to field work, a linguist can use information gleaned from the Bible translation to seed vocabulary lists and discover potentially interesting constructions, perhaps less easily elicited. Data gained in this way can quickly and easily be checked and adjusted in the field. With respect to the faster introduction of data into the research community, the Suki New Testament translation we are using was published in 1981. After almost 30 years, very little explicit linguistic analysis concerning this language has been published. This methodology provides an opportunity to reduce the length and frequency of such knowledge gaps.

1.2. Some Related Work

Related work tends to focus on one of three aspects of our methodology: improving annotations, getting more from corpora, or repurposing ambient linguistic data.

Annotation-Focused Our task has largely been occupied with finding and annotating just enough verses to provide a thorough analysis of the language. Ideally, our annotations would span the entire available text. Fully annotated texts could then serve as yet another representation of the underlying semantics for a new language to be annotated. To that end, active learning, a machine learning algorithm that self-selects training samples, has been adapted to language annotation (Palmer 2009). For a survey of the current active learning literature, see Settles (2009).

Corpus-Focused From information theory and probability theory, statistical measures such as mutual information (for more detail, see Church & Hanks 1990) allow the researcher to directly compare the distributions of two words to determine how much the presence of one word predicts the appearance of the second. Lexicographers commonly make use of these statistics to assist in the identification of word senses by examining which words occur most frequently within a specified window of another word.

Repurposing-Focused Aligning word lists cross-linguistically and with some generalized concepticon can add value to the same lists taken in isolation. As just one example of this practice, Good & Poornima (2010), in association with the LEGO Project (LEGO), leverage a sign-based model of digitized word lists to bridge the gap between these word lists and other lexical resources.

For fields like machine translation, parallel corpora are an essential tool but also prohibitively expensive to create. Some parallel corpora, such as with the European Parliament Proceedings (Koehn 2005) and the Bible, are created for alternate reasons and can be co-opted for research purposes.

2. Language General Issues

Whenever repurposing tools not completely under your domain, it is important to understand the concomitant professional and ethical responsibilities. In a special issue of *Language*, Dobrin (2009: inter alia) discuss many of the subtleties introduced by academic linguistics building upon and taking advantage of missionary resources. Namely, SIL (formerly The Summer Institute of Linguistics), in partnership with Wycliffe International, is the foremost organization undertaking language study and translation work. Combined with the relative lack of fieldwork done by academic linguists in the last several decades, the only published linguistic data available on a given language is often a translation of the Bible.

Without the traditional assurances of native speaker intuition, we have to be more cautious about the reliability of the linguistic data that is generated by this methodology. First, access to phonological levels of interpretation are limited. Second, semantic and syntactic distinctions not encoded in the English are difficult to recover (although, see Beale et al. 2005: for one work-around). Third, some themes, topics, entities, or constructions unacceptable to the Biblical genre and register will go unattested in an analysis. Fourth, wholly new concepts essential to the Bible but novel for the target culture lend themselves to the introduction of calques (see the discussion of 'sheep' in Section 4). Finally (and to be picked up again in a later section), minimal pairs are a matter of accident in Bible verses.

3. Methodological Overview

Pre-Processing The first step was to locate or generate a machine-readable corpus. Here, machine-readable means that the corpus can be treated as a plain text document for all intents and purposes. Our primary needs required easy searching and indexing such that all words were properly segmented from each other and all letters were individually recognized. Unfortunately, all our translations came in either printed form or non-machine-readable digital form.[2] Both formats can be converted into machine-readable text using Optical Character Recognition (OCR), a technique for automatically converting images into plain text.

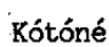
Source Scan (Folopa)		First-Pass Orthography		Target Orthography
	→	DIgNESgSI	→DIÉNESÉSI	
	→ KOt6ne, KOtOne →			Kótóné

Table 1. Two examples of Optical Character Recognition (OCR) throughput from the Folopa Bible

Table 1 documents the progression of two Folopa words ('Genesis' and 'God', respectively) from an example source image in the first column to a first-pass OCR output[3] to the ideal OCR output in the last column. In some cases, the first-pass mistakes were regular enough to be globally fixed. In other cases, manual corrections were necessary. What noise the OCR process introduced was balanced by the volume of easily searchable material even for Folopa, the language with the worst OCR accuracy.

[2] For those interested, the Rosetta Project (`http://rosettaproject.org/`), an archiving project for all the world's languages, has many of the source texts that we used. These versions suffer from the same OCR problems that we ran into when generating the plain text format manually.

[3] One particular hurdle with modern OCR software is that it is optimized for a specific language. The algorithm for deciding what letter a particular section of the image represents can use sophisticated notions of word- and letter-level co-occurrences. This approach greatly improves accuracy for the resultant text given a well-trained system. However, our languages shared neither the vocabulary nor phonotactics of the previously trained languages (e.g., English or French).

Building a Lexicon Assuming a relatively accurate machine-readable corpus[4], we could progress to the second stage: basic lexicon building. Initially, a target word was chosen based on its frequency in short English verses or because it was otherwise easy to isolate (e.g., Arabic numerals). Strings that similarly repeated were then searched for in the source language. Morphological variation present in the target language but not English complicated the matching process. Parallel root morphemes between the languages would not line up if their affixes were not also parallel.

Matthew Dryer provided a program of his own that attempted a similar comparison across an entire corpus. For every word in the source language, Dryer's program would offer a list of the most likely translations of a word based on co-occurrence and string position. Since simple morphological changes could obfuscate repeats of a word and word order is by no means universal, the success of this program varied greatly across languages. At the very least, results from the program provided a bi-directional concordance and good clues for some probable translations to investigate.

Table 2 contains the program's guesses for three Mufian words (*nemata'w*, *alipumi*, and *ilagw*) with an associated score. A higher score represents a guess with stronger evidence. The correct gloss has been bolded. The incorrect guesses are frequently easily attributable to local variation. For instance, the incorrect guesses for the word meaning 'woman' are mostly other terms for females. Likewise, the program guessed both the singular and plural form of 'cloud' in the final column.

	Mufian					
	nemata'w		alipumi		ilagw	
English Guess	**woman**	82	**disciples**	253	arriving	405
	girl	76	his	68	**cloud**	245
	her	75	they	47	overshadowed	245
	poor	61	to	45	clouds	160
	wife	59	and	41	dear	81

Table 2. English gloss guesses for Mufian words based on Matthew Dryer's concordancing software with correct guesses in bold

Because we could not ask a native speaker for a word's connotation or

[4] Because of the highly varied accuracy of the OCR data, not all languages had a fully machine-readable corpus.

denotation, we were forced to use converging evidence of sense distribution to understand what any given lexical item meant. Thus, the lexicon also sometimes lacked words that would be trivial for a traditional field worker (e.g., definite vs. indefinite determiners).

Analyzing Morpho-Syntactic Properties Given a growing lexicon, we could then investigate morpho-syntactic properties of the language. Since the distribution of syntactic constructions in the Bible is largely random, we primarily relied on manual searches to uncover useful pairs. With each researcher working on a different language, we were able to more quickly develop lists of common constructions spread throughout the Bible and lists of stories likely to include rarer constructions. As an example of the former, "X said Z to Y" is a very common construction which can usually be delineated into component parts. As an example of the latter, numerals are likely to appear in the story of the Ark and in the story of Joseph, the dream interpreter. Unfortunately, verse pairs for one language were not always useful pairs for another language.

The above description of our process separates work into clear temporal phases. In reality, work frequently switched from one domain to another: investigating a particular lexical item required positing a particular syntactic role was assigned which required understanding an allomorph of a case marker which was used with this class of lexical item, etc.

4. Methodological Metaphors

Several methodologies used in other areas of linguistic research strongly influenced our working methodology for analyzing the Biblical texts. The first metaphor comes from using meta-linguistic features to align parallel texts for decipherment. In our own work, the Bible is reliably organized around the meta-linguistic feature of verse numbers.[5] Without these, we would not be able to align our target language with its underlying semantics

[5] Unfortunately, translations group verses differently based on different criteria, such as the linguistic or discourse requirements of the target language. For instance, the original verses 2, 3, and 4 could constitute a single sentence in the target language, or may be reordered with respect to each other and therefore the verses would not be differentiated in the target translation. When faced with these many-to-one, one-to-many, or many-to-many verse number mappings, we are unfortunately forced to compare across longer segments of text.

(e.g., the translation in a known language). Three other meta-linguistic features also help align texts: spacing, capitalization, and punctuation. Spaces provide us with an initial and informal understanding of a 'word' in the target language. Regular capitalization provides one more structural designation to help uncover word classes. Finally, commas and periods provide some guidance when matching a passage to the English translation.

As a second metaphor, Jean François Champollion matched proper nouns between Greek and hieroglyphics to decipher the Rosetta Stone. In our Bible work, lining up proper nouns (based on their capitalization, if consistent, or based on their transliteration similarity) afforded insight into possible nominal affixes. These affixes then helped to boot-strap discovery of other nouns and/or to understand argument structure.

Third, words and morphemes not glossed in the prior literature could only be analyzed with respect to their positional and semantic distribution in the corpus. This assumption of distributional definition takes the Oxford English Dictionary's approach to word-definition-through-use to the extreme (for a general description, see Winchester 1998). For example, in Folopa the word *hupu* was regularly used to indicate 'pig', a very common animal throughout Papua New Guinea. However, it was also used in the compound *sipsip hupu* for 'sheep', an animal unknown in the region. Similar compounds are used in other verses to refer to a range of animals. For explanatory reasons, we assume the use of *hupu* is being treated as both the term for 'pig' and as the superordinate term for mammal. It is possible the term is used as a cultural artifact to help the local readership understand what type of beast a sheep is.

5. Language Specific Issues

Although all four of the languages that we worked on are spoken in Papua New Guinea, they are quite distinct from one another. Inhabitants separated by more than a few days' walk ($\approx$10 to 50 km) rarely have contact with each other. Mufian and Urim, spoken in the northwestern part of the country, belong to different branches of the Torricelli family. Because they are separated by 30 km and three other Torricelli languages, contact between the languages would be fairly limited. Folopa and Suki are each spoken several hundred kilometers from this area and from each other, precluding any contact. Folopa, in the Teberan Family, is spoken mainly in Gulf Province, near the south-central part of the country. Suki, in the Gogodala-Suki Family,

is spoken in Western Province, in the southwest. The latter two are both classified as Trans-New Guinea languages, though this family grouping is controversial and is more a geographical grouping than a genetic one; the structures of Folopa and Suki are quite different from each other.

5.1. Folopa: False Expectations from Prior Glosses

Glossed examples used in earlier articles (e.g., Anderson & Wade 1988; Anderson 1989) provided us with a list of approximately 250 morpheme and gloss pairings. We had hoped this seed list would serve the same purposes as collecting a translation of the 200 word Swadesh list (Swadesh 1955).[6] However, most of these morpheme/gloss pairs (e.g., *fidi* 'cassowary') did not occur in the Bible. About 30 pairs (e.g., *=né* 'ergative') could have been determined from the Bible without prior knowledge of their meaning. Of the remaining pairs, knowing a morpheme's gloss was frequently not sufficient to understand the appropriate use and implications of it from the Bible (e.g., *-uą* 'counterfactual'). Sometimes, an English lexical equivalent was used as the gloss of more abstract grammatical morphemes (e.g., *ama* was glossed as 'his' when, grammatically, Folopa does not distinguish masculine and feminine pronouns). Other times, the semantics of the morpheme were difficult to determine without clearly contrasting contexts. As a concrete example, 24 of the 250 morphemes mentioned above relate to verbal morphology. However, tense, mood, and aspect are sparsely covered in our analysis. The prior glosses provided evidence that these verbal modifiers are realized in Folopa as suffixes. However, finding evidence for these linguistic phenomena in the semantics of the translation is quite difficult.

The inability to provide a precise syntactic and semantic analysis of the verbal system did not wholly rule out any verbal morphology work. We aligned glossed examples with common suffixes to test for regular orderings. Present tense markers were all in one column while interrogative markers were in another column, etc. We found a clear ordering that was obeyed across examples. Unsurprisingly, the morphemes could be further clustered by semantic category. That is, present tense and past tense markers were in neighboring columns and thus could be combined into a single tense column. Figure 1 shows the high-level description of the verbal suffix categories revealed by this methodology: applicative markers precede aspect markers

[6] A Folopa Swadesh list does exist. However, the transcription is so significantly different from the rest of the corpus that it was largely not useful.

which precede tense markers, etc.[7]

Verb = Verb Stem + (Applicative) + (Aspect) + (Tense) + (Mood) + (Switch Reference) + (Contrastive)

Figure 1. Folopa verb morphology

In the end, we were able to use these glosses as a strong base on which to build other evidence or as clues for grammatical categories to investigate later. We could not rely on them to be sufficient evidence alone for a claim.

5.2. Mufian: Leveraging Noun Classes

Mufian's large number of noun classes simultaneously facilitated and hindered analysis. Without access to native speakers to question the meaning of words, our initial assumption was that words spelled differently were likely to have different lexical meanings, assuming those differences could not be attributed to morphological affixation. To better illustrate this, Example 1 shows the forms of the numeral 'two' in the book of Mark. Since many modifiers have a different form for each different noun class, it was not until the discovery of noun classes that we were able to correctly categorize these as instances of the same word.

(1) 'two' *biam, biagof, biagw, bias, biasa, biafin, biafina*

Although noun class agreement initially inhibited progress, it ultimately facilitated the analysis. Because noun class agreement was marked obligatorily and yet differently on many different parts of speech, we could use the morphology to determine the part of speech of an unknown word. Knowing the parts of speech for all the words in a sentence allowed us to parse sentences despite unknown vocabulary items.

As shown by the gloss in Example 2, verbs take a prefix which codes the noun class of their subject, while nominal modifiers mark agreement with the noun they modify via suffixation.

[7] While all forms except the root are listed as optional, there is most likely some interaction between which types can co-occur.

(2) Anen n-anda' waf mai-f awafi?
3S.NC4 NC4-do thing.NC8 what-NC8 bad.NC8
[Then Pilate said to them, Why,] what evil has he done? [And they cried out the more exceedingly, Crucify him.] (Mufian, Mark 15:14)

Furthermore, nouns frequently end in the same consonant used to mark agreement with them. Therefore, noun class agreement allowed verb identification based on its prefix which then helped determine the subject. It also clarified relations between adjectives, numerals, possessive pronouns and the noun being modified. In this specific example, the noun class agreement prefix *n-* helps to parse *nanda'* as a verb, while while the /f/ of *maif* and *awafi* suggests a relationship to the noun *waf*. When combined with a basic understanding of word order gleaned through examining short simple sentences from the corpus, noun class agreement morphology often clarified grammatical relations between words and allowed for further segmentation of the sentence.

5.3. Suki: Finding Pronouns

Though there were substantial difficulties in analyzing the large and complex system of morphological marking in Suki, the pronominal system was more tractable because pronouns are obligatory. English's similarly obligatory pronominal system served as a good guide.

Unfortunately, the English translation did not always match the pronoun used in the Suki. At the initial stages of analysis, the mismatch proved to be a rather large hurdle. As the data accumulated, the system became clear.

There are several possible complications which can arise when dealing with pronouns. One complication relates to the distinctions made by the pronominal system. Luckily, Suki makes the same distinctions that English does (see Table 3), though this will frequently not be the case. Suki pronouns posed a different kind of challenge.

The 2nd person singular and 1st person plural pronouns in Suki turned out to be homophonous (see Table 3). Once these base forms had been identified, it made further grammatical analysis much easier.

	Singular	Plural
1st person	ne	e
2nd person	e	de
3rd person	u	i

Table 3. Suki pronominal system

5.4. Urim: Mismatching Number Distinction

Since the researcher relies on overt distributional evidence to determine the meaning of unknown morphemes, any grammatical features encoded in the target language with no obligatory expression in English are inherently more difficult to uncover. A relatively simple example of this can be seen in one of the more interesting features of Urim: a four-way number distinction. While English distinguishes between singular and plural, Urim makes two further distinctions: dual and paucal. The second person pronoun provides an even more extreme disjoint between the English, with one category, and Urim, with four categories (see Table 4).

	2SG	2DU	2PAU	2PL
Urim	kitn	kipmekg	kipmteng	kipm
English	you	you	you	you

Table 4. The second person pronominal systems of Urim and English

Example 3 illustrates a common disjoint between the English verse and the Urim translation. Although the English translation simply uses the plural noun 'men', the Urim equivalent involves the third person paucal pronoun *tunteng* '3PAU'. While there is nothing in the English verse to specify that there are only three wise men, the Urim equivalent specifies that there are more than two men but less than approximately six. It is not until more verses are collected demonstrating dual, paucal, and plural pronouns that this difference in number distinctions can be seen.

(3) tunteng melnum ariwe
3PAU men wise
[Now when Jesus was born in Bethlehem of Judaea in the days of Herod the king, behold, there came] wise men [from the east to Jerusalem,] (Urim, Matthew 2:1)

In the end, it is sometimes necessary to look beyond the exact verse to create an analysis. The proper scope could be as simple as the surrounding verses or as complex as the situational context.

6. Comparative Results

Matthew Dryer, who supervised the writing of many grammatical sketches using our methodology, noted that those languages with nominal or word-order oriented syntax were better and more thoroughly analyzed than those with more verbal syntax (personal communication). As discussed above, many of the traditional verbal markers (e.g., mood and aspect) are very difficult to interpret via the parallel English verse. In contrast, nominal markers (e.g., case and noun class) are easier to compare across languages. Part of that bias may be tied to the underlying English semantic representation. Using a language with more verbal morphology for the underlying semantics could have made uncovering verbal regularities easier.

A second generalization not to be missed is the different scope of these constructions. The easier constructions to analyze also could be described as phrase-level (e.g., relative ordering of head noun and modifiers). In contrast, the harder constructions required clause-level understanding (e.g., relative clauses and subordination) As we acquired a larger productive vocabulary, we were able to understand and process larger constructions. Perhaps the nominal and configurational analyses could be done with a smaller window of understanding than the verbal analyses.

Comparing our analyses with traditionally generated sketches yielded favorable results. In general, our methodology resulted in a simpler analysis due to some missing or unobserved categories. For instance, Folopa's basic ergative system was clear from the Bible. However, Folopa's split intransitivity system (Anderson & Wade 1988), where ergativity is also dependent on intentionality and control, was not recoverable from the Bible. Likewise, Mufian's noun class system contains more classes than those recovered using our methodology A one-page overview of Suki is available, including a template for verbal morphology. All of the issues reported matched what was recoverable from the Bible, with the exception of the verbal system. Even after using the templatic structure to reanalyze Suki verbs, the verbal system was still far too complex to resolve based on the translation alone. Otherwise, our Bible-based description matched the other available description.

7. Future Work

We currently have three major divisions for future projects: developing a language-general back-end, introducing pointwise mutual information into the lexical analysis, and augmenting the bi-directional concordance program. A language-general back-end could help us to better leverage work done on one language in other languages. For instance, indexing rare construction environments in one language could help locate similar constructions in another language. Also, typologically and/or geographically related languages could benefit from more strongly comparative studies. Next, pointwise mutual information, and related measures, could help build up the lexicon more quickly which, in turn, allows morpho-syntactic analyses earlier. Finally, the bi-directional concordance program is still in its early stages of development. While currently each gloss decision is made independently of other decisions, and multi-word phrases are not accounted for by the comparison algorithm, these adaptations could further improve its effectiveness.

In sum, we have presented a new methodology built upon the intersection of several fields for repurposing ambient linguistic data. It can be used both as a preparatory tool for traditional field workers and as a primary tool for linguists unable to venture into the field. In either case, our end goal is a functional, if not fully complete, grammar for the source language that can be used for typological and/or comparative studies.

REFERENCES

Anderson, Neil. 1989. Folopa existential verbs. In Karl Franklin (ed.), *Studies in componential analysis*, vol. 36, 83–105. Ukarumpa: Summer Institute of Linguistics.

Anderson, Neil & Martha Wade. 1988. Ergativity and control in Folopa. *Language and Linguistics in Melanesia* 19. 1–16.

Beale, Stephen, Sergei Nirenburg, Marjorie McShane & Tod Allman. 2005. Document authoring the Bible for minority language translation. In *Proceedings of MT summit*. Phuket, Thailand.

Church, Kenneth & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29.

Dobrin, Lise M. 2009. SIL International and the disciplinary culture of

linguistics. *Language* 85(3). 618–619.

Engelbrite, Stone (ed.). 1999. *American King James Bible*. URL `http://www.crosswire.org/sword/`. Version 1.4.

Folopa Genesis. 1980. *Diénesési (Genesis)*. Ukarumpa, E. H. P., Papua New Guinea: S.I.L. Printing Department.

Folopa New Testament. 2005. *Yesu kerisoné so whị tao sere kisi fo wisi*. Wycliffe Bible Translators.

Good, Jeff & Shakthi Poornima. 2010. Modeling wordlists: Conceptual and implementational considerations. Poster presented at the 84th Annual Meeting of the Linguistic Society of America.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*.

LEGO: Lexicon Enhancement via the GOLD Ontology. Accessed: Aug. 2010. Website. URL `http://linguistics-ontology.org/project/8`.

Mufian New Testament. 1988. *Basef bu'wafi Godi (God's good talk)*. South Holland, Ill. U.S.A.: World Home Bible League.

Palmer, Alexis. 2009. *Semi-automated annotation and active learning for language documentation*. Ph.D. thesis, University of Texas at Austin.

Settles, Burr. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Suki New Testament. 1981. *The New Testament in Suki/Godte gi amkari titrum ine*. Port Moresby: The Bible Society of Papua New Guinea.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21(2). 121–137.

United Bible Society. Accessed: Aug. 2010. United Bible Society: Scripture translation. Website. URL `http://www.biblesociety.org/index.php?id=22`.

Urim Matthew. 1999. *Matyu nira yangkipm wor a la sisas krais*. Finland: Finnish Evangelical Lutheran Mission.

Urim Paul's Papers. 2003. *Wrkapm a Pol nira eng tu wrong kin kipman a Maur Wailen (Paul's papers that he wrote to the people of God)*. Finland: Hip Pocket Project - Wycliffe Bible Translators Europe and the Finnish Evangelical Lutheran Mission.

Winchester, Simon. 1998. *The professor and the madman: A tale of murder, insanity, and the making of the Oxford English Dictionary*. New York: HarperCollins Publishers.

WycliffeVision2025.org. Accessed: Aug. 2010. When we'll finish: Vision 2025. Website. URL `http://www.wycliffevision2025.org/Explore/WhenWillWeFinishtheTask.aspx`.