Guarani: A case study in resource development for quick ramp-up MT

Ahmed Abdelali, James Cowie, Steve Helmreich, Wanying Jin,
Maria Pilar Milagros, Bill Ogden, Hamid Mansouri Rad, and Ron Zacharski
New Mexico State University
Las Cruces, New Mexico 88003
{ahmed, jcowie, shelmrei, wanying, mmilagro, ogden, hamid, raz}@crl.nmsu.edu

Abstract

In this paper we describe a set of processes for the acquisition of resources for quick ramp-up machine translation (MT) from any language lacking significant machine tractable resources into English, using the Paraguayan indigenous language Guarani as well as Amharic and Chechen, as examples. Our task is to develop a 250,000 monolingual corpus, a 250,000 bilingual parallel corpus, and smaller corpora tagged with part of speech, named entity, and morphological annotations.

1 Introduction

In this paper we describe a set of processes for the acquisition of resources for quick ramp-up machine translation (MT) from any language lacking significant machine tractable resources into English. In previous work (Nirenburg et al. 1998) we developed an elicitation system that guides non-expert language informants through questions about the ecology, inflectional morphology, and syntax of their language and leads them through a lexicon development task. This information was then used to automatically generate a transfer MT system. Our current approach replaces this sequential, guided process with the more free-form acquisition of general resources, which could be used by experts to create an MT system. These resources include a 250,000 parallel text corpus, and a 250,000 word monolingual corpus, as well as smaller corpora tagged for part of speech, named entities, and morphological analysis. The collection methodology varies with the amount of web/electronic resources that can be developed, the availability and location of linguistic experts and native speakers and the political relationships with the countries where the languages are spoken. We illustrate our methodology by describing our acquisition efforts for Amharic, Chechen, and Guarani. For example, one challenge of the Guarani effort is that due to the lack of Guarani/English bilingual speakers it was necessary to use Spanish as an intermediate or "bridge" language. For Chechen, on the other hand, political ramifications had to be overcome.

2 Background

The work described here is situated in the need for quick ramp-up machine translation from languages with few machine-tractable resources (online monolingual corpora, bilingual corpora, lexicons, and analyzers) into English. Development of machine translation systems requires such resources and one could arguably make the case that the quality of the resulting MT system is greatly dependent on the quality and quantity of this language data. For example, the quality of MT systems developed by statistical methods is dependent on the size and quality of the bilingual parallel corpus. This belief is supported by recent experimental work by Banko and Brill (2001). In contrast to earlier studies (for example, Ratnaparkhi 1999 and Henderson 1999) Banko and Brill found that for statistical natural language processing it makes more sense to allocate resources to increase the size of the training corpus than it does to allocate

Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 1-9,
Cambridge, August 2006.

those resources to exploring and improving different learning methods.

The size and quality of linguistic resources also affects the quality of rule-based systems, which depend, among other factors, on access to clear information about the morphological paradigms of the language. Even translation systems that focus on the generation of text, such as Generation Heavy Machine Translation (GHMT) (see Habash 2002, 2003) require some source language resources. Thus, the development of clean machine-tractable resources of sufficient quantity is a major bottleneck in any MT effort involving low-density languages.

In Nirenburg et al. 1998 we describe an approach to this acquisition task that relied on constructing an elaborate, complex web-based system, Boas, that guided a linguistically-naive language informant through the process of acquiring descriptive knowledge about the parameter inventory for a particular language. For example, through a set of guided examples, morphological parameters such as number, gender, and case would be elicited from the informant. This was followed by elicitation of paradigms that instantiated those parameters. Once this acquisition phase was complete, an MT system could be automatically generated from the acquired parameters. While this yielded an MT system of a quality on par with other quick ramp-up approaches, it required dedicated, motivated acquirers who were willing to devote the hours required to complete the acquisition task. One benefit of this approach beyond developing MT systems is that the acquirers gained an understanding about the linguistic facts about their language.

In our work on Amharic, Chechen, and Guarani, we use an alternative approach. Instead of guiding acquirers through an elicitation process designed to gather knowledge about parameters, acquirers are used to construct basic resources for a language including:

- a monolingual text corpus of at least 250,000 words
- a parallel bilingual (with English) text corpus of at least 250,000 words
- a bilingual lexicon of at least 10,000 headwords or lemmas
- a small manually-annotated part of speech tagged corpus
- a small manually-annotated named entity tagged corpus
- a morphological analyzer

While the approach described here has very different acquisition objectives compared to our previous approach, we believe that it will lead to a quick ramp-up MT system of comparable quality.

3 Design philosophy of acquisition tools

Boas, as mentioned above, was a web-based application and this has a number of direct benefits for language preservation projects. For example, acquirers can use the system from the nearest browser. It enabled developers to fix defects promptly without having to distribute updates, which then users would need to install. Finally it facilitated the central storage of linguistic information. However, a significant downside is that it required acquirers to have adequate connection to the Internet and this could be a considerable obstacle for some acquirers living in remote areas. In addition, the modules of Boas were internally complex. Our design goal was to develop a system that could be used for any of the world's languages and this added to the complexity. In order to attain good coverage of language variation requires large parameter/value sets and as a result the work of an acquirer becomes more difficult as the set of parameters, values, and realization options grow. Moreover, users are exposed to a large number of questions that are not relevant to their particular language. As an example of this complexity, the lexical acquisition module needed to encode inherent features, account for irregular morphological forms and also account for the fact that some aspects of language realized lexically in one language might be realized as affixes or morphological features in another language. The development of the module that gently guided a language informant through the acquisition of this information required considerable effort on the part of linguists and developers (see, for example, McShane et al. 2002).

These experiences led us to several design guidelines.

1. **PC based application**. We opted to replace the "run from the nearest browser" approach used in the Boas system with a

mobile solution that enabled the acquirer to be nearly untethered from the web. This aspect was particularly important for our Guarani acquirers as high bandwidth internet connections are not widespread in Paraguay. This also allowed the system to be taken into the field. At times convenient to the acquirers, they can connect to our web-server and upload the resources they have collected to a centralized store.

2. **Favor applications and interfaces known to the acquirers**. Our previous web-based approach enabled acquirers to use the operating system with which they were most familiar to access the Boas application. Replicating this versatility is a challenge for a PC-based approach and we had several options. We could port our applications to run on the major operating systems (Linux, Macintosh OSX, Microsoft Windows). However, based on our experiences on other projects, maintaining several versions of an application and managing revisions that fix defects across platforms requires a substantial development effort and we were unwilling to invest the required resources. Another solution we considered was to use a Linux Live CD approach—one bootable CD that contains all the required language application software. This approach has the advantage of eliminating installation problems and also would enable us to have a consistent interface. We felt that this was a viable option since modern Linux Live CD distributions (for example, Knoppix) are relatively easy for naive users to use. However, there is the possibility that some unpredicted difference between Linux and the current operating system the acquirer was using would make the acquisition process difficult for the acquirer. Thus, from the acquirer's perspective, there is a high preference for tools to run on the operating system they would be most familiar with. In the case of our Guarani acquirers this was Windows 98 and our Chechen acquirers were using Windows XP SP2. Our design criterion is that all our applications will run on Windows 98 as well as Windows XP.
3. **Keep things simple**. The Boas system embodied a substantial amount of linguistic information in its acquisition tools (McShane et al. 2002). For example, the lexicon acquisition tool had knowledge of inherent features and irregular inflectional paradigms. The morphology component had knowledge of the standard morphological features and their allowable values in sufficient detail to handle the majority of the world's languages. For example, the case system had 29 case values in its realization set. A major disadvantage of this approach is the resources required for its development (the Boas system as a whole required over 12 person years to develop). While McShane et al. note this approach has a wide range of advantages, we focused on a different approach in our current project of acquiring resources for Amharic, Chechen, and Guarani. We decided to keep the tools simple and spend more effort (and more of our funding) on actual acquisition of linguistic resources. For example, in following this design principle as well as principle number 2 above it was decided to use a standard word processor as the tool used by the acquirers to enter parallel bilingual text as opposed to a specialized tool that would save the results in an XML file. In Boas, the lexicon acquisition tool was a specialized web-based application with an underlying relational database, which was completely rewritten three times during the course of the project. Our current acquisition effort makes use of a standard spreadsheet template with a handful of macros. The named-entity tagger we use is an extremely simple Python/Qt application from the Linguistic Data Consortium. Another advantage of using simple applications is that they tend to perform better than most full-featured applications on slower machines common among acquirers.
4. **Preference to open source solutions**. In an effort to have the acquisition effort continue and thrive after our initial funding expires—to have local language communities take over the effort, we prefer to use and develop tools and resources that are

open source. All the tools and resources that we develop in-house are under a Creative Commons Attribution Non-Commercial license which allows others to tweak and build upon our work. This principle also led us to use the OpenOffice.org office suite for the word processing and spreadsheet work mentioned in (3) over commercial alternatives.

4 Web-based repository

In section 2 we stated that the acquisition effort was conducted on standalone PCs, not requiring connection to the web. Once language informants using these tools have acquired linguistic resources, they are uploaded to a web-based repository commonly shared by acquirers and our customers. It is during the upload phase that any character set conversions are done. For example, our Guarani acquirers prefer working in Times Guarani. When they are collecting and producing resources, the resources are in the visual encoding matching the Times Guarani font. During the upload phase they are converted to UTF8, the standard character set of our repository. While any visitor to the site (http:crl.nmsu.edu/say) can view and download the resources, only authorized acquirers are allowed to upload materials. The system maintains a database entry for each file uploaded. These database entries includes information such as

- URL/Location of file
- Type of resource (for example, monolingual text, parallel bilingual, named-entity tagged)
- Short description
- Acquirer ID
- Date
- Length (number of words)
- Language
- Status (locked, complete, for example)
- comments

While a specific acquirer is associated with each resource, acquirers can elect to remain anonymous to other users of the system. In this case, users see only that a particular resource was collected by an anonymous acquirer.

4.1 Uploading resources

As mentioned above, during the acquisition task, language informants use simple standalone PC tools—often just a standard word processor. For example, parallel bilingual text is entered using a word processor and the source language sentence and the English equivalent are delimited by blank lines as shown below for Chechen

Кху деношкахь мелла а цхьалхадаьлла цигахь карзахе лаьттина хьал.

The other day, the alarming situation there has somewhat defused.

Коьртаниг –хиллачу зуламна оьгIаздахана, Нохчийчоьнан дозанал дехьадаьлла адам шайн цIентIе духадирзина.

The most important thing is that the people who were angry with the committed crime and left Chechnya have returned to their homes.

Иштта, цигара вухавирзина, хIинца шен балха араваьлла В.Гарсаев а.

V. Garsaev has returned from there and returned to his work.

Because this convention is not enforced in the word processor itself, during the upload process the system displays its interpretation of the file to the acquirer asking the acquirer if the file looks as expected. For example, in this Chechen case the Chechen/English pairs are displayed in a way that highlights how the system views the pairing (on the web, the English translations are displayed in blue).

1	Кху деношкахь мелла а цхьалхадаьлла цигахь карзахе лаьттина хьал.
	The other day, the alarming situation there has somewhat defused.

2	Коьртаниг –хиллачу зуламна оьгIаздахана, Нохчийчоьнан дозанал дехьадаьлла

	адам шайн цIентIе духадирзина.
	The most important thing is that the people who were angry with the committed crime and left Chechnya have returned to their homes.

3	Иштта, цигара вухавирзина, хIинца шен балха араваьлла В.Гарсаев а.
	V. Garsaev has returned from there and returned to his work.

In this way, the user can check whether the Chechen and English are aligned properly, and whether the codeset is correct. If the alignment does not look okay as in the following:

1	«Цхьаъ Бу Вайн Мохк. Юкъара Ду Дайн Кешнаш А. Ницкъ Тоьар Бац Зуламийн,
	ЭгIош Вай Вовшашна Дуьхь-Дуьхьал ХIитто»

2	“We Have One Homeland. Graves of the Ancestors Are Common. The Evil Will Not Have Enough Strength To Fool Us And Make Us Oppose Each Other”.
	Шелковски районерчу Бороздиновская станицехь бохам иккхинчу шолгIачу дийнахь дуьйна, цига командировке вахийтина, вайн Правительствон цIарах цигарчу нахаца маслаIатан болх беш вара Нохчийн Республикан къоман политикан, зорбанан, хаамийн министран заместитель Гарсаев Ваха.

the acquirer can abandon the upload, fix the file on his/her local machine and then upload this revised version.

Similarly, for the uploading of named entity tagged text, the text and tagging are displayed.

5 Geographically dispersed teams

We provide support for three different models of acquisition, depending on the language involved and thus the availability of human resources for acquisition. All models involve having a line supervisor in place in house for the language. This person need not know the language involved, but must be able to converse with the acquirers, wherever and whoever they may be. It is also expected that this person spend some time familiarizing themselves with the language, learning to speak it if at all possible. This person has overall responsibility for seeing that the acquisition takes place in a timely manner and is of acceptable quality. The supervisor also should have some computational linguistics background in order to assist in the construction, use, and evaluation of the morphological analyzer.

Beyond this supervisory level, the models differ, though they are not mutually exclusive. Instead, they function as prototypical modes of acquisition. Each language presents its own problems. The first model involves complete in-house acquisition. If we have or can hire local speakers of the language, and someone with linguistic expertise in the language (to assist with development of the lexicon, the morphological analyzer and the reference grammar) then acquisition can be done in house. However, given the nature of the languages, however, and our location, we have not yet been able to make use of this model. All of our acquisition has involved the second or third model. Currently, though, we are looking at additional languages and it seems possible that Uyghur may be the first test for this model of acquisition.

The second model involves locating acquirers and experts within the United States, and then using them as consultants or subcontractors. Often the acquirers are immigrants from the countries where the language is spoken. Experts tend to be located in various universities and research centers in the US. For instance, with Guarani, we were able to locate a professor who was a native speaker of Guarani and who had an advanced degree in Guarani and, in fact, whose dissertation was about certain syntactic aspects of the language. She will serve as our quality control monitor, our advisor on linguistic aspects of the project (development of parts of speech, morphological features and

classes, etc.), and the major contributor to the reference grammar.

This is the model for Chechen as well. This model can pose some difficulties in management since experts and acquirers may not see eye-to-eye about the language itself, and more likely, about the analysis of certain features. For example, one person we consult with for Chechen is a theoretical linguist with known expertise in Chechen. One of our experienced Chechen acquirers has considerable expertise as a professional translator. However, these individuals have wildly different views about Chechen morphology and it is a considerable challenge for the computational linguists here to consolidate these views. For our Guarani effort we have acquirers in Paraguay, and a theoretical linguist with Guarani expertise in the southwest of the United States. Part of our task then is to see these individuals as viewing the same language facts through different lenses that may be distorting the image somewhat. For the purposes of our project, some of these potential differences may not be relevant. Our task is simply to produce the resources required for producing an MT system and factors that may be theoretically elegant may serve only to distract the team.

In some cases, where there is good confidence in the abilities of an acquirer and where the acquirer is not simply a speaker of the language, but one who has studied the language as well, acquirers can serve as quality controls for each other. This is how the acquisition is primarily being accomplished for Amharic and Chechen.

The third model is the fallback model when no appropriate speakers or acquirers are available within the United States. In such cases, few, if any of potential acquirers are speakers of English. This poses some of the difficulties addressed in the next section, on bridge languages.

6 Bridge language

When acquirers and experts must be obtained outside the US (and even occasionally within the US), acquisition follows the same pattern as the second model, though it is, of course, more difficult to develop consultancies and subcontracts with people and entities outside the US. In the case of Guarani, for example, we are working with Idelguap, the Instituto de la Lingüística Guaraní del Paraguay (the language is spelled with a final accent in Spanish, but not in Guarani itself, nor in English). First we needed to negotiate a memorandum of understanding between our university and theirs, and only then could we negotiate a subcontract with the institute to provide the acquirers needed.

The major difference between Guarani and the other two languages is that the speakers of Guarani are primarily bilingual in Spanish, and very, very few are bilingual in English and Guarani and even fewer trilingual in English, Spanish, and Guarani. In cases such as this, we use Spanish as a "bridge language" both for within-project communication, and also for acquisition. For example, part of our parallel corpora contains a set of documents that were originally in English. These were translated first into Spanish by a native Spanish speaker and then translated from Spanish to Guarani by a native Guarani speaker. An example of such a document is shown here:

> A verdict handed down three years ago today brought debate from both sides of the Atlantic.
>
> Un veredicto dictado hace tres años provocó un debate hoy a ambos lados del Atlántico.
>
> Peteĩ ñe'ẽmondo ojekuaaukava'ekue ojapo mbohapy ary oporombojovake ko árape mokõive Atlántico mboypýri.
>
> In this edition of "Headliners," Bob Glascoff has the story of British au pair Louise Woodward and finds out what she's doing now.
>
> En esta edición de "Headliners," Bob Glascoff tiene la historia de la chica au pair británica, Louise Woodward, y se informa acerca de lo que está haciendo ahora.
>
> Ko'ág̃a osẽva "Headliners"pe Bob Glascoff oguereko tembiasa mitãkuña Británica au pair Louise Wordward rehegua ha oñeha'ã oikuaa mba'épa ko'ág̃a rupi ojapo.
>
> October 1997.
>
> Octubre 1997.
>
> Jasypa 1997-pe

Louise Woodward is sentenced to life in prison on charges she murdered an infant boy.

Louise Woodward es sentenciada a cadena perpetua en prisión por el cargo de haber asesinado a una criatura.

Louise Wodward oñesentensia itasã hi'arapa'ÿvape ka'irãime, ojukahague rehe peteĩ mitã.

The same Spanish-bridge approach is used in developing a lexicon. Based on word frequency lists, the Guarani acquirer adds base forms to the lexicon, providing part-of-speech and inherent features as well as a Spanish translation. This Spanish translation is then translated into English by our in-house Spanish translators. Quality control is done by a bilingual Guarani/English speaker who performs spot checks.

We expect that this bridge language method will be an increasingly common way of acquiring resources as the languages for which resources are being acquired become less and less well-known and with fewer and fewer speakers. However, the "bridge language" approach will be viable for languages spoken in large parts of the world. Arabic, for instance, can serve as a bridge language throughout the Middle East and North Africa, as well as in many countries with a large Muslim population. French, Spanish, and Portuguese can be used as bridge languages in their former colonial empires. Chinese can be used throughout a large area of south-east Asia and Russian in the territories of the former Soviet Union.

To deal with the problem of within-project communication, we insist that the in-house supervisor speak the bridge language and, in fact, be bilingual in English and the bridge language. In the case of Guarani, it was not difficult to find a bilingual English-Spanish speaker. Conference calls with the acquirers in Paraguay are conducted in Spanish, and much of the written correspondence (email) is also in Spanish. There is often as well, a good bit of printed or on-line material in the bridge language about the language to be acquired. Thus, we were able to locate Spanish grammars of Guarani and Guarani-Spanish bilingual dictionaries. In the case of Guarani, there is, in addition, a large influence of the bridge language on Guarani vocabulary as well. In such cases, knowing the bridge language is a definite advantage in understanding the language to be acquired. Guarani varies from isolated tribal dialects, though a kind of standard Guarani that is understood by all speakers but with limited lexical roots to an urban variety (jopara) with a large mixture of more-or-less nativized Spanish vocabulary and, in some cases, grammar as well. For instance, Guarani has no definite article, so the Spanish articles la and lo have been adopted in many cases. Guarani lacks gender, however, and number is only indicated occasionally on nouns through an optional suffix, so la now is often used to mark singular nouns, while lo marks plural nouns. This also raises questions about what variety of language resources to collect.

The bridge language is also used in resource collection. Thus, the parallel corpus is translated first from Guarani to Spanish and then to English, or (in the case of the English original parallel corpus) from English to Spanish to Guarani. This additional translation step poses great dangers for the quality of the resultant translation. Our Chechen acquirer spent many months correcting what had been originally a Chechen-Russian dictionary which had been turned into a Chechen-English dictionary simply by translating the Russian side of the dictionary into English.

Another area of difficulty lies in translating idiomatic expressions or non-compositional expressions out of English into Spanish, where there is no exact equivalent. For example, one English article talked about trying to bring North Korea "in out of the cold". That phrase, with its connections to the world espionage and John Le Carre, meant little or nothing if translated literally into Spanish. However, translating the sense would result in a communicative Spanish equivalent, but might skew the translation into Guarani in an unexpected direction.

With such bridge languages, then, it is vital that the quality control person and/or expert be trilingual, in English, the language to be acquired, and the bridge language. This allows the expert not only to judge the quality of the source language-English translation, but also to understand how mistakes arose through the bridge-language translation and thus to suggest structural ways to improve the translations.

If these precautions are taken, we believe that the bridge-language approach is likely to prove a fruitful method of language resource acquisition.

7 Doing morphology

In previous sections we described the tools that are used to acquire monolingual text, parallel bilingual text, and lexical entries. In this section we describe the more complex task of constructing a morphological analyzer. In creating morphological analyzers for machine translation, several development strategies are available. The analyzer can be developed by hand coding finite state rules. This approach was used, for example, by Pretorius and Bosch for Zulu (2002), Beesley for Arabic (1996), and Maxwell (2003). Alternatively, the analyzer can be developed by eliciting paradigm templates from language experts and using supervised learning techniques on these templates creating finite state morphographemic rewrite rules. This approach was used by Kemal, Nirenburg, and McShane (2001) and Zajac (2001). (Both these approaches were used in the Boas system.) In our current project, we have been using a hybrid approach that combines these two methods along with unsupervised learning. The approach follows an iterative elicit-build-test methodology. Initially, paradigm templates are developed using information from any reference grammars that may exist, and information from our language experts. Next, an initial system is built using supervised learning techniques applied to paradigm templates. This supervised approach works as follows. First, using the citation form and paradigm, we determine the best stem. We compute this using a minimum edit distance approach (Wagner and Fischer 1974). In particular, we use an approach suggested by Oflazer, McShane, and Nirenburg (2001). We examine each potential stem of a citation form. That is, for the citation form *study* we examine the citation forms *s, st, stu, stud,* and *study*. For each citation form, cf, we compute a score as follows:

score(cf) = length of cf + SUM(ed(cf, each inflected form))

where ed is the edit distance. We then select the stem with the minimum score. Once the stem is determined we develop a set of morphographemic rules using a similar minimum edit distance approach. These rules, compatible with the Xerox Finite State Toolkit, can be refined by hand as needed. Finally, the system is evaluated by comparing it to a morphological rule set developed by unsupervised learning over a monolingual corpus. The unsupervised learning uses the Linguistica processor developed by John Goldsmith (2001). Using these evaluation results and proposed rules, the developer modifies the paradigm templates and rule sets. The application of this technique is the development of morphological analyzers for low-density languages such as Amharic, Chechen, and Guarani.

8 Summary

In this paper we have described our method for collecting language resources for quick ramp-up MT. Our design principles were 1) to develop PC-based language resource collection applications rather than web-based ones; 2) to favor applications and interfaces known to the acquirers; 3) keep the tools developed simple; and 4) prefer open-source solutions. Because of these principles (particularly the 'keep tools simple' principle), more of our time and resources were spent in the actual acquisition task and less on tool development. Approximately 30% of our budget went to first-line acquirers and only a small percent went to tool development.

References

Banko, Michele, and Eric Brill. 2001. Mitigating the paucity of data problem: exploring the effect of training corpus size on classifier performance for natural language processing. Proceedings of the Conference on Human Language Technology.

Beesley, Ken. 1996. Arabic finite-state morphological analysis and generation. Proceedings of the 16th International Conference on Computational Linguistics.

Goldsmith, John. 2001. Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics. 153-189.

Habash, Nizar. Generation-Heavy Hybrid Machine Translation. INLG-02. New York.

Henderson, J. 1999. Exploiting diversity for natural language parsing. PhD thesis, John Hopkins University.

McShane, Marjorie, Sergei Nirenburg, James Cowie, and Ron Zacharski. 2002. Embedding knowledge elicitation and MT systems within a single architecture. Machine Translation 17.271-305.

Nirenburg, Sergei, Konstantin Biatov, David Farwell, Stephen Helmreich, Marjorie McShane, Dan Ponsford, Victor Raskin and Svetlana Sheremetyeva. 1999. Toward descriptive computational linguistics. manuscript.

Oflazer, Kemal, Marjorie McShane, and Sergei Nirenburg. 2001. Bootstrapping morphological analyzers by combining human elicitation and machine learning. Computational Linguistics 27.59-85.

Pretorius, L. and S. E. Bosch. 2002. Finite-state computational morphology-treatment of the Zulu noun. South African Computer Journal.

Ratnaparkhi, Adwait. 1999. Learning to parse natural language with maximum entropy models. Machine Learning 34.151-175.

Wagner, R. A. and M. J. Fischer. 1974. The string-to-string correction problem. Journal of the Association for Computing Machinery, 21.168-173.

Zajac, Remi. 2001. Morpholog: Constrained and Supervised Learning of Morphology. Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning.