A Widely Used Machine Translation Service and its Migration to a Free/Open-Source Solution: the Case of Softcatalà

Xavier Ivars-Ribes[1,2] and **Víctor M. Sánchez-Cartagena**[1]
[1]Transducens Group
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain

[2]Softcatalà
http://www.softcatala.org

xavier.ivars@ua.es and vmsanchez@dlsi.ua.es

Abstract

Softcatalà is a non-profit association created more than 10 years ago to fight the marginalisation of the Catalan language in information and communication technologies. It has led the localisation of many applications and the creation of a website which allows its users to translate texts between Spanish and Catalan using an external closed-source translation engine. Recently, the closed-source translation back-end has been replaced by a free/open-source solution completely managed by Softcatalà: the Apertium machine translation platform and the ScaleMT web service framework. Thanks to the openness of the new solution, it is possible to take advantage of the huge amount of users of the Softcatalà translation service to improve it, using a series of methods presented in this paper. In addition, a study of the translations requested by the users has been carried out, and it shows that the translation back-end change has not affected the usage patterns.

1 Introduction

The vertiginous development of the Information and Communication Technologies (ICT) has allowed an increasing number of users to access information from world-wide resources through the Internet. However, as a result of linguistic barriers, this information may not be accessed by everyone.

Tearing down such linguistic barriers is harder for minority language speakers, as only a negligible amount of information is written in their language, and there is usually a lack of tools for the assimilation of content written in other languages. In addition, when the minority language co-exists with a more widespread one, often only the widespread language is used in ICT context.

Softcatalà[1] is a non-profit association created with the aim of encouraging the usage of the Catalan language in ICT and of facing the problems which have just been presented. As Machine Translation (MT) is a powerful tool that can be useful for both assimilation and dissemination, or the creation of content, Softcatalà offers a publically available on-line translator, recently moved to free/open-source software. This paper will study the role of Softcatalà (section 2) and provide an in-depth description of the software supporting the on-line translator (section 3). Next, an analysis of the translation service usage will be performed (section 4), and the methods being used to extract knowledge from its users and improve it will be described on section 5. Finally, some conclusions will be drawn.

2 Brief History of Softcatalà

Softcatalà was created in 1998 with the aim of providing high-quality linguistic and technological resources for Catalan speakers. Catalan, spoken by about 10 million people, was totally miss-

[1]http://www.softcatala.org/

F. Sánchez-Martínez, J.A. Pérez-Ortiz (eds.)
Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation, p. 61–68
Barcelona, Spain, January 2011. http://hdl.handle.net/10609/5648

ing in ICT context, without any software or technological resources available. Some volunteers from Softcatalà firstly translated Netscape Navigator[2], and a website was created to promote such work.

Over the years, Softcatalà's website has been growing with hundreds of applications and tools to help Catalan speakers using computers. Softcatalà has carried out the translation of well-known applications, such as OpenOffice.org, the Mozilla suite, GIMP; operating systems such as Fedora and Ubuntu; and desktop environments like GNOME. In addition, a set of tools to help software developers to translate their applications into Catalan has been developed, and is still being updated. It includes a term glossary, a style guide, a translation memory and a spell checker[3].

More than ten years ago, on September, 18th 2000, Softcatalà and University of Alacant agreed[4] to allow Softcatalà's website visitors to freely access machine translation service between Spanish and Catalan (in both directions[5]), powered by interNOSTRUM, a closed-source MT engine. The MT service soon became a very important part of the website, being the most used Softcatalà's service at the moment. In fact, currently more than 70% of the 1.2 million monthly visits reach the translation service page. Such a massive usage also has a big impact on the association's resources because of income from website advertisement. In fact, advertising is the main source of funding of Softcatalà, because it is a non-profit association without any public sponsorship. Having a stable income is quite important to keep a high-quality website and allow Softcatalà members to attend conferences and perform other activities to promote Catalan.

After ten years of using interNOSTRUM, the decision was made to switch to a free/open-source MT solution. This decision is discussed in the next section. Currently, the translation form can be accessed at `http://www.softcatala.cat/traductor` (see screenshot in figure 1).

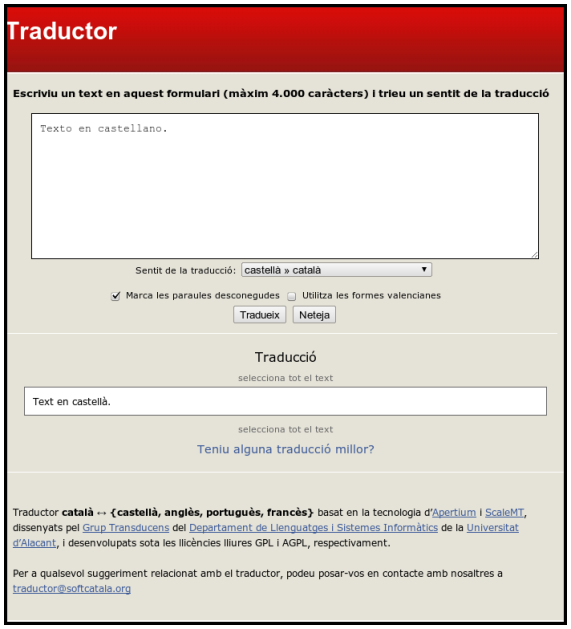

Figure 1: Translation form provided by Softcatalà.

3 Machine Translation Service: Switching to a Free/Open-Source Platform

As interNOSTRUM is a closed-source translation engine whose service is managed by University of Alacant, it was almost impossible to customise and extend according to Softcatalà's needs. The power of the great community of Softcatalà users to improve the MT service (following methods such as the ones presented in section 5) was thrown away. In addition, Softcatalà was totally defenseless against failures in the service hosted by University of Alacant.

With the aim of overcoming such problems, the association agreed to switch to a free/open-source solution, which is installed on Softcatalà's servers and does not depend on any external organisation. There are two key software pieces in the new MT platform: Apertium and ScaleMT.

The Apertium free/open-source MT platform (Forcada et al., 2009) has been chosen because of its efficiency and ease of modification. A very efficient and scalable engine is desirable in order to

[2] `http://www.softcatala.org/wiki/Navegador`

[3] All the tools are available at `http://www.softcatala.cat`

[4] `http://web.archive.org/web/20001214221300/www.softcatala.org/cgi-bin/gaudi/news/news.cgi?a=83&t=template.html`

[5] There are small differences between Catalan spoken in Catalonia and in Valencia, and the Spanish–Catalan direction can generate both variants.

address a high amount of user requests without investing huge amounts of money in hardware. Being a shallow transfer MT system makes Apertium quite efficient. In fact, it achieves translation speeds in the range of $10,000$ words per second in regular desktop computers. In addition, as in other rule-based MT systems, its rules and dictionaries can be easily updated. Both Apertium and interNOSTRUM have been developed by the Transducens research group and both follow the same translation paradigm, which ensures that the users perceive a smooth change. In addition, it can translate between additional language pairs useful for Catalan speakers: Portuguese–Catalan, English–Catalan and French–Catalan.

However, Apertium is not ready to be used by many concurrent clients out-of-the-box. For instance, it cannot run in coordination on many computers and, for each translation to be performed, spends a relatively high amount of CPU time loading resources. Fortunately, there are some free/open-source applications which mitigate these problems: ScaleMT (Sánchez-Cartagena and Pérez-Ortiz, 2010) and Apertium-service (Minervini, 2009). The first one was chosen because its architecture allows easily running the service on multiple servers and it provides a lower response time when processing many concurrent requests (Sánchez-Cartagena and Pérez-Ortiz, 2010). Apertium and ScaleMT are described in detail below.

3.1 The Apertium Free/Open-Source Machine Translation Platform

As pointed out before, Apertium[6] is an open-source platform for developing rule-based MT systems. It follows a shallow transfer approach and may be seen as an assembly line consisting of the following modules (see figure 2):

- A *de-formatter* which separates the text to be translated from the format information (RTF and HTML tags, whitespaces, etc.). Format information is isolated so that the rest of the modules treat it as blanks between words. Such special blanks are called *superblanks*.

- A *morphological analyser* which tokenises the source language (SL) text in surface forms and delivers, for each surface form, one or more *lexical forms* consisting of *lemma*, *lexical category* and morphological inflection information.

- A *part-of-speech tagger* which chooses one of the lexical forms corresponding to an ambiguous surface form.

- A *lexical transfer* module which reads each SL lexical form and delivers the corresponding target language (TL) lexical form by looking it up in a bilingual dictionary.

- A *structural shallow transfer* module (parallel to the lexical transfer) which uses a finite-state chunker to detect patterns of lexical forms which need to be processed for word reorderings, agreement, etc., and then performs these operations. In the case of less-related pairs, such as English–Catalan, a three-stage structural transfer is carried out.

- A *morphological generator* which delivers a TL surface form for each TL lexical form, by suitably inflecting it.

- A *post-generator* which performs orthographic operations such as contractions (e.g. Spanish *del=de+el*) and apostrophations (e.g. Catalan *l'oportunitat=la+oportunitat*).

- A *re-formatter* which restores the format information encapsulated by the first module.

Modules use text to communicate, which makes it easy to diagnose or modify the behaviour of the system. The Apertium MT engine is completely independent from the linguistic data used for translating between a particular pair of languages. Linguistic data is coded using XML-based formats; this allows for interoperability, and for easy data transformation and maintenance. Although engine and data are independent, both are licensed under the GNU GPL[7], which ensures the freedom of Softcatalà to modify the data.

[6] http://www.apertium.org/

[7] http://www.gnu.org/licenses/gpl.html

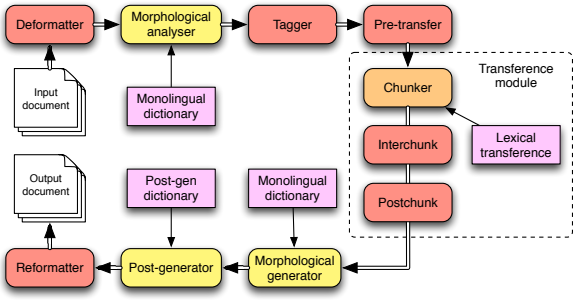

Figure 2: Main modules of the Apertium MT platform.

3.2 ScaleMT: a Free/Open-Source Framework for Building Scalable Machine Translation Web Services

It has been stated previously that Apertium is not designed to be accessed by many concurrent clients because it cannot run in coordination on many computers, and spends too many CPU cycles loading resources. ScaleMT[8] is a free/open-source framework, licensed under the GNU AGPL[9], which exposes existing MT engines as web services and avoids the aforementioned problems. Moreover, the MT engines do not need to be modified in order to act as web services.

Data from the Apertium rules and dictionaries should be kept in memory in order to avoid wasting CPU time by loading it repeatedly. As Apertium and many other translation engines commonly load the resources when they are launched, ScaleMT reuses them so that the same process performs many translations. Processes are kept in execution by simply keeping their standard input open, so that they behave as if they were translating a long text. Translation requests are queued and then written to the input of the right process (there should be, at least, one process for each language pair). These reused processes are called daemons. The different translation requests are separated by special *superblanks*, and a special option of Apertium, called *null flush*, is activated so that sending a null character after each translation request ensures that the translation is immediately available at the output after being generated, and it does not remain stored in buffers.

As well as the use of daemons improves the efficiency of the translation tasks, the architecture of ScaleMT allows the translation engine to easily run coordinately on many computers. As shown in figure 3, the ScaleMT platform consists of two main applications:

`ScaleMTSlave` runs on a machine with the translation engine installed and manages a set of running daemons; it performs the requested translations by sending them to the right daemon.

`ScaleMTRouter` runs on a web server; it processes the translation requests and sends them to the right `ScaleMTSlave` instance. The algorithm, which for each request chooses the right slave instance so that the work is fairly distributed between all the slaves, is explained in detail by Sánchez-Cartagena and Pérez-Ortiz (2010).

If the service supports many language pairs, servers may not have enough memory to run a daemon for each of pair. Consequently, the running daemons should be chosen accurately in order to achieve a good performance, and its number and machine distribution should be adapted to changing demands. ScaleMT features a placement algorithm based on the work by Tang et al. (2007) that is executed periodically and decides which daemons should run on each server.

The system is able to scale by adding new servers running `ScaleMTSlave`. These servers may be added manually, or we can let a dynamic server manager decide when to add or remove them based on the results of the placement algorithm. The servers added by the dynamic server

[8] `http://wiki.apertium.org/wiki/ScaleMT`
[9] `http://www.gnu.org/licenses/agpl.html`

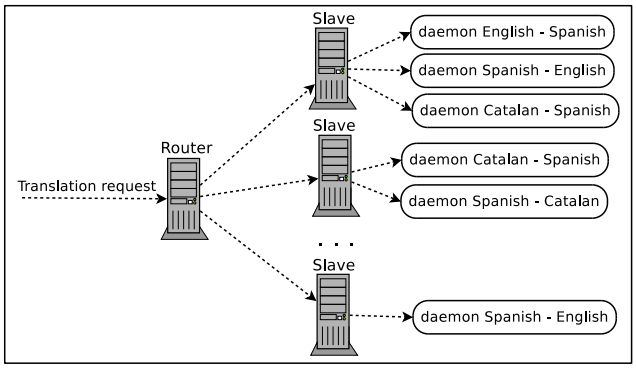

Figure 3: Architecture of the ScaleMT web service framework for machine translation.

manager may be physical machines from a local network or virtual machines from the cloud[10].

It has been decided to run only an instance of `ScaleMTSlave` on the same machine[11] as `ScaleMTRouter`. As it has enough memory to run a daemon for each language pair, the placement algorithm is not crucial. The average load[12] of the machine is around 1.15 , which is a relatively low value for a 4-processor computer. However, the architecture is ready to add more machines running ScaleMTSlave if necessary. No comparison can be established with interNOSTRUM because it runs on a different computer and load values are not comparable.

4 Translation Service Usage Analysis

As explained in section 2, the MT form is the most visited section of the website. It is accessed more than 850,000 times a month on average, with more than 3 million translations requested for 9 language pairs during October 2010, the first month after the MT platform switch. *Apertium.org*, the official Apertium website also features a translation form with 40 language pairs available, but only performed 380,000 translations during the same period.

In this section we will analyse the distribution of the translations requested among hours of the day, days of the month, and language pairs. We will also evaluate the impact of the translation engine switch and present some user opinions.

[10] `http://aws.amazon.com/ec2/`

[11] Quad core Intel(R) Xeon(R) 2.5 Ghz; 16 GiB RAM

[12] `http://en.wikipedia.org/wiki/Load_(computing)`

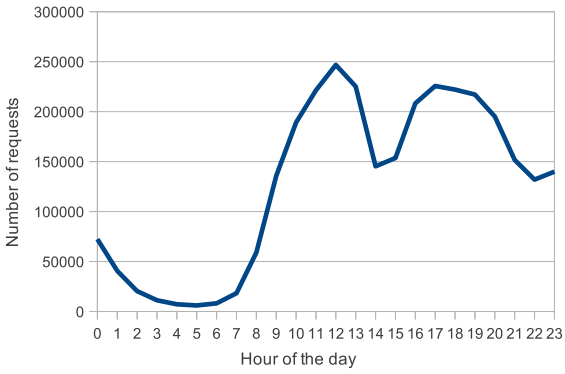

Figure 4: Hourly distribution of the translation requests received by the Softcatalà machine translation service during October 2010.

4.1 Hourly and Daily Distribution

During October 2010, the service received more than three million translation requests. Figure 4 shows the distribution of the requests over the hours of the day, and figure 5 (thick dark line) presents the number of visits the translation page received on the different days of the month. Note that a single visit may produce multiple translations, as the page is not reloaded.

It is clear that most of the translation requests are received during the working hours in Spain. The request rate is quite high from 9 to 20, with a sharp fall at lunchtime. Figure 5 also shows clearly that the translator is used much more from Monday to Friday and that usage drops during weekends. It is also worth mentioning that requests are lower on the 12th of October as it is a public holiday in Spain. The recently analysed usage patterns and the fact that feedback provided

Figure 5: Visits at the translator web page during the same 4-week period, comparing 2009 (light and thin, starting on the 28th of September) and 2010 (dark and thick, starting on the 27th of September).

comes mainly from companies and public institutions show that the service has an important professional usage.

All request logs analysed in this study can be downloaded from Softcatalà's website[13].

4.2 Impact of the Platform Switch on Service Usage

We are going to evaluate how users have reacted to the translation engine switch by comparing the visits received by the translation page before and after changing the MT engine. A comparison between requested translations would be slightly fairer, but translation requests were not logged with the old system. Figure 5 compares the daily visits received from the September, 28th 2009 to October, 25th 2009 with the ones received from September, 27th 2010 to October, 24th 2010. The compared periods start on Monday.

There is no significant variation between the 2009 and 2010 statistics, except for the 12th of October, which fell on a Monday on 2009 and on a Tuesday on 2010. As pointed out before, it is a public holiday in Spain, which explains why visits are relatively low on the 11th October 2010 (many people do not work on a Monday before a holiday) and on the 12th October 2010. In 2009, the MT service hosted by University of Alacant went down on the 11th of October, and was not fixed until two days later.

The analysed data shows that the migration of the MT platform has not affected service usage but has improved its reliability.

4.3 Language Pair Distribution

Figure 6 shows the distribution over the supported language pairs of the translation requests received during October 2010. The *Others* group contains

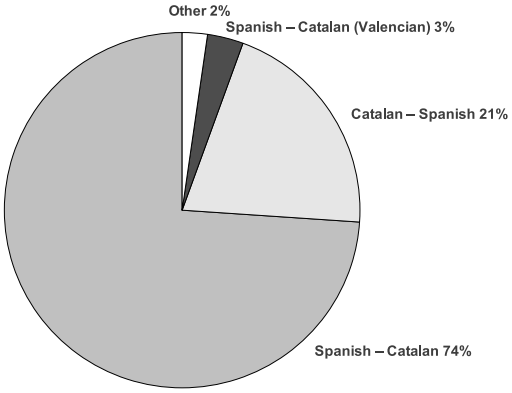

Figure 6: Pie chart showing the distribution over the different supported language pairs of the requests received by the Softcatalà translation service in October 2010.

the translations requested to the recently added language pairs: from English, Portuguese and French to Catalan and *vice versa*.

As people are used to visiting Softcatalà to translate between Catalan and Spanish, the new language pairs suffer from a very low usage. It is also notable that there are a high amount of translations from Spanish to Catalan when comparing them to translations from Catalan to Spanish.

4.4 User Opinion

Although the data analysis shows that the MT system migration has not affected its usage patterns, it is very useful to directly know the opinion of the users. Despite the fact that a exhaustive survey has not been carried out, some users have sent emails to the Softcatalà staff explaining that they had noticed an improvement in speed and translation quality.

[13] http://www.softcatala.org/apertium/logs/requests-2010-10.tgz

es-ca[1]	ca-es	en-ca	ca-en
cortadora	AMPA	nursery	penitenciari
Sócrates	Moodle	trinity	comanda
Freud	Martini	summertime	incompliment
pH	burret	default	enganxines
estiramiento	perdigot	anymore	Acta

Table 1: Some frequent unknown words for different language pairs automatically extracted from the translations requested.

5 Using the Crowd to Improve Linguistic Data

The migration from interNOSTRUM to Apertium is not only of benefit to Softcatalà, but also benefits the Apertium community. In this section we explain different ways the Apertium platform and its linguistic data are being improved thanks to the high amount of users of the translation service.

5.1 Automatic Unknown Word Extraction

When a word (surface form) is not present in the dictionaries used by the *morphological analyser* of the Apertium pipeline, it is marked as unknown with a special character (an asterisk). Searching this special character during the translation process can help us to find unknown words included in translations being requested by users.

During the deployment of ScaleMT, we slightly changed its behaviour to log every word the system is not able to translate. Then, lists of unknown words sorted by frequency may be easily obtained, and these lists can be very useful to improve coverage of the dictionaries by firstly adding the words with most impact on the translation quality delivered to the users.

Table 5.1 shows five of the most frequent unknown words in some language pairs, obtained automatically from the translation logs. Apertium dictionaries for Spanish–Catalan have a higher coverage. Consequently, the frequent unknown words from that language pairs are proper names, or some domain-restricted words. On pairs with lower coverage, as English–Catalan, more common words appear on the list. The data shown in this table has been extracted from translation requests received during five days, but a more representative amount of data would be needed in order to choose the unknown words with more impact on the translation quality.

5.2 Alternative Translations Suggested by Users

User suggestions are the most important feedback obtained from the service. There is a very active community around Softcatalà, but the majority of its members do not know how to improve the system by themselves.

We have developed a new form in the translator web page where users can suggest a better translation after receiving the translation. Users can write a better translation of the source text according to their personal criteria.

Parallel sentences, composed by the source sentence, its machine translation and the user suggestion, are stored in a database with other information as the language pair or the date and time the suggestion has been recorded.

We have also built a web interface to show the recorded suggestions. It is intended to be visited periodically by Apertium language pair maintainers, who may check the suggestions and mark them as *invalid*, if the suggestion is not correct — e.g. it has spelling errors —, *won't fix*, if it is a valid one but cannot be added to the system because of any limitation of the engine — for instance, the Apertium platform does not perform word sense disambiguation — or *solved*, if the Apertium linguistic data has been updated to fix the error. Some examples of feedback received can be seen in figure 7. The interface also provides filters to show only suggestions with one of the above tags or to show suggestions related to a single language pair. Softcatalà also offers an email address where users can send more complex suggestions.

An example of useful feedback is a wrong *ca-es* translation which helped to find a bug:

"Durant molt de temps vaig anar." → *"Durando mucho tiempo fui."*

The correct translation of *Durant* in this context is *Durante*, which is a preposition, but not *Durando*, which is a gerund. It was found that the Apertium *part-of-speech tagger* was not choosing the correct part of speech, due to bug when being invoked with the *null-flush* option, and it is currently being fixed.

168	es-ca_valencia	Bolsa	Borsa	Bossa	PENDING	
169	es-ca_valencia	cañizo	*cañizo	canyís	PENDING	
170	es-ca_valencia	solera	*solera	solera	PENDING	
171	en-ca	They adopted the Chinese writing system and created excellent bronze swords.	Van adoptar el sistema d'escriptura xinès i espases de bronze excel·lents creades.	Van adoptar el sistema d'escriptura xinès i crearen excel·lents espases de bronze.	PENDING	
175	es-ca_valencia	su labor al frente de esta entidad	la seua labor al capdavant d'aquesta entitat	la seua llabor al capdavant d'aquesta entitat	PENDING	

Figure 7: Some feedback received from users. There are semantic and syntactic ambiguity errors (168 and 171), missing dictionary entries (169 and 170) and also wrong feedback provided by users (175).

6 Conclusions and Future Work

As a conclusion, we present a review of the benefits the different stakeholders gain with the new free/open-source machine translation platform recently adopted by Softcatalà.

Softcatalà gets an up-to-date MT system and control over its deployment, which hopefully will provide a better service to their users. Additionally, there is a perfect framework for the improvement of the linguistic data. Its GPL licence will make the improvements available to the Apertium community. Moreover, the continuous improvement of the service and the scalable architecture make the growing of the user base easier.

The users of the website get a more stable translation service which improves with their suggestions. Users who frequently translate texts of the same category will notice this improvement more strongly.

Regarding future features, the suggestion web interface may be modified to make the maintainers' job easier. For instance, suggestions containing unknown words may be directly redirected to the automatic unknown word extractor, the steps of the translation pipeline may be included to clarify the error, etc.

7 Acknowledgements

We would like to thank Juan Antonio Pérez-Ortiz and Francis Tyers for his useful advice, Toni Hermoso and Pau Iranzo for his help at Softcatalà, and Jimmy O'Regan for testing the suggestion web interface. This work has been supported by Spanish Ministerio de Ciencia e Innovación through project TIN2009-14009-C02-01, Generalitat Valenciana through grant ACIF/2010/174 from VALi+d programme and European Community under the Information Society Technologies Programme (IST-1-4.1 Digital libraries and technology-enhanced learning) of the 7th framework programme - Project FP 7-ICT-2007-1.

References

Forcada, M., Tyers, F., and Ramírez-Sánchez, G. (2009). The Apertium machine translation platform: five years on. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 3–10.

Minervini, P. (2009). Apertium goes SOA: an efficient and scalable service based on the Apertium rule-based machine translation platform. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 59–66.

Sánchez-Cartagena, V. M. and Pérez-Ortiz, J. A. (2010). Scalemt: a free/open-source framework for building scalable machine translation web services. *The Prague Bulletin of Mathematical Linguistics*, 93:97–106.

Tang, C., Steinder, M., Spreitzer, M., and Pacifici, G. (2007). A scalable application placement controller for enterprise data centers. In *Proceedings of the 16th international conference on World Wide Web*, pages 331–340. ACM.