The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 221–232

Rule-Based Machine Translation for the Italian–Sardinian Language Pair

Francis M. Tyers,[a b] Hèctor Alòs i Font,[a] Gianfranco Fronteddu,[e]
Adrià Martín-Mor[d]

[a] UiT Norgga árktalaš universitehta, Tromsø, Norway
[b] Arvutiteaduse instituut, Tartu Ülikool, Tartu, Estonia
[c] Universitat de Barcelona, Barcelona
[d] Universitat Autònoma de Barcelona, Barcelona
[e] Università degli Studi di Cagliari, Cagliari

Abstract

This paper describes the process of creation of the first machine translation system from Italian to Sardinian, a Romance language spoken on the island of Sardinia in the Mediterranean. The project was carried out by a team of translators and computational linguists. The article focuses on the technology used (Rule-Based Machine Translation) and on some of the rules created, as well as on the orthographic model used for Sardinian.

1. Introduction

This paper presents a shallow-transfer rule-based machine translation (MT) system from Italian to Sardinian, two languages of the Romance group. Italian is spoken in Italy, although it is an official language in countries like the Republic of Switzerland, San Marino and Vatican City, and has approximately 58 million speakers, while Sardinian is spoken principally in Sardinia and has approximately one million speakers (Lewis, 2009).

The objective of the project was to make a system for creating almost-translated text that needs post-editing before being publishable. For translating between closely-related languages where one language is a majority language and the other a minority or marginalised language, this is relevant as MT of post-editing quality into a lesser-resourced language can help with creating more text in that language.

As described below, Sardinian is not a fully-standardised language. This means that linguistic resources are scarce, even if the orthographic norm chosen for this

Corresponding author: francis.tyers@uit.no
Cite as: Francis M. Tyers, Hèctor Alòs i Font, Gianfranco Fronteddu, Adrià Martín-Mor. Rule-Based Machine Translation for the Italian–Sardinian Language Pair. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 221–232. doi: 10.1515/pralin-2017-0022.

project was the Limba Sarda Comuna (*Common Sardinian Language*, or LSC), the one officially approved by the island's autonomous government in 2006. In fact, the main aim of the project was to create a tool that would foster text production in Sardinian, especially in areas such as administration and Wikipedia.

The remainder of the article is laid out as follows: In section 2 we provide some linguistic background to Sardinian. This is followed by a description of the platform used to build the MT system in section 3. Section 4 describes the development of the system, including resources that were reused. Then section 5 gives an evaluation of the system. Finally, we comment on possible future work in section 6 and give some conclusions in section 7.

2. Sardinian

The Sardinian language is a Romance language spoken by approximately one million people on the island of Sardinia, together with other Romance languages such as Tabarchino Ligurian (on the islands of San Pé and Sant'Antióccu), Algherese Catalan (in the city of L'Alguer), Sassarese (in the city of Sassari) and Gallurese Corsican (in Gaddùra).[1]

At the institutional level, some of these languages are recognised by the regional government. However, the use of Sardinian language is virtually non-existent at any educational level, as well as in many fields of the public sphere (media, newspapers, administration, etc.). Still, the use of Sardinian is widespread. According to (Oppo, 2007) only 2.7% per cent of the population in Sardinia does not have any competence (either active or passive) in "any local language".

Sardinian, classified as "definitely endangered" by UNESCO,[2] is spoken across most of the island despite the fact that, because of its great internal variety, two macro-varieties are often distinguished: northern (Logudorese and Nuorese) and southern (Campidanese). The existence of these two macro-varieties is one of the controversial factors when it comes to the standardisation of the language. At present, there are movements who advocate for different standardisation models and which, broadly, correspond to northern and southern regions.

On the one hand, there is a group that defends a double standard, following the Norwegian model. This model, which is basically followed in the south, has received endorsement by the provincial government of Casteddu, which has officially adopted a "southern" standard described in the document *Arrègulas po ortografia, fonètica, morfologia e fueddàriu de sa Norma Campidanesa de sa Lìngua Sarda* (Comitau Scientìficu po sa Norma Campidanesa de su Sardu Standard, 2009). On the other hand, the Limba Sarda Comuna (LSC) has been proposed as the standard form for all varieties of Sardinian. It is an evolved version of the Limba Sarda Unificada (LSU), which was in turn the result of an experts' committee called by the Sardinian government in 2001.

[1] Toponyms are written in the local languages. There are, apart from these, other linguistic islands which result from migrations, such as Venetian and Romanisku.

[2] `http://www.unesco.org/languages-atlas/en/atlasmap/language-id-337.html` and `www.unesco.org/culture/languages-atlas/en/atlasmap/language-id-381.html`

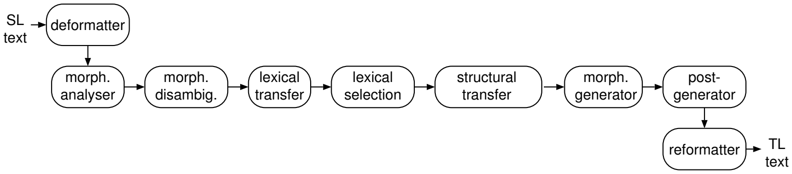

Figure 1. The modular architecture of the Apertium MT platform. Modules communicate using Unix text pipes.

In 2006, the Sardinian government adopted the LSC as a co-official language, alongside Italian, for the publication of official documents. The LSC is also the form chosen by several publishing houses and websites.

The existence of these two proposals implies that all initiatives concerning the Sardinian language must first take a stand on the issue of the standardisation model. The Sardinian Wikipedia, for instance, allows its users to mark the variety in which they write by adding a flag.

In October 2016, at the time of the writing of this article, the Sardinian Wikipedia has 5,230 content pages,[3] out of which 1,525 are written in Logudorese,[4] 776 in LSC,[5] and 295 in Campidanese.[6] Other digital products, such as Facebook (Beccu and Martín-Mor, 2017), Telegram (Martín-Mor, 2017) and Ubuntu,[7] have been partially localised into Sardinian basing mainly on the LSC model.

Indeed, according to Cheratzu (2015), textual and literary production in LSC is clearly greater in number than any other. Therefore, basing on textual production and resource availability, we decided to use LSC as the standard form of the Sardinian language in our project. Italian was chosen as the source language for our project. Despite the fact that linguistic resources (and competent writers) are scarce even for LSC, it was deemed appropriate, given the fragile situation of the Sardinian language, to facilitate the creation of contents in LSC from Italian (i.e., documents issued by the government, websites, newspapers, etc.).

3. Platform

The system is based on the Apertium machine translation platform (Forcada et al., 2011). The platform was originally aimed at the Romance languages of the Iberian peninsula, but has also been adapted for other, more distantly related, language pairs. The whole platform, both programs and data, are licensed under the Free Software

[3] https://sc.wikipedia.org/wiki/Ispetziale:Statistics

[4] https://sc.wikipedia.org/wiki/Categoria:Logudoresu

[5] https://sc.wikipedia.org/wiki/Categoria:Limba_Sarda_Comuna

[6] https://sc.wikipedia.org/wiki/Categoria:Campidanesu

[7] https://wiki.ubuntu.com/Ubuntu-Sardu/

Foundation's General Public Licence (GPL)[8] and all the software and data for the 43 supported language pairs (and the other pairs being worked on) is available for download from the project website.

3.1. Pipeline

A typical translator built with Apertium consists of 9 modules which communicate between each other using standard Unix pipes. This eases diagnosis, the insertion of new modules, etc. The modules comprise of the following:

- A **deformatter** which encapsulates any formatting (e.g. HTML or XML tags etc.) information in the input stream.
- A **morphological analyser** which for each surface form in the stream returns a sequence of possible analyses.
- A **part-of-speech tagger** which out of the possible analyses for a given word returns the most probable analysis. This is based on either first-order HMM or on HMM in combination with Constraint Grammar (Bick and Didriksen, 2015).
- A **lexical transfer** module which for each unambiguous source language lexical form returns one or more target language lexical forms.
- A **lexical selection** module which for each source language lexical form with more than one target language translation uses a set of rules operating on source-language context to choose the most adequate translation in the target language.
- A **structural transfer** module which performs syntactic and morphological operations to convert the source language intermediate representation into the target language intermediate representation. Common operations include insertion, deletion and substitution of lexical units, agreement between lexical units for e.g. gender, number and case, etc. The structural transfer module calls the lexical transfer module.
- A **morphological generator** which for each target language lexical form returns a surface (inflected) form.
- A **postgenerator** which performs orthographic operations, for example elision (such as *da+il=dal* in Italian).
- A **reformatter** which de-encapsulates any formatting, leaving it untouched.

Figure 1 gives an example pipeline. The data used by these modules are by and large specified in XML files and compiled into binary forms for use by the modules.

4. Development

The development of the Italian-Sardinian pair owes a lot to previous work on other language pairs. In this case, most of the lexical and morphological resources for Italian were taken from the Italian–Catalan pair (Toral et al., 2011), while part of the lexical and morphological resources for Sardinian was taken from the Sardinian–Catalan pair existing a prototype in the Apertium project. In parallel to our development of the

[8] https://www.gnu.org/licenses/gpl-3.0.en.html

Italian–Sardinian pair, developers from Prompsit Language Engineering were working on an Italian–Spanish pair, so we cooperated in the improvement of the resources for Italian.

4.1. Analysis

The development began with an analysis oriented to:

- collecting free linguistic resources for the dictionaries;
- collecting monolingual and bilingual corpora;
- systematically comparing the source and the target languages in order to understand what structural changes exist between them.

The contrastive analysis between Italian and Sardinian led to more than one hundred examples of translations the translator was expected to give, but a morpheme-by-morpheme translation would not, e.g.

- Nella mia terra. → In sa terra mea. ("In my land")
- Bellissimi. → Bellos a beru. ("Very beautiful')
- Darmi. → Mi dare. ("To give me")

These observed differences were used in creating the transfer rules.

4.2. Morphological dictionaries

The Italian morphological dictionary is, for the most part, the one used in the Italian–Catalan translator. However, some work has been done to extend and fix verbal paradigms. In addition, some 2,000 lemmas were added from the free/open-source resource *Morph-it* (Zanchetta and Baroni, 2005).

A first version of the Sardinian morphological dictionary already existed. It was based on the "experimental" norms of LSC (Regione Autonoma della Sardegna, 2006). It was augmented with data from the spell checker provided by the regional government of Sardinia.[9]

An important lack of proper nouns in the spell checker was detected, so we partially solved it adding a few hundreds of the most common person and family names in Sardinia, as well as the names of all Sardinian municipalities and Italian regions. It is worth adding that many place names are not yet standardised, e.g. the names of the countries and capitals. We added a few of the most common.

4.3. Morphological disambiguation

Romance languages have a fair amount of morphological ambiguities. Fortunately for developers of rule-based machine translation systems between these languages, they share most ambiguities, so most of the time selecting the wrong morphological analysis does not imply a bad translation, a *free ride*. For instance, this is generally the case for words finishing in -ista (like *comunista*, 'communist') that may be both adjectives or nouns. Since this ambiguity happens to be in both the source and the

[9] `http://www.sardegnacultura.it/cds/cros/`

Dictionary	Entries
Sardinian	51,743
Italian	35,099
Sardinian–Italian	25,484

Table 1. Dictionaries in the MT system. The final translator is assembled as the intersection of the entries in these dictionaries.

target language, e.g. a wrong analysis of *comunista* as a noun in *il partito comunista* would still give a good translation as *su partidu comunista*.

Probably the most frequent ambiguity in Italian, which is shared by French, Spanish and Catalan too, is *la* that can be both a definite article (feminine *the*) or a pronoun (*her*). In Sardinian these two analyses have different forms so it was necessary to resolve the ambiguity.

In addition to training the tagger on a corpus of 17,000 words from TED talks and Wikinews,[10] we added a set of 30 rules using rules written using Constraint Grammar (CG) (Bick and Didriksen, 2015). CG rules for Italian mainly deal with the disambiguation between imperative verbal forms with enclitic pronouns and adjectives (e.g. *centrali* as 'central', masculine plural, or 'centre them'), and contractions of prepositions and determiners (e.g. *dalle* as 'from the.F.PL' or 'give.IMP.2.SG them'; *dai*, 'from the.M.PL', 'give.IMP.2.SG' or 'give.PRI.2.SG'; *dei*, 'of the.M.PL' or 'gods').[11]

Not every morphological ambiguity can be easily solved. A clear case is *sono*, which can be "I am" or "they are". This ambiguity does not exist in Sardinian: "I am" is *so*, while "they are" is *sunt*. Both Italian and Sardinian are pro-drop languages, the subject pronoun can be omitted since it can be almost always inferred from the context (especially from the verb form). So it happens that we often have to guess whether it is about "I" or "they" when dealing with *sono*. By default we assume third person based on our target domain of encyclopaedic texts.

4.4. Transfer lexicon

The transfer lexicon was one of the tasks of the project that has taken longer because of the lack of free bilingual dictionary. In total 25,484 lemmas have been added to the bilingual dictionary, about a half of them by hand using frequency lists of words. Most of the time Antonino Rubattu's *Universal Dictionary Italian-Sardinian* and Mario Casu's *Logudorese-Italian vocabulary* were consulted. However, when using the dictionaries we made efforts to choose a form which was also found in the LSC spell checker.

[10] Corpus provided by Prompsit Language Engineering, `http://www.prompsit.com`

[11] = masculine, F = feminine, SG = singular, PL = plural, IMP = imperative, PRI = present of indicative, 2 = second person.

4.5. Lexical selection

Because of the short time in which the translator was developed only 35 lexical selection rules have been added. The lack of bilingual corpora did not allow us to automatically infer any rules. For instance, a difficult case is the word *corso*, which may be both "street" and "Corsican". Both meanings are found often in similar contexts and have different translations in Sardinian. Rules define that, if the noun is found in plural or is preceded by the preposition "in", "Corsican" is preferred, otherwise "street" is chosen.

4.6. Structural transfer rules

Apertium, as a rule, translates lemmas and morphemes one by one. Obviously, this does not always work, even for closely related languages. Structural transfer rules are responsible for modifying morphology or word order in order to produce "adequate" target language. In all, we have defined 89 such transfer rules.

4.6.1. Noun-phrase internal agreement

Most of the rules deal with noun-phrase internal agreement both in gender and number. Two situations have to be distinguished. On one hand, the target language has combinations of gender and/or number that do not exist in the source language. About 8% of the nouns have been labelled in the bilingual dictionary as requiring that the gender or the number needs to be determined when translating from Italian into Sardinian. In this case, the actual gender and/or number is obtained from other words in the noun phrase.

On the other hand, a noun in the target language may have a gender and/or a number different than in the source one. This is the case for 7% of the nouns in the bilingual dictionary. In this case, the gender and/or the number of the other words of the noun phrase must be modified to agree with the name.

4.6.2. Possessives

Possessives also require a correct delimitation of noun phrases since they must be moved from its beginning to the end (1).

(1) La sua apparente indifferenza .
S' aparente indiferèntzia sua .
"His apparent indifference."

4.6.3. Tenses

Tenses in Sardinian tend to be often analytical. A number of tenses which are synthetic in Italian, as well in most of the Romance languages, are conjugated in Sardinian by means of verbal periphrasis, e.g. the future (2a) and conditional (2b) and historical. In addition, LSC does not have the absolute past tense of Italian, and uses the present perfect (2c).

(2) a. Canterò
Apo a cantare
“I will sing”

b. Canterei
Dia cantare
“I would sing”

c. Cantai
Aia cantadu
“I sang”

All these transformations have been done by means of specific transfer rules.

4.6.4. Clitic pronouns

In Italian clitic pronouns must be placed after the verbs in infinitive, imperative and gerund forms, as well as with past participles when used as past gerunds. Instead, in Sardinian in infinitive forms clitics should be placed before the verb. As a result, for instance *cantarla* (“to sing it”) must be translated as *la cantare*.

4.6.5. Change of the auxiliary verb

In Italian the present continuous construction uses the auxiliary *stare*, while in Sardinian the auxiliary *èssere* is used instead of *istare* (3).

(3) Io sto studiando.
Deo so istudiende.
“I am studying.”

4.7. Post-generation rules

After the generation of the raw version of the translation some additional processing has to be done. In most of the cases, this means to apostrophise. For instance, *l'accumulazione* (“the accumulation”) is translated first of all as *sa acumulatzione*, where a special symbol is produced by the morphological generator, warning that the word *sa* is liable to receive modifications. A set of rules define in which case words in Sardinian are apostrophised. In the same way, the Sardinian words *no* and *ne* (“no” and “nor”) may be changed to *non* and *nen* according to the context.

5. Evaluation

The system has been evaluated in two ways. The first is its coverage.[12] The second is the error rate of two pieces of text produced when comparing with a post-edited version of them.

[12] Here coverage is defined as naïve coverage, that is for any given surface form at least one analysis is returned. This may not be complete.

Corpus	**Tokens**	**Coverage** (%)
Wikipedia 10%	34,736,257	89.3
UD Italian	285,199	96.4

Table 2. Naïve vocabulary coverage. This is the percentage of tokens which receive at least one analysis from the morphological analyser. The coverage of Wikipedia is lower due to the large number of proper nouns and foreign words.

Words	**Unknown words**	**WER**	**TER**
2,033	9.4%	9.9%	6.3%

Table 3. Word Error Rate and unknown words over the 2,033 word test corpus.

5.1. Coverage

Table 2 presents the lexical coverage of the system over two corpora. The first was a subset of the Italian Wikipedia, which was created by randomly selecting 10% of the sentences from the Italian Wikipedia as of May 2016. The second corpus is the text from the Italian treebank in the Universal Dependencies project.[13]

5.2. Translation quality

We measured translation quality using two metrics: Word error rate (WER), which is based on the Levenshtein distance (Levenshtein, 1966) and was calculated for using the `apertium-eval-translator` tool; and Translation Error Rate (TER, Snover et al. (2006)). Metrics based on word error rate have been chosen for a number of reasons. Firstly we would like to be to compare the system against systems based on similar technology, and to assess the usefulness of the system in a real setting, that is of translating for **dissemination**. Secondly, the reference translation is a postedition, whereas most MT evaluation metrics use pre-translated references. Using a more commonly used metric in an uncommon setting would give deceptively good results.

A corpus of 2,033 words (53 sentences) was extracted from Wikipedia. The average length of a sentence was 42 words. This was the first paragraphs of the last two texts put in the section "vetrina" ("showcase") at the time of the GSoC final evaluation (more or less 1000 words per text). Wikipedia texts were selected, as this is one of the major uses for Apertium translators, especially as they are used by the Wikimedia Content Translation Tool.[14] The section "vetrina" is a pseudo-random selection (not done by the machine translator developers) of quality Wikipedia articles.

[13] http://universaldependencies.org

[14] https://www.mediawiki.org/wiki/Content_translation

The vast majority of unknown words are proper names (foreign person, family and place names) as well as foreign words (e.g. in French or English).

The scores are similar to or slightly better than those for other translators in the Apertium platform for Romance languages, for example the Catalan–-Occitan system achieves a WER of 9.6% (Armentano-Oller and Forcada, 2006) and the Spanish–Aragonese 16.8%, (Martínez Cortés et al., 2012).

5.3. Qualitative evaluation

Along with the quantitative evaluation of post-edition effort, we also performed a qualitative evaluation to determine where the system can be improved. Based on the final evaluation text, we have detected two major issues: 1) incorrect disambiguation of the verb *avere*; and 2) the absolute past tense transfer rule. In the examples that follow, the Italian phrase is presented on the first line, followed by the current translation into Sardinian produced by the system on the second, the correct translation on the third, and an English translation on the fourth.

5.3.1. Incorrect disambiguation of "avere"

The Italian verb *avere* ("to have") may be both an auxiliary and a lexical verb. These have different translations in Sardinian (4). The distinction between both verbs *avere* is done in the tagger. Nevertheless, it happens that when the auxiliary is separated from the participle by an adverb, *avere* is wrongly tagged as a lexical verb (5).

(4) a. Ho cantato.
Apo cantadu.
"I have sung."

b. Ho un gatto.
Tèngio unu gatu .
"I have a cat."

(5) Non aver adeguatamente protetto la Francia.
* Non tènnere in manera adeguada amparadu sa Frantza.
Non àere in manera adeguada amparadu sa Frantza.
"Not having adequately protected France".

This issue has to be solved in the morphological disambiguation step, for example using CG rules.

5.3.2. Absolute past tense

As seen before, an absolute past tense exists in Italian, but not in LSC, in which the present perfect is used instead. A transfer rule constructs the past perfect adding the Sardinian auxiliary verb *àere* ("to have") with the same person and number as the Italian verb and the past participle of the Sardinian translation of the lemma. Nevertheless, in Sardinian, as well as in Italian, several verbs are conjugated with the auxiliary verb "to be", particularly the verbs of movement and the verb "to be" itself. The current transfer rule is too simple and does not take into account this fact 6a, so needs to be improved.

(6)

a. Sfuggì.
* Aiat isfugidu.
Fiat isfugidu.
"He escaped."

b. Fu.
* Aiat istadu.
Fiat istadu.
"He was."

6. Future work

Aside from fixing the problems outlined in section 5.3, we would also like to see more translation systems for Sardinian. We have an experimental system for Sardinian–Catalan which is particularly relevant as Catalan is one of the larger languages in direct contact with Sardinian. We are also interested in working on Corsican as it is also spoken in Sardinia.

7. Conclusions

We have presented the first ever MT system from Italian to Sardinian. The performance is similar to other translators created using the same technology. It translates texts sufficiently well for post edition, although there remains a lot of work to do with respect to improving lexical coverage, and some work to do on improving the disambiguation and transfer rules. The system is available as free/open-source software under the GNU GPL and the may be downloaded from Apertium SVN.[15]

Acknowledgements

We would like to thank Mikel Forcada for his constant encouragement and comments on an earlier version of this manuscript. Thanks also go out to Diegu Corràine for clarifications on the LSC standard. The project was partially funded by a stipend from the Google Summer of Code. We would also like to thank Gema Ramírez Sánchez, and the anonymous reviewers.

Bibliography

Armentano-Oller, Carme and Mikel L. Forcada. Open-source machine translation between small languages: Catalan and Aranese Occitan. In *5th SALTMIL workshop on Minority Languages*, pages 51–54, 2006.

Beccu, A. and A. Martín-Mor. Sa localizatzione de Facebook in sardu. *Revista Tradumàtica*, 14, 2017.

[15] `http://www.apertium.org`

Bick, Eckhard and Tino Didriksen. CG-3 – Beyond Classical Constraint Grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, pages 31–39. Linköping University Electronic Press, Linköpings universitet, 2015.

Cheratzu, Francesco. Sa Chirca. In Mura, Riccardo and Maurizio Virdis, editors, *Caratteri e strutture fonetiche, fonologiche e prosodiche della lingua sarda. Il sintetizzatore vocale SINTESA*. 2015.

Comitau Scientìficu po sa Norma Campidanesa de su Sardu Standard. Arrègulas po ortografia, fonètica, morfologia e fueddàriu de sa Norma Campidanesa de sa Lìngua Sarda, 2009.

Forcada, M. L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, 2011.

Levenshtein, Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

Lewis, M. Paul, editor. *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, sixteenth edition, 2009.

Martín-Mor, A. La localització de l'apli de missatgeria Telegram al sard: l'experiència de Sardware i una aplicació docent. *Revista Tradumàtica*, 14, 2017.

Martínez Cortés, Juan Pablo, Jim O'Regan, and Francis Tyers. Free/Open Source Shallow-Transfer Based Machine Translation for Spanish and Aragonese. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

Oppo, Anna. Conoscere e parlare le lingue locali. In Oppo, Anna, editor, *Le lingue dei sardi: una ricerca sociolinguistica*, chapter 1, pages 6–45. Regione Autonoma della Sardegna, 2007.

Regione Autonoma della Sardegna. Limba Sarda Comune. Norme linguistiche di riferimento a carattere sperimentale per la lingua scritta dell'Amministrazione regionale, 2006. URL `http://www.regione.sardegna.it/documenti/1_72_20060418160308.pdf`.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the Conference of the Association for Machine Translation in the Americas*, 2006.

Toral, Antonio, Mireia Ginestí-Rosell, and Francis M. Tyers. An Italian to Catalan RBMT system reusing data from existing language pairs. In Sanchez-Martínez, F. and J.A. Perez-Ortiz, editors, *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 77–81, 2011.

Zanchetta, Eros and Marco Baroni. Morph-it! A free corpus-based morphological resource for the Italian language. *Corpus Linguistics 2005*, 1(1), 2005. ISSN 1747-9398.

Address for correspondence:
Francis M. Tyers
`francis.tyers@uit.no`
Giela ja kultuvvra instituhta
UiT Norgga árktalaš universitehta,
N-9018 Romsa,
Norway