Shooting at flies in the dark: Rule-based lexical selection for a minority language pair

Linda Wiechetek[1], Francis M. Tyers[2], and Thomas Omma[3]

[1] Giellatekno
Romssa Universitehta, Norway
linda.wiechetek@uit.no
[2] Dept. Lleng. i Sist. Inform., Universitat d'Alacant, Spain
ftyers@dlsi.ua.es
[3] Divvun, Sámi Parliament, Norway
thomas.omma@uit.no

Abstract. This paper presents a set of rules which form the prototype lexical selection component of a rule-based machine translation system between two closely-related minority languages, North Sámi and Lule Sámi. While the languages have comprehensive monolingual computational linguistic resources, they lack bilingual resources. One-to-one relations in the lexicon dominate, but there are also more complex relations that require lexical selection using both lexical and syntactico-semantic context. An evaluation was performed over a set of 11 word pairs, which shows that constructing lexical selection rules and doing research on a North Sámi–Lule Sámi contrastive lexicon is an interrelated process. Other lesser-resourced language pairs will benefit from the use of lexical selection rules as the relevance of lexical selection increases with the divergence of the languages.

Key words: Sámi languages, rule-based lexical selection, MT

1 Introduction

North Sámi and Lule Sámi belong to Sámi group of languages which is a subfamily of the Finno-Ugric language family. They are spoken in the north of the Nordic countries. North Sámi has between 15,000 and 25,000 speakers and Lule Sámi has around 2,000 speakers.

The languages are neighbours and are mutually intelligible, although often the majority language is used as a *lingua franca*. Despite their mutual intelligibility, orthographic differences impede reading comprehension between the two languages.

A large amount of word roots and morphosyntactic categories (cases, inflectional and derivational patterns etc.) are shared between the languages. However, despite the relatedness, there are a number of challenges in machine translation for the two languages.

Both languages have phonemic orthographies, with differences resulting from differing conventions in standardisation. These differences can largely be handled by rules, thus a bilingual dictionary between the two was created from scratch by simply converting the orthography.

This process provides an adequate lexicon, but when inspecting the resulting translations with a native Lule Sámi speaker, the situation was found to not be as simple as originally assumed.

On the syntactic level, there are obvious differences, such as an asymmetry in the case system (North Sámi locative being expressed by Lule Sámi elative and inessive) and differences in word order, Lule Sámi tends towards an OV word order, where North Sámi tends towards VO.

Other differences are less obvious, while North Sámi can express the semantic notion of path with both a *-ráigge* 'along' compound construction[4] (1-a) and a genitive case adverbial (1-b), Lule Sámi (1-c) lacks the simple genitive construction. Therefore, the translation of(1-a) is .

(1) a. Soai boahtiba geainnoráigge. (North Sámi)
They-DU come this way-along
'They come along the way.'

b. Soai boahtiba dán geainnu. (North Sámi)
They-DU come this way-GEN
'They come along this way.'

c. Såj boahteba gæjnnorájge. (Lule Sámi)
They-DU come this way-along
'They come along this way.'

The languages also diverge on the lexical level, and although the automatically constructed bilingual lexicon often provides adequate translations, in many cases word use is actually quite different. In some cases historical word roots are different, in other cases, words in one of the languages have acquired a new sense which does not exist in the other language or appear in specific syntactic or semantic construction which is resolved differently in the other language.

The need for lexical selection[5] came up when seemingly straightforward translations were not accepted by Lule Sámi native speakers. The errors could not be fixed by correcting the bilingual dictionary because the translation error lies not in any erroneous entries, but in the one-dimensionality of the entries. Less related languages will profit from lexical selection to an even larger extent as natuarlly semantic concepts will diverge more the further languages and the speech communities differ from each other.

[4] The latter part of the lexicalised construction is originally a genitive too.

[5] Lexical selection is defined by [1] as the "principled selection of a) lexical items and b) the syntactic structure for input constituents, based on lexical semantic, pragmatic and discourse clues available in the input."

2 Objectives

The objectives behind the development of a machine translation (MT) system between the two languages are largely guided by the sociolinguistic situation.

Following [2], applications of machine translation can be divided in two main groups with different requirements: *assimilation*, that is, to enable a user to understand what the text is about; and *dissemination*, that is, to help in the task of translating a text to be published.

Assimilation may be possible even when the text is far from being grammatically correct; however, for dissemination, the effort needed to correct (*post-edit*) the text must be lower than the effort needed to translate it from scratch.

A majority to minority language system will mainly be used for dissemination purposes, where post-editing the output should be faster than translating from scratch and intelligibility is less important.

In a minority to majority language system on the other hand, intelligibility is the main goal as MT is mainly used for assimilation, for instance, to answer vital questions such as "what are they writing about me in the minority language newspaper?".

The system described in this paper falls outside the usual majority–minority continuum, as both languages can be considered minority languages (one of which again is a minority language in the Sámi context), and the system has a dual focus.

On one hand, it should be able to produce Lule Sámi texts appropriate for post-edition from North Sámi texts, for example to translate educational materials.

On the other hand, it should also be useful for assimilation, to give Lule Sámi speakers the opportunity to follow news in North Sámi (for example from the daily published newspaper *Ávvir*[6]).

3 Technical background

This section gives a brief overview of the two main technologies used in the construction of the prototype system,[7] Apertium,[8] a rule-based machine translation platform, and Constraint Grammar, [3] a rule-based framework for the disambiguation and annotation of text.

3.1 Apertium

The Apertium platform was originally aimed at the Romance languages of the Iberian peninsula, but has also been adapted for other language pairs, such

[6] `http://avvir.no`

[7] The prototype system may be tested online at: `http://victorio.uit.no/cgi-bin/francis/index.php`.

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

as Welsh [4] and Basque [5]. The whole platform, both programs and data, is available from the project website under the GPL licence.[9]

The engine largely follows a shallow-transfer approach to machine translation [6]. Finite-state transducers [7] are used for lexical processing, first-order hidden Markov models (HMM) and optional Constraint Grammar are used for part-of-speech tagging, and finally multi-stage finite-state based chunking is used for structural transfer.

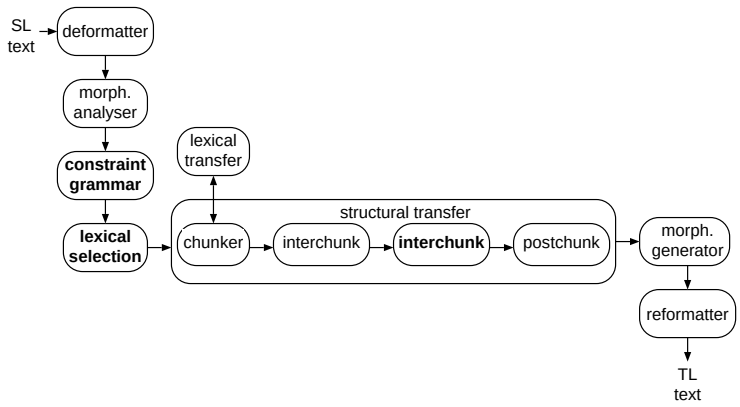

Fig. 1. Modular architecture of the Apertium MT platform. Bold indicates adjustments made for the North Sámi to Lule Sámi pair. The HMM-based part-of-speech tagger has been replaced with a constraint grammar which also provides syntactic labelling, a lexical selection module, also based on a constraint grammar has been inserted, along with an extra *interchunk* (structural transfer) module to deal with longer distance reordering. For example reordering co-ordinated noun-phrase objects, in the first stage of transfer, the co-ordinator and the noun phrases are chunked, then in the first interchunk stage, the two noun phrases are chunked together with the co-ordinator, and finally in the second interchunk stage, they are moved as one unit.

As this paper focuses on the lexical selection aspect, a more detailed description of the pipeline (figure 1) will not be made.

3.2 Constraint grammar

The formalism used for both disambiguation and annotation is Constraint Grammar, which is a linguistically-based approach used for the bottom-up analysis of running text. The Sámi Constraint Grammar performs both morphological and syntactic disambiguation, and annotates syntactic and dependency labels. It uses the VISL-CG3[10] implementation that has been further developed to include support for annotating dependency relations.

[9] `http://www.fsf.org/licensing/licenses/gpl.html`
[10] `http://visl.sdu.dk/constraint_grammar.html`

The lexical selection module is also implemented in Constraint Grammar, and annotates words which are ambiguous in translation in a disambiguated source language sentence with references to their translation in the target language.

This method is inspired by other MT systems including rule-based lexical selection, such as the *Dan2Eng* system [8] which successfully uses 17,000 hand-written lexical transfer rules.

4 Lexical selection

4.1 Potential candidates

The bilingual North Sámi–Lule Sámi lexicon contains many one-to-one relations such as *eadni* → *ieddne* 'mother', and many nouns seem to have fairly straightforward translations.

Upon closer inspection one-to-many, many-to-many, and many-to-one relations are found in the bilingual dictionary. However, many-to-one translations will not be dealt with in this paper as the translation currently is limited to the direction North Sámi→Lule Sámi. Some lexical entries are polysemous in both directions. The North Sámi verb *ráhkadit* 'make' translates into *dahkat* 'make, do' and *stiellit* 'prepare'. Lule Sámi *dahkat* on the other hand translates both into North Sámi *ráhkadit* and *dahkat* 'do'. The current set of one-to-many wordpairs includes verbs, nouns, adjectives and adverbs. Any word with two or more (unrelated) meanings (homonymy or polysemy) is a candidate for lexical selection.

In both languages there is substantial homonymy and polysemy between inflected forms of words, but little between lemmata, although it does exist, for example, *luohkká* 'hill' or 'class, grade' and *giella* 'language, snare, lasso-ring'. In some cases, traditional words which have acquired a modern meaning, e.g. North Sámi *cuozza* 'skin/membrane, transparency', originally only 'membrane'. This polysemy is partly preserved in Lule Sámi (as in *giella*). In other cases, Lule Sámi uses different words for the different senses of the North Sámi noun (as in *luohkká*).

Some words have acquired a metaphorical sense *jámas* 'dead' as in *jámas dolkan* 'dead sick of' – possibly under the influence of the Scandinavian languages, i.e. *dødslei* 'dead-sick.of' in Norwegian.

In other cases, words that can be used in a wider, narrower or otherwise different domain need to have lexical selection rules written, for example the adjective *boaris* 'old' which can only be translated as *boares* for inanimate objects, and otherwise translates as *vuoras*.

4.2 Rules

Within the Constraint Grammar, semantic information is encoded in semantic sets within the lexical selection module. The bilingual lexicon specifies one or more alternative translations, the default labelled with `SO`, and the alternatives

labelled with consecutive numbers from one. The rules make use of morphological, syntactic and semantic information. Rules were inspired by comments by a native speaker of Lule Sámi about incorrect lexical choice in the translations made by the MT system. A number of word-pairs with context-dependent translations were identified. A native speaker of North Sámi with passive knowledge of Lule Sámi was asked to translate a number of Lule Sámi sentences including the different variants so that the contexts for each translation variant could be refined.

The example in (2) shows how PoS/morphological information can be used to create a rule to distinguish the translations of *luohkká* 'hill, class or grade', where the sense *klássa* 'class or grade' is used with a preceeding ordinal. The other translation *luohkka* 'hill' on the other hand is less likely to be the correct one if the word is preceeded by an ordinal.

(2) Sii leat vuosttaš luohkás. (North Sámi)
They are first grade-LOC
'They are in first grade.'

The rule selecting the translation *vuoras* for *boaris* 'old' makes use of the fact that personal pronouns in first and second person usually denote a human. Syntactic information is specifically used with the polysemous verb *orrut* 'stay, seem', which translates into *vuojnnet* before a noun/adjective in essive case[11] or a predicative, as in example (3).

(3) Orru leamen buorre. (North Sámi)
Seem be-ACTIO.ESS good-PRED
'It seems to be good'

The last type of constraints are narrower lexical or even idiosyncratic constructions. The adjective *buorre* 'good' is translated into *jasskat* before particular nouns, such as *ieš dovdu* 'self-confidence'. Example (4) shows an example of this kind of constraint, in the example North Sámi is given on the first line and Lule Sámi on the second line.

(4) ... addin dihte *buori* ieš dovddu. (North Sámi)
... vattátjit *jasska* iesjdåbdov. (Lule Sámi)
'... in giving good self-esteem'

Semantic information is used in a number of rules. The noun *luohkká* 'hill, class' is translated into *klássa* 'class' in a sentence containing members of a set of words related to education. The verb *ráhkadit* 'make, prepare' is translated into *stiellit* in sentences that have a grammatical object from a set of words related to food. The rule translating *boaris* 'old' into *vuoras* makes use of a set generalising over nouns denoting humans, see figure 2. The adverb *jámas* 'dead' translates into *sælldát* in connection with *psych-verbs*,[12] for example *ballat* 'fear' *dolkat* 'be sick

[11] The essive case expresses a temporary state or quality.
[12] Psych-verbs are those verbs which designate a psychological state or process.

of' *suhttat* 'get angry at', where it gets a metaphorical meaning which is not conveyed by the word *jámas* in Lule Sámi.

Lexical selection can be handled by rules that pick a certain sense of a word. This sense is chosen in a certain syntactic or semantic context and then receives a particular translation in the target language.

North Sámi *boaris* 'old' can be translated with both *vuoras* and *boares*. *vuoras* is used only for humans and animals. The set ANIMAL includes nouns that denote animals such as *ealga* 'elk' and *rievssat* 'ptarmigan'. The set HUMAN includes male, female proper nouns, surnames and other nouns denoting humans such as *áddjá* 'grandfather' and *oahpaheaddji* 'teacher'. The rule in 2 selects sense 1 `(S1)` *boares* when the adjective `(A)` has an attributive form `(A Attr)`, and the noun `(N)` it modifies is not in the set of nouns that denote humans or animals `(LINK NOT 0 HUMAN OR ANIMAL)`.

```
SET ANIMAL = "ealga" "rievssat" ...;
SET HUMAN  = (Prop Mal) (Prop Fem) (Prop Sur)
             "áddjá" "oahpaheaddji" ... ;

SUBSTITUTE (A S0) (A S1) ("boaris"ri A Attr)
           (*1 N BARRIER NOT-Attr LINK NOT 0 HUMAN OR ANIMAL);

SUBSTITUTE (IV) (IV S1) ("orrut"ri V) (1 (@←SPRED));
```

Fig. 2. Two constraint grammar lexical selection rules to select between two translations of *boaris* 'old' and *orrut* 'stay, seem'.

The second rule selects sense 1 `(S1)` of the intransitive verb `(IV)` *orrut* 'stay, seem' if there is a subject predicative `(@←SPRED)` one position to the right as in example .

(5) a. Orru buorre. (North Sámi)
Seems good.
'It seems good.'

b. Årru buorre. (Lule Sámi)
Seems good.
'It seems good.'

5 Evaluation

For the evaluation, the North Sámi side of the New Testament was tagged and sentences with the target words were extracted.[13] Both equivalent and non-

[13] The corpus of test sentences may be downloaded from: `http://www.dlsi.ua.es/~ftyers/sme-smj.testsentences.tar.gz`.

equivalent translations were considered, but only equivalent translations were included when calculating the percentage of correct translations.

Equivalent constructions are those where the lexical item is translated by a possible equivalent of the same part-of-speech. Derivations which do not change the lexical category of the word (e.g. Noun → Noun) and compounds are permitted. In some cases it was difficult to decide whether the translation is a possible equivalent or a different word. As is the case with the North Sámi verb *muitalit* 'tell'. Non-equivalent constructions are those where the lemmata in question are not possible lexical equivalents, the syntactic construction differs completely or the lexical equivalent is simply left out, compare the aligned translations in (6) and (7).

(6) Muhto ii son eallán suinna ovttas (North Sámi)
But not he/she live+PP him+COM together
'But she did not live together with him'

(7) Valla ittjij suv duohtada (Lule Sámi)
But not+PRT he/she+ACC touch+CONNEG
'But she did not touch him'

Rather than aiming at a system that translates (6) into (7), the aligned sentence should be discarded in favour of a more literal translation.

Some potential Lule Sámi equivalents have better North Sámi equivalents than *muitalit* 'tell' and it is questionable in how far these can be considered equivalent constructions: Lule Sámi *sárnnot* is usually translated with *dadjat* 'say, tell' in North Sámi and *hållat* with *hupmat* 'talk, speak' or *hállat* 'talk, speak'.

These sentences were run through the rules and the chosen translation was checked against the translation in the Lule Sámi New Testament. The difficulty lies in finding appropriate text for the evaluation. The number of parallel texts for North Sámi and Lule Sámi is very limited. For evaluation the New Testament is used, which exists as parallel text for North and Lule Sámi, but is based on different originals, in different languages. That means that the closest aligned sentences do not neccessarily contain lexical equivalents, possibly not even corresponding syntactic constructions.

A number of other problems can also be predicted: the Biblical language might not catch the newer (metaphorical) senses of a word and in general not reflect the language to which the MT system is targetted and the New Testament might miss out contexts in which word senses could be used for different reasons.

What is more, Lule Sámi does not have a strict norm (as opposed to North Sámi). Lexical dialect borders are fuzzy, which makes it even harder to decide if different word choice is simply due to dialect variation with a use restricted to a dialect context or words that can be distinguished in their use by means of linguistic constraints.

Word	Gloss	Type	Translations	Type	Equiv.	Correct	(%)
jámas	'dead'	Adv.	jámas, sælldát	Sem	2	2	100%
láhkái	'kind, type'	Adv.	láhkáj, muoduk	Synt	2	2	100%
čeahppi	'smart'	Adjec.	smidá, tjiehppe	Lex	2	2	100%
buorre	'good'	Adjec.	buorre, jasskat	Lex	218	192	88%
eallit	'to live'	Verb	viessot, iellet	Lex	147	115	78%
orrut	'to stay, to seem'	Verb	årrot, vuojnnet	Synt	87	50	57%
ráhkadit	'to make'	Verb	dahkat, stiellit	Sem/Synt	33	16	48%
muitalit	'to tell'	Verb	subtsastit, mujttalit	Lex	102	28	27%
boaris	'old'	Adjec.	vuoras, boares	Sem/Synt	31	8	25%
hui	'very'	Adv.	huj, sieldes	Synt	8	0	0%
jaska	'quiet'	Adv.	sjávot, jasska	Lex	0	0	-
luohkká	'class, hill'	Noun	luohkka, klássa	Sem/Cat	0	0	-

Table 1. Evaluation of the lexical selection rules over the New Testament. The first column gives the word in North Sámi, and the fourth column the possible translations into Lule Sámi from the bilingual lexicon with the default underlined. The *Type* column shows the type of rule (lexical, word category, syntactic, semantic). The *Equiv.* column gives the number of equivalent sentences which were found for the word in both the North and Lule Sámi New Testaments, while the *Correct* column gives the number of correct translations produced by the rules.

6 Discussion

As can be seen in table 1, the rules perform best on words where the non-default scope is quite narrow, such as the rule for *buorre* 'good', where the non-default is only picked in some lexical contexts. Bad performance of some of the other rules is due to the selection of the wrong default as e.g. in the case of *boaris* 'old', the existence of several variants that have not been considered (and might be even restricted to a Biblical context), as in *muitalit* 'tell', where *giehttot* and not *subtsastit* or *mujttalit* get most hits, and difficulties in excluding synonymy.

It is also due to the inclusion of various potentially deviating contexts in the total number of equivalent sentences as in the case of *muitalit* 'tell'. Categorising the rules with regard to their linguistic level and complexity, the simple lexical rules (referring to nearly idiosyncratic contexts) are written very quickly and make up the ones performing best (with the exception of the rule for *láhkái* 'kind, type' and *jámas* 'dead', which are hard to evaluate since each of them only occurs twice.). In table 1, most of the syntactico-semantic rules perform worst. The higher the level of abstraction, either in terms of semantic sets or syntactic contexts, the more carefully the rules need to be made to achieve good performance. The higher the level of abstraction is, the better one needs to know positive and negative contexts of the word on the one hand and all the possible equivalents of a word on the other hand.

The evaluation has helped to specify the contexts of lexical selection in some cases. The rule selecting *iellet* as a translation for *eallit* 'live, be alive' originally had a very narrow context, which now can be extended to other contexts than *agálaččat* 'forever'. The verb *iellet* is translated by [9] as *leva (vanligen i andlig betydelse)* 'live/be alive (usually in a spiritual sense)'. Typical and recurring constructions for the religious sense of *eallit* are *ealli čáhci* 'living water', *ealli Ipmil* 'living God', and *ealli sátni* 'living word'. The adjective *vuoras* is typically selected as a translation for *boaris* 'old' after the word *jahki* 'year'.

Secondly, we found additional equivalents for North Sámi words. The verb *muitalit* 'tell' has been translated as *subtsastit*, *mujttalit*, *giehttot*, *sárnnot*, *ságastit*, *javllat*, *hållat*, *diededit*. The verb *ráhkadit* has been translated as *dahkat*, *stiellit*, *tsieggit*, *gárvedit*. The adjective *boaris* 'old' does not only have the translations *boares* and *vuoras*, but also *oames* in contrastive phrases about *ådå* 'new', *varás* 'fresh' and *oames* 'old', in food contexts, and about clothes 'worn out'. Other than in SMT, in RBMT, one (most) suitable translation can be picked generalizing over the variation that can be found in parallel texts.

The improved rule set picks out *oames* if the sentence contains either *ådå* 'new' or *varás* 'fresh', and *boaris* 'old' stands in attributive position to a noun from the semantic set of clothes or food.

The rule for the verb *orrut* 'stay, seem' performs well in selecting the non-default *vuojnnet* 'seem', but often gets translated not only with *årrot* 'stay' or *vuojnnet* 'seem', but simply with *liehket* 'be' as in (8-b).[14] The verb *liehket* has a closer equivalent in North Sámi, *leat* 'be'. The goal of the machine translation system is not to get caught up in possible correct variants. Linguistically on the other hand this finding is interesting, as it could hint at a more frequent use of *liehket* in Lule Sámi than in North Sámi.

(8) a. Ehpet go dieđe [...] ahte Ipmila Vuoiŋŋa orru din siste?
Not ø-qst know [...] that God-gen Spirit stays you inside?
(North Sámi)

'Don't you know that God's Spirit lives in you?'

b. Ehpit gus dádjada [...] Jubmela Vuojŋŋanis dijájn le?
Not ø-qst know [...] God-gen Spirit you-LOC PL stays?
(Lule Sámi)

'Don't you know that God's Spirit is in you?'

In some cases the default needs to be reconsidered. The adjective *boaris* 'old' is translated more frequently as *boares* than as *vuoras* (as originally predicted). A bigger corpus might be needed to test and compare the frequency of non-human vs. human use of the word.

Some of the word pairs seem to be very difficult to distinguish and have rather a synonymy status as even seemingly clear contexts included both variants (*eallit*

[14] 1 Corinthians 3:16

agálaččat both *iellet* and (a few) *viessot*). In other cases it is more important to distinguish (*orrut*, *boaris*).

The New Testament examples showed that even in a seemingly straightforward word pair, the realisation in text can diverge in both directions. This may result in several alternative translations, partly synonymous.

Writing lexical selection rules does not only help to pick the correct equivalent, but also to acquire knowlege about the correct equivalent.

Even though North Sámi and Lule Sámi are closely related languages and a large amount of lexical transfer is fairly straightforward, a number of cases require lexical selection. And the variation found in the New Testament shows that preferences with regard to lexical and syntactic constructions can vary substantially. The reasons for that can lie in individual preferences of the translator, but can also be a general linguistic tendency of the language. Two independently translated texts with different source languages do not provide this information. A lot of native speaker competence and a more bilingual corpus material is needed in further work.

In a non-standardised language like Lule Sámi there is much lexical variation. This makes it difficult to match parallel texts, as texts show no consensus as to which term is the appropriate translation. For each translation one equivalent must be chosen, and here, RBMT may do just that.

Simple and accurate rules, if possible with a high level of abstraction, can improve the output of MT on a lexical level considerably, even between closely-related languages. A rule-based approach to lexical selection also has further benefits where the languages in question are under-studied as it provides an opportunity to do research into lexicography and semantics in both languages.

7 Conclusion

The paper has explored the use of lexical selection in machine translation to improve lexical choice in translation. In the case of such little researched language pairs as North Sámi–Lule Sámi, this does not only include technical work, but also linguistic pioneer work as one cannot base decisions on an existing bilingual dictionary, but rather write parts of the bilingual dictionary oneself. Both linguistic research and technical solutions for lexical selection in machine translation between related under-resourced languages are needed. But of course, lexical selection is necessary to a much larger extent for less related languages. The current approach shows that despite (linguistic) difficulties, the inclusion of a lexical selection module based on Constraint Grammar rules can be included in a fairly simple manner into a rule-based machine translation system such as Apertium.

Acknowledgements

Many thanks to Trond Trosterud, Lene Antonsen and the anonymous reviewers for their helpful comments in improving this paper. This work has also received

the support of the Spanish Ministry of Science and Innovation through project TIN2009-14009-C02-01.

References

1. James Pustejovsky and Sergei Nirenburg. Lexical selection in the process of language generation. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 201–206, Morristown, NJ, USA, 1987. Association for Computational Linguistics.
2. Denis Gachot. Assimilation or dissemination? that is the question. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, Montreal, Quebec, Canada, 1996.
3. Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila. *Constraint Grammar: A language independent system for parsing unrestricted text.* Mouton de Gruyter, 1994.
4. Francis M. Tyers and Kevin Donnelly. apertium-cy: A collaboratively-developed free RBMT system for Welsh to English. *Prague Bulletin of Mathematical Linguistics*, (91):57–66, 2009.
5. Mireia Ginestí-Rosell, Gema Ramírez-Sánchez, Sergio Ortiz-Rojas, Francis M. Tyers, and Mikel L. Forcada. Development of a free Basque to Spanish machine translation system. *Procesamiento del Lenguaje Natural*, 43:187–195, 2009.
6. Mikel L. Forcada, Francis M. Tyers, and Gema Ramírez-Sánchez. The free/open-source machine translation platform Apertium: Five years on. In F.M. Tyers J.A. Pérez-Ortiz, F. Sánchez-Martínez, editor, *Proceedings of theFirst International Workshop on Free/Open-Source Rule-Based Machine Translation FreeRBMT'09*, pages 3–10, November 2009.
7. E. Roche and Y. Schabes. Introduction. In E. Roche and Y. Schabes, editors, *Finite-State Language Processing*, pages 1–65. MIT Press, Cambridge, Mass., 1997.
8. Eckhard Bick. Dan2eng: Wide-Coverage Danish-English Machine Translation. *Proceedings of Machine Translation Summit XI, 10-14. Sept. 2007, Copenhagen*, pages 37–43, 2007.
9. Olavi Korhonen. *Báhkogirjje: julevusámes dárrui dáros julevusábmái.* Jokkmokk, Sámij åhdadusguovdásj, 2007. '*Dictionary: Lulesámi to Norwegian, Norwegian to Lulesámi*'.