An Experiment of Use and Reuse of Verb Valency in Morphosyntactic Disambiguation and Machine Translation for Euskara and North Sámi

Linda Wiechetek
Giellatekno / Romssa universitehta
linda.wiechetek@uit.no

Jose Mari Arriola
IXA / Euskal Herriko Unibertsitatea
josemaria.arriola@ehu.es

1 Introduction

There are a number of well known resources dealing with verb valency including *PropBank* (Palmer et al., 2005), *VerbNet* (Kipper et al., 2006) and *VALLEX* (Hajič et al., 2003). These include thematic roles, morpho-syntactic specifcations and selection preferences. A comparatively wide definition of valency including subcategorization information on all mentioned linguistic levels is applied here. However, these resources have not often been used in rule-based NLP tasks such as machine translation or disambiguation. Bick (2000) uses syntactic verb valency tags specifying e.g. transitivity and selection preferences for various NLP tasks. The use of verb valency is on a high level of grammatical analysis and requires other elaborated linguistic resources. Bick (2000) uses tags specifying transitivity preferences such as "preferably transitive, but potentially intransitive" but also selection preferences, e.g. specifying a human accusative. Agirre et al. (2009) successfully apply valency information, i.e. case subcategorization information, to the Spanish->Euskara MT system *Matxin* in order to improve NP/PP translation. They present different kinds of tests enriching their machine translation system with different techniques. In all cases, the combinations of techniques that include valency information produce the best results especially in recall and F-score.

This paper describes an experiment for the application of verb valency in Euskara and North Sámi rule-based NLP applications, i.e. morpho-syntactic disambiguation and machine translation. 10 frequent verbs each are annotated to improve the analysis, and later the effects on the application is evaluated.

The main objective of the experiment is improving linguistic resources for North Sámi and Euskara taking advantage of pre-existing existing resources in one language and transfering them to the other language. Other works (Antonsen et al., 2010) have shown that the reuse of grammatical resources between both related and unrelated endangered languages is possible and provides useful results, especially on a high level of linguistic analysis. In Antonsen et al. (2010) especially the reuse of the dependency grammar is described.

A number of problems that syntax alone cannot handle can be resolved by semantically richer information included in verb valency. Verb valency annotation is applied on a high level of linguistic analysis and is therefore useful for reuse even for unrelated languages.

2 The experiment

The test cases used in this experiment regard morpho-syntactic disambiguation and machine translation and improve the analysis / translation by making use of valency information.

In many cases, pure syntactic information is not sufficient for the morpho-syntactic disambiguation of nouns, and richer linguistic information is needed. The same counts for machine translation, where morpho-syntactic generation of nouns and polysemy resolution can require high-level linguistic analysis.

The languages in question are lesser-used languages, with 15,000 to 25,000 North Sámi speakers and 775,000 Euskara speakers. North Sámi and Euskara are unrelated languages. Euskara is a language isolate, while North Sámi belongs to the Finno-Ugric language family. One major similarity is their morphological complexity: Euskara is an agglutinative language and North Sámi has both agglutinative and inflective features. They also both have a medium to large sized system of affixed case markers/postpositions. North Sámi has 7 cases (nominative, genitive, accusative, locative, illative, comitative, essive), while Euskara has 17 affixed cases/postpositions (ergative, absolutive, possessive genitive, local genitive, dative, allative, ablative, inessive, destinative, partitive, prolative, instrumental, sociative, motivative, directional and terminative) [1]. In North Sámi, two of the main ambiguities are genitive-accusative and comitative-locative with a significant impact on the F-Score of the analysis.

In Euskara, the homonymy of absolutive plural and ergative singular cause approximately 40% of the ambiguity left after morpho-syntactic disambiguation.

[1] The definition of the terms case/postposition is disputed. In the current terminology only ergative, absolutive and dative are considered cases, while the others are considered affixed postpositions.

(1) Nekez lortu nuen zure ezpainak ikustea.
Hardly achieve do-I-IT.PAST your lip-ERG.SG/ABS.PL see-VNOUN.ABS.SG
'I hardly managed to see your lips.'

In example (1), the ergative singular/absolutive plural ambiguity can be seen in the word *ezpainak 'lips'*. The form can potentially be a subject, object or predicate in absolutive case and a subject in ergative case. In this sentence the object reading holds, therefore absolutive case should be selected. The ergative interpretation can be discarded based on valency requirements of the nominalized verb *ikustea 'seeing'*. The ambiguity is not resolved in the current version of the Euskara analyzer, but can be resolved by similar methods as in North Sámi.

2.1 Technical and linguistic background

The Euskara and Sámi sentence analysis NLP tools built at the University of the Basque country and the University of Tromsø have a similar structure. They contain finite-state transducers for the morphological analysis compiled with the Xerox compilers `twolc` and `lexc` (Beesley and Karttunen, 2003). They can alternatively be compiled with the open-source compilers HFST (Lindén et al., 2009) (North Sámi) and foma (Alegria et al., 2010) (Euskara). For syntactic analysis and morpho-syntactic disambiguation, Constraint Grammar parsers are being used (Karlsson, 2006).

One main difference between the systems is that while for North Sámi, morphological and syntactic modules are strictly separated, for Euskara most of the syntactic tags are annotated in the same module that adds morphological information to the lemmata. The thought behind that was that morphology and syntax are closely related and in a number of cases the syntactic function can be unambiguously mapped to the morphological representation.[2] In North Sámi mapping the tags at the same time would lead to imense overgeneration, as there is a huge amount of homonymy. The syntactic mapping and disambiguation is organized a bit differently / takes different philosophies as their basis. North Sámi has a large mapping section where by means of context specifications, secure syntactic tags are mapped and disambiguated at an early phase. In Euskara, most of the syntax is first introduced with all ambiguity regardless of the context, and later select and remove rules take care of disambiguation. Another difference is that Euskara has several seperate modules that treat different syntactic tasks, in North Sámi, one grammar handles all of the syntactic tag mapping except for explicit dependencies. Some of the modules that are used for Euskara, mainly the chunking module, which is introduced to handle dependencies, do not exist for North Sámi. In North Sámi, recognizing chunks (such as relative clauses etc.) is done implicitly by selecting barriers for phrases and making classes of clause-boundary identifiers.

North Sámi machine translation uses the open-source rule-based machine translation platform Apertium[3] (Forcada et al., 2009). There are existing prototypes for North Sámi Lule Sámi (sme-smj), South Sámi (sme-sma), Finnish (sme-fin), Norwegian (sme-nob) and Euskara (eus-sme). Apertium works with shallow transfer and uses finite-state transducers, hidden Markov models (HMM), Constraint Grammar and finite-state based chunking.

Most of the ambiguity for North Sámi can be resolved by means of sets for verb valency, semantic prototype sets for nouns and linguistic rules that make use of those (Trosterud and Wiechetek, 2007). There are approximately 60 sets that categorize the verbs according to the syntactic cases they subcategorize for and approximately 160 sets of nouns according to their semantic properties. CG morpho-syntactic rules apply this information and rule out either one of the cases (Trosterud and Wiechetek, 2007).

A set specifying the syntactic subcategorization is for example **LOCV** containing verbs like *ballat 'fear'* and *jearrat 'ask'*. It is used in a rule asking for an argument in locative case. A set specifying the semantic subcategorization on the other hand is **PLACE-V** containing verbs such as *čuožžut 'stand'* and *orrut 'live'*. It is used in a rule typically selecting locative instead of comitative case if the argument is a noun denoting a place.

For Euskara, a few general semantic features derived from the machine translation system MATXIN (Mayor et al., 2011) such as **ANIMATE, HUMAN, TIME, MATERIAL, VEHICLE** and **LANGUAGES** are used. North Sámi on the other hand has also more specific sets, such as **EDUCATION** containing words like *skuvla 'school' giellagursa 'language class'* and **PLACE** containing words like *jeaggi 'swamp'*, *luossabáiki 'salmon fishing place'* and *gávpot 'city'*.

For complex NLP tasks, often a systematized way of storing valency information is desirable. Subcategorization information of verbs (and other PoS as well) is more complex than simple semantic categorization of nouns as it includes morphological, syntactic and semantic information, which is related not only to the verb itself but to a number of arguments that are potentially related to the verbs. Multiple dimensions need to be considered when working with valency.

Sets to encode subcategorization information for verbs encode the information in a fairly one-dimensional way. The main disadvantage of the codification of valency information in sets is that

[2] The ergative plural suffix *-ek* is always a subject *@SUBJ*.

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

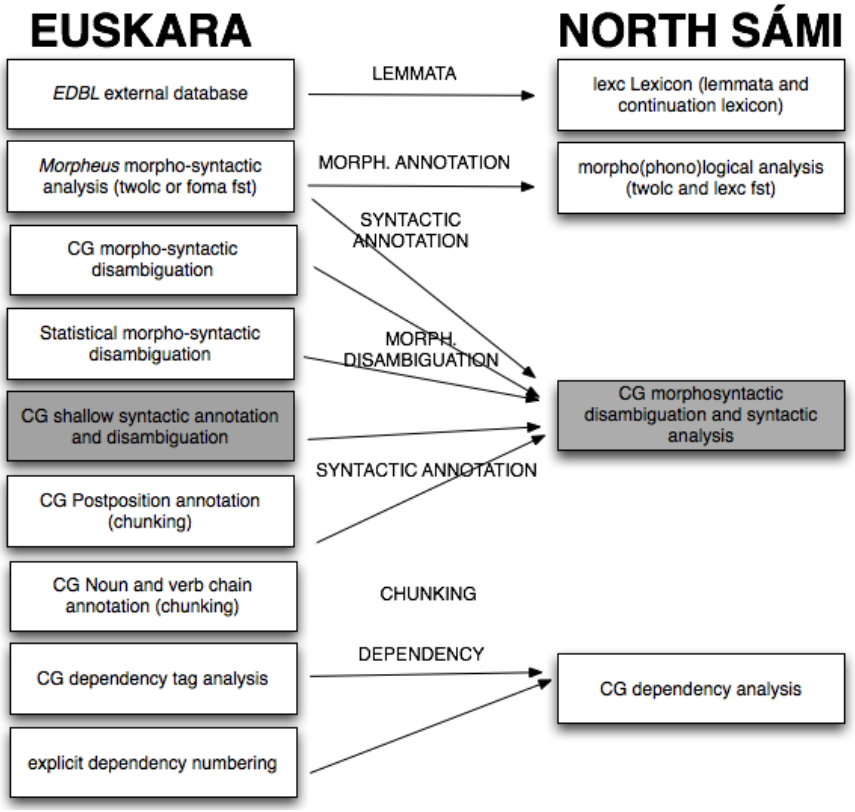

Figure 1: Comparison of the chains for Euskara and North Sámi; the shaded areas mark the places where valency annotation is included

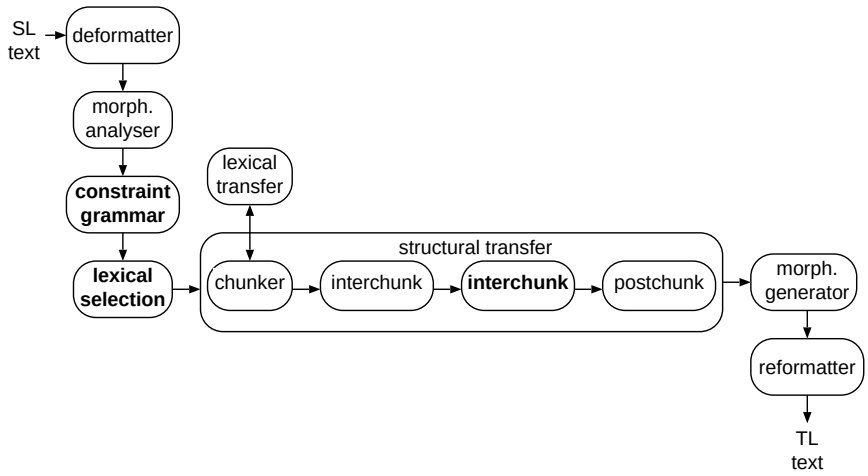

Figure 2: The chain of modules in Apertium

```
<verb lm="hil">
  <frame id="1">
    <ex>Miren hil da.</ex>
    <glosses>
      <gloss lang="eng">Miren has died.</gloss>
      <gloss lang="sme">Miren lea jápmán.</gloss>
    </glosses>
    <theme>
      <case>abs</case>
      <syn>subj</syn>
      <sem>animate</sem>
    </theme>
  </frame>
  ...
```

Figure 3: Verb valency information

the information cannot always be accessed as a whole. In some cases the lemmata in the sets are polysemous, and only one of the meanings is relevant in a certain context. For example, a rule applying the previously named set of place-verbs **PLACE-V** hits for the Sámi verb *orrut '1. stay, live 2. seem*, where it should only hit when the first sense *'stay, live'* is used. If a full valency specification of the verb is available, this problem can be avoided e.g. in tasks as MT as the word senses can be distinguished together with their valencies. Therefore a multi-dimensional representation of valency information is desired.

For Euskara on the other hand, an elaborated database of 100 verbs originally developed for the Euskara PropBank Aldezabal et al. (2010) already exists. It contains rich valency information, i.e. semantic frames including semantic roles and morpho-syntactic information. Here the valency aspect is approached very much from the linguistic side and not so much based on NLP problems (e.g. ambiguity, lexical selection etc.). The challenge of applying the database to NLP tasks lies in the adaptions that are necessary in order to resolve NLP specific problems. As Bick (2000) claims, his categories are made to distinguish meaning, not to define it.

2.2 Annotation of frames

In the experiment, (Aldezabal et al., 2010) verb valency information is converted into valency tags for both Euskara and North Sámi that can be used for disambiguation and machine translation. Code 3 illustrates the way the information is encoded in the database that is meant to be used at a later stage of development.

Each verb can have several frames of argument constellations. The Euskara verb *hil 'die; kill'* for example has two frames, one for the sense *die* and the other for the sense *kill*. In the first sense it has only one argument, in the second it has two.

```
1 hil V Thcase\_Abs Thsyn\_Subj Thsem\_Ani
2 hil V Agcase\_Erg Agsyn\_Subj Pacase\_Abs
        Pasyn\_Obj Pasem\_Ani
```

Arguments are ordered by semantic roles (e.g. agent, theme, topic, patient, location) because they are more unique[4] than syntactic arguments (it is very common to have several adverbials in one sentence). The semantic role level is furthermore perceived as being more abstract and therefore more language-independent, which makes it suitable for reuse for other languages. Arguments have 3 possible attributes: case (or postposition) such as (nominative, accusative, ergative), syntactic function (subject, object, adverbial), and selection restrictions (human, concrete, place). In the case of the verb *hil 'die; kill'*, the first argument, characterized by the semantic role theme, has the three attributes **Thcase_Abs** (absolutive case) **Thsyn_Subj** (syntactic function subject) **Thsem_Ani** (selection restriction animate).

For each verb in Euskara, each frame had to be matched to a verb in North Sámi. In many cases, a lemma in Euskara could not be directly translated to a Sámi verb, the valency frames had to be taken into account to find the correct equivalent(s). When the equivalent in Sámi had been found, the frames were copied to Sámi, and in a second step adaptions were made. While roles in principle stayed the same, cases and to a smaller degree also syntactic functions had to be changed.

For test purposes, we found that the easiest way to annotate valency information was by means of Constraint Grammar rules.

The rules adding the valency tags to the verb *lortu 'achieve, get'* have the following format:

```
ADD (Agcase_Erg Agsyn_Subj Thcase_Abs
   Thsyn_Obj) TARGET (ADI) IF (0 LORTU);
```

This format is sufficient for the annotation of a small amount of verbs for testing purposes. For a large-scale annotation of verbs we would like to include the tags automatically from the verb database in figure 3.

2.3 Disambiguation

Verb valency information is used for syntactic disambiguation. In Euskara, both morphological and syntactic ambiguity exist, i.e. one word recives multiple analyses. Morphological ambiguity in Euskara includes e.g. categorial ambiguity e.g. typically noun/verb ambiguity. For agglutinative languages there are additional sources of ambiguity (number, case, etc.). One of the most pervasive ambiguities is the one related to the suffix *–ak*, it can codify absolutive plural or ergative singular. Additionally the suffix *-a* causes significant ambiguity.

Syntactic ambiguity is added on top of morphological ambiguity. Disambiguation of subject or object functions is needed to detect agreement errors. Concerning the previously mentioned suffix *–ak* the following ambiguity is given: absolutive case can be subject, object or predicative, ergative case on the other hand can only be a subject.

[4] Constellations with e.g. two themes are possible but not that frequent.

This suffix can be attached mainly to nouns and also finite verbs (e.g. *etorri den-ak 'the one who has come'*) or non-finite verbs (e.g. *etortze-ak 'the coming'*), which are converted into subordinate clauses. Here, the ambiguity appears in much more complex contests: finite or non-finite verbs with subject, object or predicative function. The same ambiguity is caused by the suffix *–a*.

In order to improve morpho-syntactic disambiguation, valency information is used to reduce absolutive/ergative ambiguity and syntactic ambiguity. Both morpho-syntactic and syntactic ambiguity are closely related. For that reason, during the first step absolutive or ergative case is selected, and in the second step the correct syntactic tag is chosen. This is done in a second constraint grammar module. This module contains disambiguation rules that make use of the valency. In the case of *lortu 'achieve, get'* in example (1), the ambiguity between the predicative and the object reading of *ezpainak 'lips'* is resolved by means of the valency of *ikusi 'to see'* and the object reading is selected by means of the following rule.

```
SELECT (@OBJ) IF (0 ABS LINK 0 (@OBJ))
  (NOT 0 ERG)(*1 Thcase_Abs BARRIER
  ADI/ADL/ADT LINK 0 Thsyn_Obj );
```

The other ambiguity in the sentence consists in the readings of the the non-finite verbal noun *ikustea 'seeing'*, which can be a subject, an object or a predicate. In order to select the object reading the rule checks if there is a verb, here *lortu 'to achieve'* to its left, that has an object in its valency.

```
SELECT (@-NON-FINITE-VERB_CLAUSE_OBJ)
  IF (0 NON-FINITE-VERB)
  (*-1 Thsyn_Obj BARRIER ADI/ADL/ADT);
```

Even though semantic roles are not explicitly annotated to the verbs' arguments in running text, the CG-rules make use of semantic role information as in *Thsyn_ Obj 'the theme has the syntactic function object'*. In order to annotate semantic roles, Bick and Valverde (2009) uses morphological information (PoS, case etc.), syntactic information (subj etc) and semantic information. In addition, barriers that identify beginning and end of a phrase, are necessary to define the dependency between verbs and their arguments, especially to find long-distance dependencies in case there are relative (subordinate) phrases etc. In general, it can be said that by means of two elements of the "triple" valency, semantic roles and dependencies, the third one can be identified.

Therefore, it makes sense to refer to semantic roles of the arguments of the verbs, even at this stage of the analysis. Furthermore, semantic roles are currently being annotated in the corpus of Euskara, and will be available in the near future.

2.4 Machine Translation

In the case of machine translation, the use of valency is meaningful for two subtasks. In the first case, a default argument realization is corrected to the one that suits the valency requirements/restrictions in the target language. Sociative case in Euskara usually corresponds to comitative case in Sámi (cf. Table 1).

Euskara	**North Sámi**
ergative	nominative
absolutive	accusative, (nominative, essive)
genitive	genitive
inessive	locative
ablative	locative
dative	illative
allative	illative
benefactive	illative
instrumental	comitative
sociative	comitative

Table 1: Default correspondences between a number of relevant cases in Euskara and North Sámi

In some cases, the verb valency in the target language deviates from the default case correspondence as in example (2-a), where the case of the experiencer is illative as assigned by a substitution rule.

(2) a. Zergatik haserretzen zara nirekin?
Why get.angry do.you I.soc.sg

'Why do you get angry with me'

b. Manin don suhtat munnje?
Why you get.angry I.ill.sg
'Why do you get angry with me'

The following substitution rules assign valency within a separate Apertium valency module to the verbs in Euskara and North Sámi.

```
SUBSTITUTE (V) (V Caucase_Erg Caucase_Soc
  Causyn_Subj Expcase_Abs Expsyn_Obj
  Expsem_Hum) ("haserre");

SUBSTITUTE (V) (V Caucase_Nom Causyn_Subj
  Causem_Hum Expcase_Ill Expcase_ala
  Expsyn_Advl Expsem_Hum) ("suhttat");
```

Another substitute rule in a constraint grammar valency module replaces the Euskara valency frame for *haserre* by the North Sámi valency frame and a transfer rule matches the correct case to the Sámi noun based on the case attributes in the valency frame.

As a default, a transfer rule as shown in 5 selects a the most frequent corresponding case, e.g. comitative, for a particular case, here sociative, in Euskara.

The following rule picks a valency-based case, if a verb valency tag asks for it. It sets case to +Acc if Thcase is Thcase_Acc.

```
SUBSTITUTE (%Val Cacase_Erg Cacase_Soc
    Casyn_Subj Excase_Abs Exsyn_Obj Exsem_Hum)
    (%Val Cacase_Ill Cacase_ala Casyn_Advl
    Casem_Hum Excase_Nom Exsyn_Subj Exsem_Hum)
    ("haserre");
```

Figure 4: Substitution rule in the valency module

```
<choose>
  <when>
      <test><equal><clip pos="1"
         side="sl" part="case"/>
         <lit-tag v="Soc"/></equal></test>
  <let><clip pos="1" side="tl" part="case"/>
         <lit-tag v="Com"/></let>
  </when>
</choose>
```

Figure 5: Transfer of default cases

In the other case, i.e. lexical selection, depending on the valency frame of the verb, a specific lexeme is chosen in the target language. The regular case when translating from one language to the other, i.e. from Euskara to North Sámi, is that there is more than one possible translation depending on the context, i.e. in most cases the valency frame of the verb. The verb *hil 'die, kill'* translates into *jápmit 'die'* with only the theme role realized. With an animate patient object, it translates into *goddit 'kill'*.

The lexical selection module helps to pick the correct equivalent. The lexicon specifies the possible lexical variants by means of numbers.

```
hil jápmit (die)
hil:1 goddit (kill)
```

A lexical selection rule picks the non-default reading *goddit 'kill'* if it finds an animate absolutive item to the left of it.

```
SUBSTITUTE ("hil") ("hil:1") ("hil")
  (0 (Pacase_Abs Pasyn_Obj Pasem_Ani)
  LINK *§PA LINK 0 ANIMATE BARRIER FAUX
  OR S-BOUNDARY2);
```

3 Evaluation

3.1 Translation of frames

100 Euskara verbs were translated into 187 North Sámi verbs on a frame-to-frame basis, i.e. a polysemy of at least 1,87 meanings per Euskara verb as can be seen in table 2. Careful lexicography work would of course increase the number of possibilities. Of the 187 translations some doubles were found, e.g. *mannat* 'go' (6x), *boahtit* 'come' (5x), *leat* 'be' (3x), *borrat* 'eat' (4x), *šaddat* 'become' (3x). The 100 verbs have 219 listed frames, 184 of those correspond to frames of North Sámi verbs, 35 do not. The correspondence is based on semantic roles, not on syntactic correspondence or on case correspondence.

```
<choose>
  <when>
    <test><equal><var n="Thcase"/>
        <lit-tag v="Thcase_Acc"/>
        </equal></test>
    <let><clip pos="1" side="tl" part="case"/>
        <lit-tag v="Acc"/></let>
  </when>
</choose>
```

Figure 6: Transfer: valency-based case selection

Of the 35 that do not correspond, there are different types: some of the verbs lexicalize different parts of the argument structure. While in Euskara, *barkatu 'forgive'* and *afaldu 'have dinner'* consist of only one lexical unit, in North Sámi part of the verb is realized as an argument *addit ándagassii 'forgive'* and *borrat eahketbiepmu 'have dinner'* and therefore changes the argument structure quantitatively. Verbs that do not correspond, both quantitatively and qualitatively are motion verbs such as *atera 'go out'*, *etorri 'come'*,*igo 'ascend, rise'*, *iritsi 'arrive'*, *pasatu 'go by'*, *sartu 'enter'* *eraman 'bring'*. While in Euskara, typically both source and destination are defined, in North Sámi only either one belongs to the argument scheme.

	Euskara	**N. Sámi**
verbs	100	187
frames	219	-
- corresponding		184
- not corresponding		35

Table 2: Valency-based polysemy and correspondence between verb frames

Typically, a change in valency also corresponds to another translation equivalent. In some cases, all frames of one verb in the source language are translated with one verb in the target language, as is the case for the verb *elkartu 'meet'*, which translates into *deaivvadit*. In other cases polysemy is not related to a distinction in frames. The verb *jo* in its sense 'hit' for example translates into *časkit* if the agent is a human. If the agent is a e.g. a horse, it translates into *nordadit*. Here semantic selection restrictions are necessary for a lexical selection. But in general, semantic role based valency seems to be very useful for a basic sense distinction and lexical selection in machine translation.

3.2 Disambiguation

10 of the most frequent verbs for disambiguation in Euskara were tested and evaluated. The test corpus contains 177 verbs, the verbs evaluated in the experiment represent 5,6% of the verbs of the sample.

The valency frames of 10 verbs were annotated by means of 48 mapping rules. The grammar contains 47 disambiguation rules that resolve absolutive / ergative, absolutive sg./ absolutive indefinite, object, subject and predicative ambiguity for Euskara. The rules can refer to valency tags rather

than the verb lemma and therefore apply to any verb with the characterisitics that appear in the context specification of the rule. The testcorpus is running text and includes sentences without the verbs that have been annotated. This leads to low coverage on the one hand, but takes into account the general impact of the annotation with respect to the frequency of the selected verbs on the other hand. The precision and recall for the rules involved in the disambiguation of the absolutive-ergative syncretism case are 72% and 72% respectively. While the overall analysis improves by 5.5 %, the figures for the ambiguity resolution aimed at are higher. As can be seen in table 3, abs.pl. - erg.sg. ambiguity resolution improves by 18 %, and abs.sg. - abs. indef. ambiguity by 24 %.

precision	72 %
recall	72 %
disambiguation of	
... abs.pl. - erg.sg.	18 %
... abs.sg. - abs. indef.	24 %
... abs./erg. - abs. sg./indef.	46.4 %
improvement for	
... overall analysis	26.9 %
... abs./erg. - abs. sg./indef.	46.4 %

Table 3: Evaluation of morpho-syntactic disambiguation for Euskara

In 13 of 83 cases, the subcategorization information for the verbs is missing completely, in the remaining 19 cases the existing rules do not manage to disambiguate correctly. Wrong applications of rules are mainly due to the occurrence of several verbs with different valencies in one sentence and scope mistakes of the rules and low coverage of semantically annotated nouns.

In 13 of those 19 erroneously applied rules, the rule disambiguates based on the valency information of an unrelated verb, in the 6 remaining cases the semantic information of the nouns is missing. In order to improve the results generalising and extending the subcategorization information to more verbs, refining the disambiguation rules based on verb subcategorization and finally improving the semantic noun sets to meet the lexical selection restrictions of the verbs will be necessary.

3.3 Machine Translation

Valency information has been used for two distinctive tasks in machine translation, syntactic transfer (e.g. picking the correct morpho-syntactic realization of the arguments) and lexical selection (picking the correct lexical equivalent in the target language). Since the basic free resources that are necessary for a complete analysis are not available, they cannot be included in the open-source Apertium machine translation system and only the lexical selection rules have been evaluated. 10 verbs have been annotated and 29 rules refering to valency information have been made for lexical selection. Test sentences to evaluate the lexical selection rules are taken from a newspaper corpus of Euskara. This evaluation will be aimed at improving the existing hand-written rules. As such it will be qualitative not quantitative. A full quantitative evaluation is not possible as the non-availability of existing grammatical resources prevents automatic analysis. The evaluation is focussed on explaining why in some cases where the rules do not apply. Usually this is not because it is impossible to formulate a rule for a given context, but rather that a linguist is not able to forsee all possible contexts without real-life sentences and extensive corpus analysis.

The lexical selection rules are built in the following manner: They refer to a possible right and a possible left context with a semantic role often linked to a case or syntactic function. The context is restrained by a barrier taking into consideration possible markers of borders of clauses such as other finite verbs, punctuation and subordinators. It is obvious that these rules could easily be too simple and that their constraints may have to be modified. With a dependency annotation of the the relations in the sentence between the arguments and the verb would be explicit and barriers would not be necessary.

Rules for lexical selection that refer to quantitative valency differences (differentiating between translation equivalents by means of the number of arguments) as in the case of *hil 'die; kill'* seems to be pretty straightforward. The only difference is that one has only a theme, while the other has an agent and patient. In case of a missing agent, the *jápmit `die'* reading could be selected. The difficulty is that in Euskara the agent does not have to be explicit. Both subject, object and indirect object can be dropped. The auxiliary on the other hand is explicit about the number of grammatical arguments, if the agent is missing another form of the auxiliary is being used. But the auxiliary can be missing too, either when it has the form of a nominalization as in *hiltzea '(the) dying/killing'* or when preceeding a postposition as in *29 lagun hil ondoren 'after 29 people had died' or 'after they had killed 29 people'*.

When the decisive differences for lexical selection are qualitative rather than quantiative, e.g. for *asmatu 'guess, invent, think'* which can be translated as *árvidit 'guess'* or *fuomášit 'invent, come up with, think'* subject/object drop can become a problem. If it is translated as *árvidit 'guess'* it has a theme role while *fuomášit 'invent, come up with, think'* can have a product role. Furthermore it needs to be taken into account that semantic roles can also be carried by clauses such as *itua bete betean asmatzen zutenak 'the ones that guessed the aim exactly'*, where the auxiliary *zutenak 'the ones that did'* carries the semantic role. It makes therefore more sense if rules refer to syntactic functions

rather than morphological cases as carrier of semantic roles.

With regard to barriers, it is important to take into account how far the dependencies of a verb span. In some cases the valency spans far (3), in others they do not (4).

(3) - ...zer-nolako harrera egin-go zion asmatzera jarri zen
what-how welcome make-FUT do-PAST.SUBJ.3.SG.OBJ.3.SG.IOBJ.3.SG think bring be-PAST.SUBJ.3.SG
she/he started thinking what kind of welcome he/she would make her

Here a whole clause *zer-nolako harrera egingo zion* is the argument of the nominalized verb asmatu, and another finite verb *zion 'she/he did to her/him'* is its argument instead of being a barrier.

In the following case on the other hand, the subclause marker '-ela' tells that the arguments of asmatu cannot be outside the subclause and the following auxiliary and main verb are barriers to the span of potential arguments.

(4) Ez duzu-la asmatu esan-go dizute, baina badaezpada galdetu egiten du aurretik.
Not do-SUBJ.2.SG.OBJ.3.SG.-SUBCLAUSE guess tell-FUT do-SUBJ.3.PL.OBJ.3.SG.IOBJ.2.SG., but just in case he/she ask him/her beforehand
They will tell you that you have not guessed it, but just in case he/she ask him/her beforehand

Rules can therefore be improved by taking into consideration possible differences in restricting contexts when nominalizations are being used or auxiliaries are missing. Without a dependency annotation of the text, barriers need to be carefully chosen and take into account possible subclauses and clausal arguments of (nominalized) verbs, and they need to distinguish between the two contexts.

4 Conclusion and future work

The experiment has shown that high-level grammar resources encoding deep linguistic analysis such as verb valency information can be reused even for unrelated languages (such as Euskara and North Sámi) and do not need to be built from scratch. Even though language specific adaptions with regard to syntax and morphology need to be made, semantic role specifications can mostly be transferred without changes. Verb valency information is necessary for both linguistically based disambiguation and machine translation tasks.

North Sámi constraint grammar disambiguation rules that make reference to valency information and semantic sets served as a model for developing Euskara disambiguation rules. Grammar rules based on valency frames provide an efficient way to reduce syntactic ambiguity as they manage to select the correct syntactic function in cases where the syntactic context itself remains ambiguous, but the argument specifications of the verb resolve this ambiguity. In machine translation on the other hand, syntactic transfer involving valency-dependent case realizations of the verb's argument can be accomplished by means of linguistic rules that have access to valency information. Additionally, we have seen that polysemy is frequently related to a distinction in valency, which is why valency information has a key function in picking out the correct argument realization and selecting the correct lexical variant in the target language.

Developing parallel resources for two distinctive and unrelated resources does not only benefit NLP, we gain insights in contrastive grammar of understudied languages in general, and the work can serve as a model for the devolopment of linguistic resources for other languages.

Future plans involve extending both the resources and linguistic rules for disambiguation and machine translation. We want to annotate more verbs with valency specifications, which existing general rules apply to, and evaluate the results and improvements. Automatic dependency annotation and semantic role labelling are currently under development and will not only serve the development of grammar rules including valencies, but also benefit from it. Inducing valencies automatically and thereby extending valency resources is another future task.

5 Acknowledgements

The research of this project has been supported by the Department of Education, Universities and Research of the Basque Government (IT344-10) and University (UPV/EHU) (GIU09/19), Giellatekno (Sámi language technology) at the University in Tromsø and the NILS mobility project (Universidad Complutense de Madrid). We would also like to thank Francis Tyers for his helpful critical remarks and corrections.

References

Agirre, E., A. Atutxa, G. Labaka, M. Lersundi, A. Mayor and K. Sarasola (2009), Use of rich linguistic information to translate prepositions and grammar cases to basque, *in* L.Màrquez and H.Somers, eds, 'BEST PAPER AWARD of the XIII Conference of the European Association for Machine Translation EAMT 2009', Barcelona, pp. 58–65.

Aldezabal, Izaskun, María Jesús Aranzabe, Arantza Díaz de Illarraza, Ainara Estarrona and Larraitz Uria (2010), 'Euspropbank: Integrating semantic information in the basque dependency treebank', *Computational Linguistics and Intelligent Text Processing* pp. 60–73.

Alegria, I., I. Etxeberria, M. Hulden and M. Maritxalar (2010), 'Porting basque morphological grammars to foma, an open-source tool', *Finite-State Methods and Natural Language Processing Lecture Notes in Computer Science* **6062**, 105–113.

Antonsen, Lene, Linda Wiechetek and Trond Trosterud (2010), Reusing grammatical resources for new languages, *in* 'Proceedings of the International conference on Language Resources and Evaluation LREC 2010', The Association for Computational Linguistics, Stroudsburg, pp. 2782–2789.

Beesley, Kenneth R. and Lauri Karttunen (2003), *Finite State Morphology*, CSLI publications in Computational Linguistics, USA.

Bick, E. (2000), *The Parsing System 'Palavras': Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, Aarhus University Press, Aarhus.

Bick, Eckhard and Pilar Valverde (2009), Automatic semantic role annotation for spanish, *in* 'Proceedings of NODALIDA 2009', Vol. 4 of *NEALT Proceedings Series*, Tartu University Library, Tartu, pp. 215–218.

Forcada, Mikel L., Francis M. Tyers and Gema Ramírez-Sánchez (2009), The free/open-source machine translation platform Apertium: Five years on, *in* F. T.J.A. Pérez-Ortiz, F. Sánchez-Martínez, ed., 'Proceedings of theFirst International Workshop on Free/Open-Source Rule-Based Machine Translation FreeRBMT'09', pp. 3–10.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová and Petr Pajas (2003), Pdt-vallex: Creating a large-coverage valency lexicon for treebank annotation, *in* 'In: Proceedings of The Second Workshop on Treebanks and Linguistic Theories', Vaxjo University Press, pp. 57–68.

Karlsson, Fred (2006), *Constraint Grammar - A Language-Independent System for Parsing Unrestricted Text*, Mouton de Gruyter, Berlin.

Kipper, Karin, Anna Korhonen, Neville Ryant and Martha Palmer (2006), Extending verbnet with novel classes, *in* 'Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy'.

Lindén, K., M. Silfverberg and T. Pirinen (2009), Hfst tools for morphology—an efficient open-source pacakge for construction of morphological analyzers, *in* 'Proceedings of Workshop on Systems and Frameworks for Computational Morphology', Zürich, Switzerland.

Mayor, Aingeru, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi and Kepa Sarasola (2011), 'Matxin, an open-source rule-based machine translation system for basque', *Machine Translation Journal (to appear)* .

Palmer, Martha, Paul Kingsbury and Daniel Gildea (2005), 'The proposition bank: An annotated corpus of semantic roles', *Computational Linguistics* **31**.

Trosterud, T. and L. Wiechetek (2007), 'Disambiguering av homonymi i Nord- og Lulesamisk', *Suomalais-Ugrilaisen Seuran Toimituksia = Mémoires de la Société Finno-Ougrienne. Sámit, sánit, sátnehámit. Riepmočála Pekka Sammallahtii miessemánu 21. beaivve 2007* **253**, 375–395.