An FST grammar for verb chain transfer in a Spanish-Basque MT System

Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, Aingeru Mayor and Kepa Sarasola

IXA Group, University of the Basque Country

Abstract. We are working on the construction of a Spanish-Basque (es-eu) machine translation tool using the traditional transfer model based on shallow and dependency parsing. Spanish and Basque languages are very different and one of the main problems is the translation of the verb chain. We will present the features of each language verb chain, and the design of the verb chain structural transfer. We have decided to use FSTs for the transfer rules, and the design and the realization of this module are explained on this paper .

1 Introduction

This paper presents the current status of development of an FST (Finite State Transducer) grammar we are developing in a project of MT (Machine Translation) between Spanish and Basque using the traditional transfer model and based on shallow and dependency parsing. The project is based on the previous work of our group [9] but integrated in OpenTrad initiative, a larger government-funded project which will also include MT engines for translation among main languages in Spain, shared among different universities and companies [8]. The main objective is the construction of an open, reusable and interoperable framework. The whole system will be released at the beginning of 2006.

The MT architecture proposed for the Romance languages uses finite-state transducers for lexical processing, hidden Markov models for part-of-speech tagging, and finite-state based chunking for structural transfer, and is largely based upon that of systems already developed by the Transducens group such as Inter-NOSTRUM (Spanish-Catalan, [6]) and Traductor Universia (Spanish-Portuguese, [10]); these systems are publicly accessible through the net[1] and used on a daily basis by thousands of users.

The design of the architecture for the Spanish-Basque pair (es-eu) has differences due to the high distance between both languages, but we try to reuse modules and interchange-formats used for more similar languages. In the design of the verb chain structural transfer the possibility of using FST rules was studied, and the design and realization of this module is explained in this paper.

The following sections give an overview of similar works and the architecture of the MT system (sec. 2), the structure of the verb chain in Spanish and

[1] http://www.internostrum.com; http://copacabana.dlsi.ua.es/es/index.php

Basque and the translation between them (sec. 3), and the FST grammar we have developed (sec. 4). Finally, the concluding remarks and the planned future work are pointed (sec. 5).

2 Architecture of the MT system and related works

2.1 General architecture

The es-eu MT engine is a classic transfer based system composed by three main modules: analysis of the Spanish text, transference, and generation of the Basque target text. In order to ensure the modularity and interoperation among modules, interchange-formats based on XML have been proposed in the output of each module.

The analyzer of Spanish text is a free available analyzer, FreeLing [7], basically a shallow-parser, which has been augmented with a dependency parser. This module links the dependences among tokens in the chunk, and among chunks in the sentence. The output is an XML structure where the main elements are the chunks in the sentence, and nodes in the chunks.

The transfer and generation phases work in three levels: lexical form or node, chunk and sentence.

In the transfer process, first, lexical transfer is carried out using a bilingual dictionary compiled into a finite-state transducer. Then, structural transfer at sentence level is applied, and some information is transferred from some chunks to others and some chunks can disappear. Finally the structural transfer at chunk level is carried out. This process is quite complex for verb chains and it is the main subject of this paper.

A solution using an FST grammar was designed and tested. In the actual implementation we are using XRCE Finite States Tools [5], but in the near future it will be integrated with the open software developed in OpenTrad.

The syntactic generation process at sentence level reorder the chunks using a Context Free Grammar. At chunk level it uses a simple hash table to map the change in the order of its nodes; additionally, it transfers morphological information to the last node, which has to agglutinate all the morphological information of the chunk. Finally, an FST based morphological generator [2] will generate the surface representation of each word.

No semantic disambiguation is applied, but a large number of co-occurrences, entities and terms will be included in the bilingual dictionary in order to minimize this lack.

2.2 Related work

The success of the use of Finite State Transducers in NLP tasks has extended their application to different areas including syntax and machine translation.

There are several well known papers about FST approaches to parsing. Approaches from Abney [1], Roche [13] and Oflazer [12] have been main references

to our work. There are similarities between Oflazer's work [12] and our proposal. Both work with dependency trees, the rules for establishing dependency links have a similar flavour, and the underlying tool they use (XFST) is the same.

The use of FST for specific tasks in MT is not very usual but there are some substantial contributions [14], [3], [4]. Most of them use the engine in order to identify corresponding syntatic units in bilingual corpora. The AT&T group [3], [4] uses Weighted Transducers, but always in a statistic or example-based framework.

Application of FST for transliteration of named entities is frequent too.

3 Spanish and Basque verb features and their translation

We want to translate Spanish verb chains into Basque. We will focus on the translation of Spanish non-finite forms (21%), indicative forms (65%) and periphrases (6%). This means that we will cover 92% of all possible cases. The rest (8%), mainly subjunctive forms, will be treated in the near future. This section describes the features of Spanish and Basque verb forms, and then, how we do the translation from Spanish into Basque.

3.1 Spanish verb

Spanish finite verbs are composed by several morphemes, and these morphemes give information about voice, mood, aspect, tense, person and number.

Mood expresses the attitude of the speaker toward the action/state of the verb. The Spanish language has finite or non-finite forms depending on the moods. Indicative, subjunctive, and imperative moods present finite forms making distinction among grammatical persons. Infinitive, participle and gerund moods have a single form for all grammatical persons.

Moods and tenses work together to give verbs a precise meaning. When a verb is finite, first the appropriate mood is chosen and then a tense is added to it. Spanish verb conjugations can be divided into two categories: 'simple tenses' and 'compound tenses'. Simple tenses have only one part (*(yo) como / (I) eat*), and in the indicative there are five different tenses: present, imperfect, past tense, future and conditional. Compound perfect tenses have two parts: the finite auxiliary verb *haber (to have)* and the past participle (*comido / eaten*): *(Yo) he comido / (I) have eaten.*

Spanish verb periphrases. A small set of lexical verbs have secondary specialized uses when combined with tenseless verbal forms to form periphrases. Semantically most of these verbs are delexicalized and they function as aspectual modifiers of the action expressed by the formally subordinate tenseless verb.

Usually we find in periphrases the following elements that work all together as a unit: a finite auxiliary verb, an optional particle, and the main verb (non-finite form, infinitive or gerund, giving the meaning): *Tengo que ir ((I) have to go), voy a vender(lo) ((I) am going to sell (it)), debes trabajar ((you) must work), ando buscando ((I) am looking for).*

3.2 Basque verb

Finite verbs, in Basque, can be synthetic, consisting of a single word (*noa / (I) am going, dakit / (I) know it*) or analytical, consisting of a participial form and an auxiliary (*joaten naiz / (I) go, jakingo dut / (I) will know*).

The structure of finite forms (synthetic and auxiliary verb) in Basque is rather complex. The main affixes that appear in finite forms are: agreement markers, pluralizers, dative marker and past/subjunctive suffix. Depending on the transitivity or intransitivity of the verb the sequence of affixes that appear on the synthetic form or in the auxiliary is different. The simplest forms of intransitive verbs are monovalent and only mark agreement with the subject. They can also have bivalent forms adding a dative argument. Finite transitive verb forms are minimally bivalent, marking agreement with an ergative argument (the subject of the clause) and an absolutive argument (the direct object). In addition, there are trivalent forms that add agreement with a dative argument.

Basque language has two auxiliaries for the indicative mood: an intransitive auxiliary, *izan* which as independent verb has the value of *to be*; and a transitive auxiliary, **edun*, which by itself has the value of *to have*.

In Basque, subordination may be marked with a final suffix in the finite verb form. For example, the sentence *jakingo dut ((I) will know)* becomes *jakingo dud**alako** (**because** I will know)* when it is a causal subordinate sentence.

Synthetic verbs. These verb forms have only one part, and our system translates all the nine main synthetic verbs: *izan (to be), ukan (to have), eduki (to have), egon (to be), etorri (to come), joan (to go), eraman (to carry), ekarri (to bring), and jakin (to know).*

Analytical verbs. Most Basque verb expressions are analytical. Analytical verbs expressions contain a participle and a finite auxiliary. Participles carry information about meaning, aspect and tense, whereas auxiliaries convey information about argument structure, tense and mood.

Four participial forms are used: the perfective participle (*erosi / bought*), the imperfective participle (*erosten*), the prospective or future participle (*erosiko*), and the verb radical (*eros / buy*). These nonfinite forms are combined with finite auxiliaries (transitive or intransitive) to give rise to the different basic tenses, aspects and moods.

Additionally some nominalizations are possible. A verbal noun is formed with the suffix *-tze/-te* attached to the radical. The form in *-tze/-te* has the morphological and syntactic properties of a singular noun and can be inflected as such (*auto bat eros**tea** komenigarria da / buy**ing** a car is helpful*).

It can appear more suffixes around the verb form to determine its aspect and its tense. Two of them are: *-tzear/-tear*, which means *about to* (*katua iris**tear** dago / the cat is about to arrive*), and *-ta*, which is added to the perfective participle in predicatives and adverbial expressions (*leihoa apurtu**ta** dago / the window is broken*).

3.3 Spanish-Basque verb chain translation

Depending on the Spanish form of the verb its translation into Basque should be obtained in a different way.

Spanish non-finite forms. We translate the Spanish infinitive with a verbal noun (*-tze/-te*)[2], the gerund with a locative verbal noun (*-tzen/-ten*), and the participle with the perfective participle and a modal suffix (*-tuta*).

Spanish simple tense verbs. A Spanish simple tense verb can be translated into Basque as synthetic or as analytical depending on the verb and its tense.

If a verb is translated as analytical, the main verb will get one of the four participle forms. For example, if we have to translate *(yo) como (manzanas) / (I) eat (apples)*, we will translate it with the imperfective participle form (*jaten*) of the verb *jan (to eat)*, and a transitive auxiliary (*ditut*) which has the information about the agreement with the subject (*I*: 1st sg.) and with the object (*apples*: 3rd pl.).

Spanish complex tense verbs. In the translation of some Spanish complex tenses, we need a new element, which is a dummy auxiliary (*izan*) and its aspect. For example, we need the perfective form of *izan* (*izango*) to translate the Spanish future perfect *habré comido (manzanas) / (I) will have eaten (apples) / (sagarrak) janak* ***izango*** *ditut.*

Spanish infinitive and gerund periphrases. We will try to translate most of the periphrases known by the Spanish analyser.

Spanish periphrases and their Basque translations have a different word order, as we will explain with this example: *habré terminado de comer / bazkaltzen amaitu izango dut ((I) will have finnished eating).*

In this Spanish periphrasis, we have an auxiliary (*habré*), a finite auxiliary (*terminado*) a link element (*de*), and the infinitive form of the main verb (*comer*).

In Basque we first have the main verb (*bazkaldu*) with an imperfective aspect (*bazkaltzen*); after, we have the translation of the finite auxiliary (*amaitu*), which also has to carry the perfective aspect (*amaitu*). In some cases, as in this example, the third element will be a dummy verb (*izan*) with a prospective aspect (*izango*). All aspects depend on the Spanish finite verb tense. In the last position, we have another auxiliary verb (*dut*) which depends on the transitive or intransitive feature of the main verb in some cases, and in other cases of the auxiliary verb.

To perform the translation we focus on the Basque syntactic behaviour of these periphrases, and we join them in four sets.

[2] There are some exceptions, as infinitive forms following perception verbs. In this case the infinitive is translated with a locative verbal noun (*-tzen/-ten*): *(te) veo* ***comer / jaten*** *ikusten zaitut / (I) see (you)* ***eating***. We are studying them to be integrated in next versions.

Most of infinitive periphrases, which have been joined in set 1 (P1), have the schema we have just explained above. In Basque there are some verbs, which work as semiauxiliaries like in the case of Spanish *deber, tener que, haber de (must, have to)* where the correspondence in Basque is the semiauxiliary *behar*. In the same set there is a Spanish verb *querer (to want)* which is not periphrastic, but we have to work with it as we do with *deber*, because both are translated using semiauxiliaries into Basque, and both take transitive auxiliary. For example *(artikulu horiek) irakurri nahi/behar ditut (I want/need to read (those articles))*.

In this set we have also joined all gerund periphrases because the schema is suitable to translate them.

Other set of verbs, set 2 (P2), do not have a periphrastic form or a semiauxiliary verb in Basque, but a modal particle. This particle is attached immediately to the left of the tensed verbal form (whether auxiliary or synthetic verb) to modulate and validate the information conveyed by the clause.

In the third set (P3) there is only a verb (*volver a*) in which translation there is not a particle immediately on the left of the tensed verbal form, but an adverb just on the left of the participle form.

The last set (P4) includes periphrastic verbs which are translated as non-periphrastic.

4 The FST grammar

The FST grammar for verb chains we present here takes as input the morphological information of the nodes of the Spanish verb chain, the Basque form corresponding to the Spanish main verb of the chain, agreement information about the objects (absolutive and dative) and the type of subordination of the sentence. Its output is the list of the nodes of the corresponding Basque verb chain, each one with the information necessary to decide the order of the words, and to realize the morphological generation.

4.1 Rules

The grammar contains three kinds of rules: identification of types, replacing of attributes and cleaning of output.

Identification and markup rules. These rules identify the type of the Spanish verb chain, and add a different schema for the Basque verb chain depending on the type. They use the left-to-right, longest match markup operator and the format is the following:

```
[ esVerbChainType @-> ... BORDER euVerbChainSchema ]
```

There are six rules of this kind, corresponding to the different verb types, that adds one of the five different Basque verb schemas: one for the non-finite forms (NF), one for the non-periphrastic (`NP`) verbs and the periphrasis of the set 4 (`P4`), and one for each of the other three periphrasis set (`P1,P2,P3`):

```
NF>    (main) Aspm
NP/P4> (main) Aspm           /Dum Aspd /Aux TMood SubjObjDat +RelM
P1>    (main) Aspm /Per Aspp /Dum Aspd /Aux TMood SubjObjDat +RelM
P2>    (main) Aspm /Partic   /Dum Aspd /Aux TMood SubjObjDat +RelM
P3>    (main) Aspm /Adverb   /Dum Aspd /Aux TMood SubjObjDat +RelM
```

Attributes replacement rules. These rules replace attributes in the Basque schema with their corresponding values, depending on the values of some attributes in the Spanish verb chain and/or in the Basque schema, which are separated by a BORDER tag. These rules use the left-to-right, longest match conditional replacement operator.

```
[ "euAttribute" @-> "euValue"
                    || ?* esValues ?* BORDER ?* euValues ?* _ ]
```

Depending on the possible contexts there are 18 possible replacement rules for the aspect of the main (`main`)verb (`Aspm`), 18 rules for the periphastic verbs (`Per`) and its aspects (`Aspp`), for particles (`Partic`) and for adverbs (`Adverb`) attributes, 15 rules for the dummy auxiliary (`Dum`) and its aspect (`Aspd`), 8 rules for the last auxiliary (`Aux`) and its tense and mood (`TMood`), 20 rules for the information about subject, object and dative(`SubjObjDat`), and 5 rules for the relation morpheme (`RelM`)

The grammar composes (`A .o. B`) the possible replacement rules for any attribute in different contexts, instead of using the most adequate union operator (`A | B`), because doing the union the grammar generates too big transducers when it is compiled it.

Cleaning rules. Finally two rules remove the unnecessary information, giving the desired output.

4.2 Example

We will illustrate the process using the following example: *porque no habré tenido que comer patatas (because I won't have to eat potatoes).*

The system analyses the Spanish sentence using Freeling, returning this information[3] (each line shows a chunk):

```
subordinate_conjunction: porque[cs]
negative: no[rn]
verb_chain: haber[vaif1s]+tener[vmpp0sm]+que[cs]+comer[vmn]
noun_chain: patatas[ncfp]
```

The input of the verb chain FST transfer module, needs more than the morphological information of the verb chain provided by Freeling in the third chunck. In the lexical transfer phase the system obtains the translation of the main verb (*comer/to eat* is translated by *jan*), and gets the information about transitivity of this verb, (in this case it is transitive [tr]). During the sentence level structural

[3] Freeling uses EAGLES tags: http://www.lsi.upc.es/ nlp/freeling/parole-es.html

transfer, the system gets the number and person information of the absolutive object (*potatoes*: third plural [3p]), and the type of subordination (*because* [cause]), from the other chunks in the sentence.

All this information is the input for the sequence of transducers that will perform the transfer of the verb chain:

```
haber[vaif1s]+tener[vmpp]+que[cs]+comer[vmn]/[tr][3p][caus]/jan
```

The first rule identifies the input of a Spanish verb chain that has a periphrastic of type 1 (`esVerbChainPerif1`), and adds the schema for the Basque verb for this type (`euVerbChainSchemaP1`):

```
[ esVerbChainTypePerif1 @->... BORDER euVerbChainSchemaP1 ]
```

where the symbol `esVerbChainPerif1` recognizes in the input verb chains with periphrastic verbs of type 1, and gives this output:

```
haber[vaif1s]+tener[vmpp]+que[cs]+comer[vmn]/[tr][3p][caus]/jan
==>P1> (main)Aspm/Per Aspp/Dum Aspd/Aux TenseM SubjObjDat +RelM
```

The next rules replace one by one the attributes of the Basque verb schema. For instance, the rules that replace the aspect of the periphrastic verb (Aspp) in the periphrasis of set1 are the following:

```
[     [ "Aspp" @-> "[partFut]" || ?* [VMIF|VMIC|VAIC] ?*
                                                BORDER "P1" ?* _ ]
  .o. [ "Aspp" @-> "[partImp]" || ?* [VMIP|VMII] ?*
             BORDER "P1" ?* [{hasi}|{amaitu}|{utzi}|{joan} ] _ ]
  .o. [ "Aspp" @-> "[verbRad]" || ?* BORDER "P1" ?* {ari} ?* _ ]
  .o. [ "Aspp" @-> "[partPerf]" || ?* BORDER "P1" ?* _ ]
] ;
```

The arc labelled with "`Aspp`" is replaced by "`[partFut]`" (participle future), when the tense of the Spanish main verb is future indicative (`VMIF`) or condicional (`VMIF`) or the tense of auxiliary verb is conditional (`VAIC`); however "`Aspp`" is replaced by "`[partImp]`" (participle imperfective) when the tense of the Spanish main verb is present indicative (`VMIP`) or imperfect (`VMII`) and the Basque periphrastic verb is one of *hasi, amaitu, utzi, joan*, and ; otherwise "`Aspp`" is replaced by "`verbRad`" (verb radical) when the Basque periphrastic verb is *ari*; and in any other case "`Aspp`" is replaced by "`partPerf`" (participle perfective).

Table 1 shows all the replacements for the example, and the context that constraints them.

The output after all these replacements is:

```
haber[vaif1s]+tener[vmpp]+que[cs]+comer[vmn]/[tr][3p][caus]/jan
   BORDER
P1> (main)[partPerf]/behar(per)[partPerf]/izan(dum)[partFut]
      /edun(aux)[indPres][subj1s][obj3p]+lako[causal morpheme]
```

The last transducer eliminates the information of the input, and gives the desired output:

```
jan(main)[partPerf]/behar(per)[partPerf]/izan(dum)[partFut]
      /edun(aux)[indPres][subj1s][obj3p]+lako[causal morpheme]
```

The MT system gets this output for the next phase of the translation. The information between parenthesis will be used to decide the order of the words

Attribute	Value	Context
Aspm	[partPerf]	?* “tener” ?* “que” ?* BORDER “P1” ?*
Per	behar(per)	?* “tener” ?* “que” ?* BORDER “P1” ?*
Aspp	[partPerf]	?* VAIF ?* BORDER “P1” ?*
Dum	izan(dum)	?* [“tener” ?* “que” & VAIF] ?* BORDER “P1” ?*
Aspd	[partFut]	?* VAIF ?*
Aux	edun(aux)	?* “tener” ?* “que” ?*
TenseMood	[indPres]	?* VAIF ?*
SubjObjDat	[subj1s][obj3p]	?* “tr” ?* “pl” ?*BORDER?* “edun(aux)” ?* “1s”
RelMor	lako[causal morpheme]	?* caus

Table 1. Replacements of the example and contexts that constraint them

in the syntactic generation phase and the information between brackets will be used in order to do the morphological generation.

The translation obtained in the output of the system after the generation phase is the next sentence: *“ez ditudalako patatak jan behar izango”*

4.3 Compilation and use of the grammar

When compiling the grammar, the inverse of any set of rules is calculated and then each set is saved in one diferent transducer. The **lookup** utility applies the transducers to look up words. We use a lookup-strategy script that indicates all our transducers are to be applied one after the other, in a simulated composition.

The compilation of the grammar generates 25 transducers, occuping in all 2795 Kbytes.

In execution our module treats 250 verb chains/second.

5 Conclusions and future work

We have presented an FST grammar that faces the main problems that presents the structural transfer of verb chains from Spanish to Basque, giving a solution that works efficiently, achieving a coverage of 92%. We are studying the transfer of subjuntive and imperative forms, and the possible exceptions in the translation of non-finite forms.

This task is embedded in a transfer based Spanish-Basque MT framework and in the near future we want to include as much as possible all the FST technology and reuse the tools developed for the main project.

Finally, we are designing a new architecture based on a framework combining the different paradigms in MT.

6 Acknowledgments

This work has been partially funded by OpenTrad project FIT-340101-2004-3 (Spanish Ministry of Industry, Commerce and Tourism). Thanks to the other

research groups and companies in the project, and especially to Mikel Forcada and Sergio Ortiz-Rojas for their help in the design of the system.

Thanks to David Martinez for the helpful advices.

References

1. Abney S.: Partial parsing via finite-state cascades. Natural Language Engineering, Cambridge University Press (1996)
2. Alegria I., X. Artola, K. Sarasola, M. Urkia: Automatic morphological analysis of Basque. Literary and Linguistic Computing (1996)
3. Alshawi H., Bangalore S., Douglas S.: Learning Dependency Translation Models as Collections of Finite State Head Transducers. Computational Linguistics (2000)
4. Bangalore S. & Riccardi S.: A Finite-State Approach to Machine Translation. 2nd Meeting of the North American Chapter of the ACL(2001).
5. Beesley K. & L. Karttunen: Finite-State Morphology. CSLI Publications, Stanford, California (2003)
6. Canals-Marote, R., A. Esteve-Guillén, A. Garrido-Alenda, M.I. Guardiola-Savall, A. Iturraspe-Bellver, S. Montserrat-Buendia, S. Ortiz-Rojas, H. Pastor-Pina, P.M. Peŕez-Antn, M.L. Forcada: The Spanish-Catalan machine translation system interNOSTRUM. in B. Maegaard, ed., Proceedings of MT Summit VIII: Machine Translation in the Information Age, 73–76 (2001)
7. Carreras, X., I. Chao, L. Padrò and M. Padrò: FreeLing: An Open-Source Suite of Language Analyzers. Proceedings of the 4th International Conference on Language Resources and Evaluation -LREC'04 (2004)
8. Corbí-Bellot M., M. L. Forcada, S. Ortiz-Rojas, J. A. Perez-Ortiz, G. Ramirez-Sanchez, F. Sanchez-Martinez, I. Alegria, A. Mayor, K. Sarasola: An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain. Proceedings of the EAMT2005 (2005)
9. Díaz de Ilarraza A., Mayor A., Sarasola K.: Reusability of Wide-Coverage Linguistic Resources in the Construction of a Multilingual Machine Translation System. MT 2000. University of Exeter, United Kingdom (2000)
10. Garrido, A., A. Iturraspe, S. Montserrat, H. Pastor, M. L. Forcada (1999): A compiler for morphological analysers and generators based on finite-state transducers. Procesamiento del Lenguaje Natural, **25**, 93–98.
11. Hualde J.I., Oyharabal B., Ortiz de Urbina J.: Verbs. A grammar of Basque. Edited by Hualde J.I. and Ortiz de Urbina J. (1980) 155–250
12. Oflazer K.: Dependency Parsing with an Extended Finite-State Approach. Computational Linguistics (2003)
13. Roche E.: Parsing with Finite-State Transducers. in Finite-state language processing. MIT Press (1997)
14. Wu D.: Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. Computational Linguistics (1997)

This article was processed using the LaTeX macro package with LLNCS style