Towards a Spatial Annotation Scheme for Basque based on ISO-Space

Ainara Estarrona and Izaskun Aldezabal
IXA NLP group, University of the Basque Country
{ainara.estarrona}{izaskun.aldezabal}@ehu.eus

Abstract

The purpose of this paper is to create a preliminary spatial annotation scheme for Basque based on ISO-Space. To do this, we have first analysed in depth the ISO-Space annotation scheme, and then checked its suitability to apply it to Basque in order to expand the semantic tagging that is being developed in the IXA group. Although the typology of English and Basque are very different, we conclude that the model is also useful for Basque. However, as we went further on the scope of spatial structures, we noticed that this research field *per se* is still in an early stage and sometimes it was difficult to understand some concepts in the annotation scheme. Therefore, we have made our own proposals to tackle some of the problems that we have encountered.

Key Words: Semantic annotation, ISO-Space, spatiotemporal annotation, spatial relations, Basque.

1- Introduction

The annotation of spatial information is an important challenge for the development of advanced tools and applications such as machine translation, language learning or text summarization, among others. The aim of this paper is to create a preliminary scheme of spatial information for Basque based on ISO-Space. This paper is part of a more general ongoing project the IXA group[1] is pursuing about corpus-tagging frameworks. At the semantic level, the nouns have so far been tagged with Basque WordNet senses [2, 10] and the verbs have been annotated following the PropBank/VerbNet model [4][2]. Regarding spatiotemporal annotation, a corpus that contains temporal information has been created (*EusTimeBank*) following the EusTimeML mark-up scheme [3]. Our goal now is to lay the foundations to continue this ongoing project with the annotation of spatial information.

The present paper develops as follows: Section 2 presents a brief outline of ISO-Space; in Section, 3 we analyse the adequacy of the ISO-Space model to Basque. Section 4 discusses the theoretical problems and practical difficulties that we have found when analysing the ISO-Space scheme. Finally, Section 5 will outline some conclusions.

[1] http://ixa.si.ehu.es

[2] This semantic tagging makes use of the EPEC corpus (*Euskararen Prozesamendurako Erreferentzia Corpusa-Reference Corpus for the Processing of Basque*) [1], which contains 300,000 words of standard written text.

2- Brief overview of ISO-Space

With the aim of creating a model for labelling spatial information for Basque, we have first analysed the work carried out in this field for English.

2.1- Development of the spatial annotation process within SemEval

The annotation of spatial information is a shared task (SpaceEval) within SemEval since 2012 [6]. We have reviewed the development of the annotation process starting with the SpatialML [8] mark-up scheme, following with the Space Role Labeling [5] and ending with ISO-Space [10], the annotation scheme adopted as a standard since SemEval 2015 [11][3].

SpatialML was the basis of the annotation of spatial information. It provide a robust platform for the subtask of geolocating geographic entities and facilities in text, and to do that, it uses basic tags to identify locations and toponyms in the text. However, the complexity of spatial language, motivates a more expressive mark-up scheme.

The Space Role Labeling (SpRL) in SemEval 2012 [6] had a focus on the main roles of trajectors, landmarks, spatial indicators, and the links between these roles which form spatial relations. The formal semantics of the relations are divided into three types: directional, regional (topological), and distal. The annotated corpus, contained mostly static spatial relations. In SemEval 2013, the SpRL task was extended to the recognition of motion indicators and paths, which are applied to the more dynamic spatial relations.

The annotation scheme was extended enriching the semantics in static and dynamic spatial information and including new tags and relations. In this way the ISO-Space annotation scheme was created and it is the standard adopted since SpaceEval 2015.

2.2- The ISO-Space mark-up scheme

The descriptive mechanism of ISO-Space consists of a set of six basic entities and a set of four spatial relations over them. The basic entities are: *place, path, spatial_entity, motion_event, spatial_signal* and *measure.* All these entities are described by four spatial relations: *qualitative spatial link (qslink), orientation link (olink), movement link (movelink)* and *measure ling (mlink)* (see Table 1).

[3] The IXA group participated in SemEval-2015 task8 (SpaceEval) for the automatic recognition of spatial information following ISO-Space [13].

Table 1. The ISO-Space mark-up scheme.

Basic entities		Spatial relations
Location tags	**Non-location tags**	
Place	Spatial_entity	
Path	Spatial_signal	QSLINK
		OLINK
	Motion	MOVELINK
	Measure	MLINK

The *place* tag is used for annotating geographic and administrative entities (lakes, mountains, towns, countries...). The *path* tag is used for locations whre the focus in on the potential for traversal or funtions as a boundary (road, coast, Pacific Coast Highway...). The *spatial_entity* is a named entity that is not a location, but one which participates in an ISO-Space link tag. A *motion* tag is an spatial event involving change of location. The *spatial_signal* tag annotates typically prepositions or other function words that trigger spatial relations between two ISO-Space elements. A *measure* is a tag that captures distances and dimensions and it consists of a numerical component and a unit component or of a relative measurement term (*near, close, far...*).

The tags for spatial relations capture information about relationships between those tagged elements that we have mentioned in the previous paragraph. There are four link tags:

a) *qslink:* This tag is used to capture the topological relationship between two spatial objects and it is triggered by *spatial_signal* tags.

b) *olink:* This tag covers the relationships that are not topoligical and its trigger is a *spatial_signal*.

c) *movelink:* This tag connects all of the elements that participate in a motion event and it is introduced by a triggering *motion* tag.

d) *mlink:* This tag can be used to capture the distance between two objects and also to describe the dimensions of a single object and it is commonly accompanied by a *measure* tag (but this is not a requirement).

The annotation scheme also specifies a list of attributes and their values for each of these entities and relations.

3- Descriptive adequacy of the ISO-Space annotation scheme to Basque

In order to create an annotation scheme of the spatial information for Basque, we have based on the annotation guidelines for SpaceEval 2015[4] and we have

[4] http://jamespusto.com/wp-content/uploads/2014/07/SpaceEval-guidelines.pdf

translated the annotation scheme into Basque to be able to use it in our internal works. However, we have kept the tags in English (see Table 2).

Table 2. The attributes and values for the *spatial_signal* tag.

Attribute	Value	Extent
Id	s1, s2, s3...	Postposition
Cluster	Identifies the sense of the postposition[5]	
Semantic_type	DIRECTIONAL (1) TOPOLOGICAL (2) DIR_TOP (3)	
Ex.: - *Boston New York iparraldean dago* ('Boston is north of New York') - *Donostia Gipuzkoan dago* ('Donostia is in Gipuzkoa') - *Edalontzia mahai(aren) gainean dago* ('The glass is on the table')		

We have created similar tables for each tag, but in this case we have focused on this one because it gives us the opportunity to talk about one of the points in which English and Basque differ. In English spatial signals are typically prepositions while in Basque, which is an agglutinative language, they are postpositions. Therefore, words like *iparraldean* ('north of') in Basque should be segmented into lemmas and suffixes (postpositions) before annotating them (*iparralde* [lemma] + *an* [inessive suffix]). Although this typological contrast makes the annotation different, this does not represent a problem for the ISO-Space mark-up scheme, as demonstrated by Lee *et al.* [7] for another agglutinative language such as Korean.

Lexical differences may also result in dissimilar annotations as we can see with the motion verbs in Basque and English. In English some motion verbs such as 'bike' or 'walk' contain the information on the manner of moving inside them, but in Basque the manner of moving is usually expressed explicitly. For instance *oinez ibili* ('on foot go' = 'walk') or *bizikletaz ibili* ('by bike go' = 'bike'). This lexical feature can be annotated by ISO-Space without problems, because *motion* tag (see Table 3) has attributes for motion type (path or manner) and motion class (move internally, move externally, leave, reach, follow...). Therefore, ISO-Space mark-up scheme is adequate for the detailed annotation of various features associated with motion verbs [7].

[5] Preposition in English.

Table 3. The attributes and values for the *motion* tag.

Attribute	Value	Extent
Id	m1, m2, m3...	Verb
motion_type	MANNER, PATH, COMPOUND, GOAL[6]	
motion_class	MOVE, MOVE_EXTERNAL, MOVE_INTERNAL, LEAVE, REACH, DETACH, HIT, FOLLOW, DEVIATE, CROSS	
motion_sense	LITERAL, FICTIVE, INTRINSIC_CHANGE	
mod	A spatially relevant modifier	
countable	TRUE/FALSE	
Ex.: *Jon bizikletaz iritsi zen eskolara* ('Jon arrived at school by bycicle')		

In the same way that in English some verbs of movement contain the way of moving within themselves, in Basque some verbs of movement may contain in themselves the information about the *final_location* or *goal* of the movement (***etxera***_tu_ = 'home-to go'; ***zelaira***_tu_ = 'the field-into go'). We propose to add a fourth value for the *motion_type* attribute in order to tag this type of movement verbs in Basque. Therefore, we would have four values in the annotation scheme for Basque: *manner, path, compound* and *goal* (see Table 3).

4- Discussion

In this section we will focus on three topics. On the one hand, we will analyse both the '*non-consuming'* tags and the *motion_class* attribute of the *motion* tag together, because we think that they are closely related; and on the other hand, we will talk about the *spatial_signal* tag and the links that it triggers.

Non-consuming tags were created to capture spatially relevant locations or entities that are not directly referenced in the text. The extent of *non-consuming* tags is a null or empty string. In SpaceEval Annotation Guidelines[7] it is said that, generally, non-consuming tags are not necessary to capture relevant spatial objects and relations and that for this reason they should be used sparingly. In fact, they mention three situation where the use of these tags is necessary: i) locations referenced by a measure; ii) locations

[6] The *goal* value does not appear in ISO-Space. It is a value that we have added to tag a particular type of verbs that exist in Basque, the verbs that contain in themselves the information about the final location or goal of the movement.

[7] http://jamespusto.com/wp-content/uploads/2014/07/SpaceEval-guidelines.pdf

implied by ‘cross’ and ‘across’; and iii) sets whose members are mentioned[8]. The reason to use these tags in such situations and not in others is that these tags are necessary to fill the value of certain attributes in other tags or links. We may agree with this assumption, but we do not understand why in other situations, for example, in the case of *movelink* tags the *non-consuming* tags are not necessary. In the *motion* tag there is an attribute that is *motion_class* (see Table 3 in section 3). The guidelines [12] specifies which attributes are required in the *movelink* tag for each *motion_class* (see Table 4). We assume that if these attributes are required, they should be specified using a *non-consuming* tag, because they are necessary to fill the *source, goal, midpoint* or *ground* attributes of the *movelink* tags (see Table 5). However, the guidelines do not specify anything about it.

Table 4. Required attributes in the *movelink* tag for each *motion_class*.

***motion_class* of trigger**	**Required Attributes**
move	-
move_external	landmark[9]
move_internal	landmark
leave	source
reach	goal
detach	source
hit	goal
follow	pathID[10]
deviate	source
cross	source, midPoint, goal

[8] This third situation is not mentioned in Pustejovsky [12].

[9] *Ground* in Pustejovsky [12].

[10] *Goal* in Pustejovsky [12].

Table 5. The attributes and values for the *movelink* tag.

Attribute	**Value**	**Trigger**
Id	mvl1, mvl2, mvl3...	Movement verb
trigger	ID of a MOTION that triggered the link	
source	ID of a location/entity/event tag at the beginning of the event-path	
goal	ID of a location/entity/event tag at the end of the event-path	
midPoint	ID(s) of event-path midpoint location/entity/event tags	
mover	ID of the location/entity/event tag whose location changes	
ground	ID of a location/entity/event tag that the mover participant's motion is relative to	
goal_reached	TRUE, FALSE, UNCERTAIN	
pathID	ID of a PATH tag that is identical to the event-path of the trigger MOTION	
motion_signalID[11]	ID(s) of (an) MOTION_SIGNAL tag(s) that contributes path or manner information to the trigger MOTION	
Ex.: *[Jonek$_{se2}$[12]] [autoz$_{ms3}$] [bidaiatu$_{m1}$] zuen* ('Jon$_{se2}$ traveled$_{m1}$ by_car$_{ms3}$') trigger=m1; mover=se2; motion_signalID=ms3		

In order to help the annotators know when to use these *non-consuming* tags and when not, we think it is necessary to establish precise criteria. Hence we have made a proposal based on the verbal subcategorization of Basque that

[11] *AdjunctID* in Pustejovsky [12].

[12] *Se* = spatial_entity, *ms* = motion_signal and *m* = motion.

we have collected in our verbs lexicon *Basque Verb Index* (BVI)[13]. We propose to annotate with *non-consuming* tags all the subcategorised elements or arguments of a given movement verb that are not explicitly referenced, but which can be retrieved from the text using coreference[14] (1).

(1) *Mikel Indiara joan zen oporretan.*
Mikel to-India go.partc was on-holiday
'Mikel went on holiday to India'

[Indiara] *iritsi bezain_laster damutu zen.*
[To-India] arrive.partc as soon as regret.partc was
'As soon as he arrived [to India] he regretted it'

In (1) we would create a non-consuming tag for the final location argument of the verb *iritsi* ('to arrive'), because we can retrieve it from the previous sentence. Nevertheless in (2) we would not create a non-consuming tag, because the final location of the verb *joan* ('to go', 'to leave') is underspecified and can not be retrieved from the text.

(2) *Halaxe joan zen mundu honetatik Joanes, Bargotako Brujoa.*
Like this go.partc was world this-of Joanes, Bargota-of Sorcerer-the
'This is how Joanes, the sorcerer of Bargota, left this world'

Following with the verbs of movement, the classification of the movement classes and their event-path structures presented in the guidelines is very interesting for crosslingual studies. As mentioned above, each *motion_class* has an argument which is focused depending on the structure of the event-path. For example, while the focus for the 'leave' class is the 'source', the focus for the 'reach' class is the 'goal'. In our case, we have not yet done this analysis for Basque, and therefore, we find it difficult to classify the Basque verbs based on such criteria. For instance, the verb *joan* ('to go') can be both a 'leave' class and a 'reach' class verb depending on the underspecified locative arguments, that is, when the underspecified locative argument is the *goal,* it will be a 'leave' *motion_class* verb. In fact, in this case the verb *joan* would be translated as 'to leave', and not as 'to go', in English. However, we think that in order to carry out a motion-class study for Basque it is necessary to annotate previously a sample of corpora following the proposed criteria, since both the omission and underspecification phenomena can only be identified at surface syntax level.

[13] http://ixa2.si.ehu.es/e-rolda/index.php

[14] In some cases, we will need a wider context than the phrase in order to identify these elided elements.

Finally, the *spatial_signal* tag has caused us problems, since it has been sometimes difficult to differentiate between the three semantic types that are mentioned in the guidelines (*topological, directional* and *dir_top*) and the relations that they introduce (*qslink* and *olink*). In Pustejovsky [12] when measuring the inter-annotator agreement for each tag, the tags *qslink* and *olink* are the ones that get the lowest results. This suggests to us that perhaps the distinction between them is not clear enough and that depending on the particular task or application in which the annotation will be applied it is likely that this distinction will not be necessary, or at least not in that degree of detail.

5- Conclusions

In this paper we have presented a brief overview of the ISO-Space annotation scheme and we have analysed its adequacy for labelling spatial structures in Basque.

The main conclusion we have drawn is that the ISO-Space mark-up scheme is suitable for tagging spatial information in Basque. However, as Basque is an agglutinative language, the identification of markables needs sometimes to resort to smaller segments than word forms. In addition, we have enriched the annotation scheme with a new type of movement for the *motion* tag, the motion type *goal*, necessary to annotate this type of verb that exists in Basque.

In this paper we have presented the first attempt to adapt the ISO-Space mark-up scheme to Basque, and therefore, although this model can be taken as a general theoretical framework, in the future and with concrete applications in mind, it will be necessary to delimit the annotation scheme depending on those specific applications.

Acknowledgements

This research has been supported by the University of the Basque Country (IXA group, (GIU16/16)) and the Ministry of Economy, Industry and Competitiveness of the Spanish Government (TUNER: TIN 2015-65308-C5-1-R; MUSTER: PCIN-2015-226).

References

[1] Aduriz, Itziar, Aranzabe, María Jesús, Arriola, Jose. María, Atutxa, Aitziber, Díaz de Ilarraza, Arantza, Ezeiza, Nerea, Gojenola, Koldo, Oronoz, Maite, Soroa, Aitor and Urizar, Ruben (2006) Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing. In Andrew Wilson, Paul Rayson and Dawn Archer (eds.), *Corpus Linguistics Around the World.* Book series: Language and Computers. Vol. 56, 1-15. Rodopi (Netherlands).

[2] Agirre, Eneko, Aldezabal, Izaskun, Etxeberria, Jone, Izagirre, Izaskun, Mendizabal, Karmele, Pociello Eli and Quintian, Mikel (2006) A methodology for the joint development of the Basque WordNet and Semcor. In *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC)*. Genoa, Italy.

[3] Altuna, Begoña, Aranzabe, María Jesús and Díaz de Ilarraza, Arantza (2017) EusHeidelTime: Time Expression Extraction and Normalisation for Basque. *Procesamiento del Lenguaje Natural,* n.º 59, 15-22.

[4] Estarrona, Ainara, Aldezabal, Izaskun, Díaz de Ilarraza, Arantza and Aranzabe, María Jesús (2016) Methodology for the semiautomatic annotation of EPEC-RolSem, a Basque corpus labelled at predicate level following the PropBank/VerbNet model. Edward Vanhoutte (ed.) *Digital Scholarship in the Humanities* (2016) 31 (3): 470-492. DOI: http://dx.doi.org/10.1093/llc/fqv010. First published online: 17 June 2015 (23 pages). Published by Oxford University Press on behalf of EADH: The European Association for Digital Humanities.

[5] Kordjamshidi, P., VanOtterlo, M. and Moens, M.F. (2010) SpatialRoleLabeling: Task Definition and Annotation Scheme. *Proceedings of the Seventh conference on International Language and Resources and Evaluation (LREC'10).* 413-420. European Language Resources and Evaluation (ELRA).

[6] Kordjamshidi, P., Bethard, S. and Moens, M.F. (2012) SemEval-2012 Task3: SpatialRoleLabeling. *In Proceedings of the 6th International Sorkshop on Semantic Evaluation (SemEval),* 365-373.

[7] Lee, Kiyong, Fang, Alex C. and Pustejovsky, James (2011) Multilingual Verification of the Annotation Scheme ISO-Space. *Fifth IEEE International Conference on Semantic Computing,* 449-458.

[8] Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J., Wellner, B., Mardis, S. and Clancy, S. (2010) SpatialML: annotation scheme, resources, and evaluation. *Language Resources and Evaluation,* Volume 44 (3), 263-280. Springer.

[9] Pociello, Eli, Agirre, Eneko and Aldezabal, Izaskun (2010) Methodology and Construction of the Basque WordNet. *Language Resources and Evaluation Journal,* 45:2, 121-142. Springer.

[10] Pustejovsky, James, Moszkowicz, Jessica and Verhagen, M. (2012) A Linguistically Grounded Annotation Language for Spatial Information. *Traitement Automatique des Langues (TAL),* Volume 53 – nº 2/2012, 87-113.

[11] Pustejovsky, James, Kordjamshidi, P., Moens, M.F., Levine, A. Dworman, S. and Yocum, Z. (2015) Semeval-2015 task 8: Spaceeval. *Proceedings of the 9th Internationa Workshop on Semantic Evaluation.* 884-894.

[12] Pustejovsky, James (2017) ISO-Space Annotating Static and Dynamic Spatial Information. Nancy, Ide and James, Pustejovsky (eds.) *Handbook of Linguistic Annotation,* 989-1024. Springer Netherlands.

[13] Salaberri, Haritz., Arregi, O. and Zapirain, Beñat (2015) IXAGroupEHUSpaceEval: (*X-Space*) A WordNet-based approach towards the Automatic Recognition of Spatial Information following the ISO-Space Annotation Scheme. *Proceedings of the 9th*

International Workshop on Semantic Evaluation (SemEvan2015), 856-861. Denver, Colorado, June 4-5, 2015. Association for Computational Linguistics.