eman ta zabal zazu

UNIVERSITY OF THE BASQUE COUNTRY (UPV/EHU)
Department of Basque Language and Communication

PhD dissertation

Analysis of Basque Temporal Constructions and Creation of a Corpus

Begoña Altuna Díaz

2018

eman ta zabal zazu

UNIVERSITY OF THE BASQUE COUNTRY (UPV/EHU)
Department of Basque Language and Communication

Analysis of Basque Temporal Constructions and Creation of a Corpus

This summary is a shortened and translated version of the dissertation entitled “Euskarazko denbora-egituren azterketa eta corpusaren sorrera”, written by Begoña Altuna Díaz under the supervision of Dr. Arantza Díaz de Ilarraza and Dr. María Jesús Aranzabe.

Donostia, 2nd of October 2018.

You never learn to swim until you are in the water.

Jack Ma

Amari

Acknowledgements

Hasteko, eskerrak eman nahi dizkiet Maxuxi eta Arantzari tesi-lan hau egin bitartean emandako laguntza guztiagatik eta ikerkuntza zer den irakasteagatik. Eskerrik asko bost urtean ni aguantatzeagatik.

Eskerrik asko Ixa taldean nirekin egon zareten guztioi! Bereziki Kortako jendeari, zuek izan zaretelako nire zalantza gehienak argitu dituzuenak eta zuek egin duzuelako (Korta mundutik aparte badago ere) inoiz bulegoz aldatu nahi ez izatea.

Ezin ditut ahaztu tuppertxokokoak. Elkarrizketa bizarro guztiengatik, informazio extraofizial guztiagatik, bazkalosteko atsedenengatik, harrikoa egiteko teknikarik optimizatuena irakasteagatik eta beste hainbatengatik, eskerrik asko!

Dena ez da Kortan eta fakultatean gertatzen, ordea. Eskerrik asko pintxopotera animatu izan zareten guztiei (Pintxopote Buru-Batzarrari eta Pintxopotearen Zaindariei bereziki) askotan osteguna nire asteko egunik gogokoena bihurtzeagatik. Eskerrik asko Matia kaletik kanpo egin ditugun planetara etorri zareten guztiei ere; Orhiseak ez bagintuen akabatu, ez gaitu lanak hobiratuko!

Eskerrik asko nirekin bizi izan zareten guztiei ere, bai Casa Mertxeko nire arkitekto maiteei, bai Tolosa 87koei (alfabetikoki, inor haserretu ez dadin, Ander, Manex, Mara eta Olatz). Oso guay egon dira pisu-poteak, afariak eta karaokeak.

Eskerrik asko familiari azken urtea baino ez duelako pasatu tesia noiz defendatuko dudan galdetzen. Plazera da lanari buruz hitz egiten ez duen

jendearekin egotea, azken urtean pelma batzuk izan bazarete ere. Noski, eskerrik asko Pedromariri tupperrak egiteagatik eta Itsasori nire Deustuko gela oraindik bere billar gela ez bihurtzeagatik.

Eskerrik asko Deustuko (eta Bakioko) jendeari, kuadrillari, mendi taldeari (eta Izabako jendeari) eta filologoei (eta Marruskilariei). Nahiz eta AP-8ak banatu zuek beti egon zaretelako ni hartzeko prest.

Vorrei ringraziare la gente della FBK, sopratutto Manuela e Anne-Lyse, ma anche Mara, Mohammed e Roldano e tutti gli altri. Vorrei salutare anche la gente della FB3 al Sanba (e tanti altri), perché mi hanno accolto dal primo momento.

Tesia amaitzen ari naizen honetan inor ahaztu badut, barkatu. Zuei ere, eskerrik asko!

Official acknowledgments

This dissertation was funded by the PRE_2013_1_959, PRE_2014_2_242, PRE_2015_2_0284 and PRE_2016_2_294 grants awarded by the Basque Government's Education, Universities and Research Department, as well as by the DETEAMI (Basque Government, ref.: 2014111003) and PROSA-MED (Economy and Business Ministry, Spanish Government, ref.: TIN2016-77820-C3-1-R) proyects.

Abstract

In this research work we have taken the first steps in temporal information processing for Basque. For this, we have based our research on similar works done for other languages as well as on the linguistic analysis of temporal constructions in Basque. Then, we have employed that knowledge on the recognition of the most relevant linguistic features of Basque temporal constructions and we have built the EusTimeML mark-up language to normalise them. In addition, we have manually annotated the EusTimeBank corpus following EusTimeML in order to create a temporal information corpus.

That corpus, apart from being used for linguistic phenomena analysis, has been employed in automatic tool development and evaluation. In fact, in this research work we have developed EusHeidelTime for time expression extraction and normalisation and we have built KroniXa, the automatic timeline creation system from Basque texts. We have settled the first steps on the integration of those tools in the Basque processing chain, in order to be able to add temporal information to automatic Basque processing.

Contents

INTRODUCTION

1 Introduction

This work pertains to the Natural Language Processing (NLP) field; more precisely to the temporal information processing of Basque. This work is part of the research done in Ixa Group, which has worked for the last 30 years developing tools and resources for NLP in Basque.

Humans continuously deal with narrations full of events and temporal references and are able to distinguish what happened and when easily. For computers, instead, understanding textual information requires multi-layer linguistic analysis. In this dissertation we will describe the way temporal information is expressed in Basque, the mark-up language created for making temporal information explicit for its processing, the annotated corpus we have built and the tools that process and take advantage of the extracted temporal information.

1.1 Temporal Information

Temporal information is the one concerning the expression of what happens and when it happens. For example, in (1) it is said that *the landing* happened at *18:40*.

(1) Lehen Airbus A380a *18:40an (GMT+8)* **lurreratu da** Singapurren.
(The first Airbus A380 **has landed** at *18:40 (GMT+8)* in Singapur.)

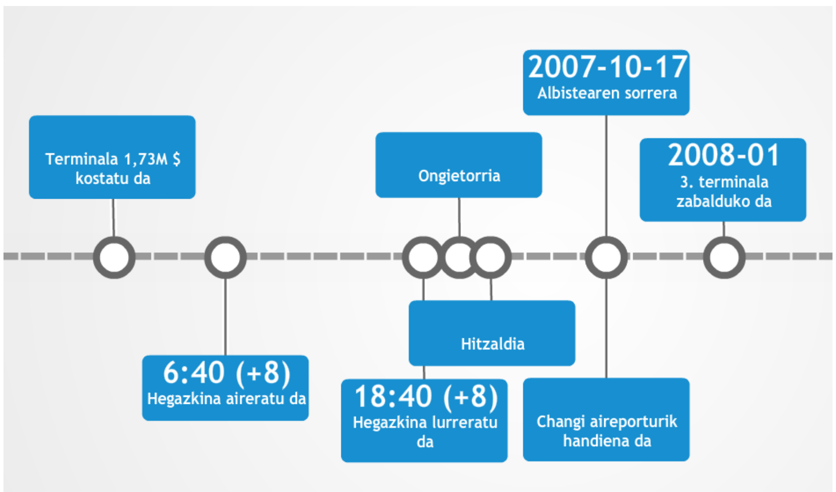
We call the actions and situations that occur *events* and the moments and durations in text are *time expressions*. Furthermore, we know they do not happen individually in text; there are relations between those words. The relations that connect two events and/or time expressions and imply certain ordering in time of the two instances are *temporal relations*. Consequently, one can say that if the *landing* in (1) happened at *18:40* there is a simultaneity relation between both elements.

Nonetheless, temporal information is not only relevant at sentence level. Normally humans produce much longer texts and more events and time expressions are present in those. For example, the sentence in (1) is part of the following piece of news:

> **Lehen A380a hasi da zerbitzu komertziala ematen**
>
> 2007ko urriaren 17a
>
> Hegazkingintzaren historian, sekulako lorpena izan da: *lehen Airbus A380a 18:40an (GMT +8) lurreratu da Singapur-en*, Changi nazioarteko aireportuan, Airbus-en bidalketa-zentrotik atera eta 12 orduko hegaldia egin ostean. Hegazkinari 400 bat gonbidatuk egin zioten ongietorria laster zabalduko den Changi Nazioarteko aireportuko 3. terminalean.
>
> Tartean zen Lee Hsien Loong, Singapur-eko lehen ministroa. Hitzaldian, lehen ministroak Singapore Airlines (SIA) aire-linea hartzailea txalotu zuen, eta aipatu zuen arrakasta Singapur-eko biztanle guztiak harro egoteko modukoa zela. Gehitu zuen lurreratzea gertaera gogoangarria zela, ez bakarrik SIArentzat, baita Changi nazioarteko aireportuarentzat ere, zeina herrialdeko aireporturik handiena baita.
>
> Orain arteko terminal guztiak prest daude A380a hartzeko, eta 3. terminalak (1,75 milioi dolar singapurtar kostatu da) datorren urteko urtarrilean zabalduko ditu ateak.
>
> A380ak ongintzazko hegaldi berezi bat egingo du urriaren 25ean, Singapur-etik Sydney-ra, eta Singapur-Sydney ibilbideari dagokion ohiko zerbitzua urriaren 28an hasiko du.

Temporal information helps ordering textual information. In the previous news text three main events are mentioned: i) the landing after the maiden flight of the Airbus, ii) the fact that all the terminals in the airport are ready and iii) a charity flight the Airbus will do. Those can be placed in a timeline according to the times they happen in.

In timelines, time can be represented as a linear axis that goes from the past to the future and events can be placed linked to the temporal anchors they happen in. Equally, events with no explicit anchors can be ordered relative one to another and then placed in an approximated place in the timeline. The information in the piece of news in this section has been arranged chronologically in the timeline displayed in Figure 1.1.



1.1 Figure – Timeline extracted from the piece of news *Lehen A380a hasi da zerbitzu komertziala ematen* (The first Airbus A380 has started offering commercial service)

1.2 Motivation and Goals

As mentioned earlier, we aim at exploiting temporal information in Basque. In fact, temporal information can be useful in tasks such as chronology creation (Yan *et al.*, 2011; Minard *et al.*, 2015; Bauer *et al.*, 2015), causal information processing (Mirza and Tonelli, 2014), question answering (QA) systems (Pustejovsky *et al.*, 2003a; Saurí *et al.*, 2005; Mirza and Minard, 2015) or in query systems (Chieu and Lee, 2004; Setty *et al.*, 2010). Additionally, temporal information can also be employed in event prediction (Radinsky and Horvitz, 2013) and future forecasting (Kawai *et al.*, 2010).

The main goal of this work is developing the resources for temporal information processing in Basque in order to be able to use the extracted and normalised temporal information in more advanced tools. For that, we have defined the following intermediate goals:

- Analysis of the forms that convey temporal information.
- Development of a mark-up language for temporal information in Basque.
- Creation of an annotated corpus for experimentation from news and historical texts.
- Creation of tools for temporal information extraction and normalisation.
- Development of a timeline creation tool.

1.3 Dissertation Structure

After this introductory section, we will present the works that have been done in temporal information processing in Chapter 2. Then, we will focus in temporal information processing for Basque.

In Chapter 3 we describe the ways temporal information is expressed in Basque. In Chapter 4 we present the resources we have created for temporal information processing in Basque: the EusTimeML mark-up language and the EusTimeBank corpus. In Chapter 5 we present the tools we have developed for temporal information processing: EusHeidelTime for time expression extraction and normalisation and KroniXa for automatic timeline creation. To conclude, we present the contributions and conclusions extracted from our research work in Chapter 6.

1.4 Works Attached to the Dissertation

The work on this research has been presented in many meetings and publications. We list them below according to their publication place and we order them chronologically.

Journals

- Altuna B., Soraluze A., Aranzabe M.J., Arregi O., and Díaz de Ilarraza A. KroniXa: Timeline Creation from Basque Texts. *Journal of Information Processing and Management*, Under revision

- Salaberri H., Altuna B.A., Aranzabe M.J., Arregi O., and Díaz de Ilarraza A. bTime: a Hybrid Architecture for Capturing Temporal Information in Basque. *Knowledge-Based Systems*, Under revision

- Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. EusHeidelTime: Time Expression Extraction and Normalisation for Basque. *Procesamiento del Lenguaje Natural*, 59(0):15–22, 2017a. `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5488`

- Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. Euskarazko denbora-informazioaren tratamendu automatikoa TimeMLren eta HeidelTimeren bidez. *Ekaia*, 30:153–165, 2016b

- Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. *Linguamática*, 6(2): 13–24, 2014a. `http://linguamatica.com/index.php/linguamatica/article/view/v6n2-1`

Chapters in books

- Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. An Event Factuality Annotation Proposal for Basque. In Frank A.U., Ivanovic C., Mambrini F., Passarotti M., and Sporleder C., editors, *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*, 1 lib., 15–24, Vienna, Austria, 2018b. `https://www.oeaw.ac.at/ac/crh2/proceedings/`

- Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. Euskarazko ezeztapenaren tratamendu automatikorako azterketa. In Alegria I., Latatu A., Ormaetxebarria M.J., and Salaberri P., editors, *II. IkerGazte, Nazioarteko Ikerketa Euskaraz: Giza Zientziak eta Artea*, 127–134, Bilbo, 2017b. `http://ixa.si.ehu.eus/sites/default/files/dokumentuak/8863/IKERGAZTE.2017.GIZAZIENTZIAKetaARTEA.pdf`

- Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. Adapting TimeML to Basque: Event Annotation. In Gelbukh A., editor, *Computational Linguistics and Intelligent Text Processing*, 565–577, Cham, Switzerland, 2018a

- Altuna B. Análisis de la información temporal en euskera. In Martínez Barco P., Navarro Colorado B., Vázquez Pérez S., and Romá Ferri

M.T., editors, *Actas del XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Universidad de Alicante, Alicante, Spain, 2015a

- Altuna B. Euskarazko denbora-egituren tratamendu automatikorako azterketa. In Alegria I., Latatu A., and Omaetxebarria M.J., editors, *I. Ikergazte, Nazioarteko ikerketa euskaraz*, 46–53, Bilbao, Spain, 2015b

Technical reports

- Altuna B., Aranzabe M., and Díaz de Ilarraza A. Euskarazko denbora-egiturak etiketatzeko gidalerroak v2.0. Barne-txostena, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2016, 2016a. `https://addi.ehu.es/handle/10810/17305`

- Altuna B., Aranzabe M., and Díaz de Ilarraza A. Euskarazko denbora-egiturak etiketatzeko gidalerroak. Barne-txostena, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2014, 2014b. `http://ixa.si.ehu.es/Ixa/Argitalpenak/Barne_txostenak/1414871293/publikoak/Denbora-egiturak%20etikeatzeko%20gidalerroak`

Apart from the research on strictly speaking Basque temporal information processing, we have done some complementary research. The works on those additional issues are presented below.

- Altuna B., Minard A.L., and Speranza M. The Scope and Focus of Negation: A Complete Annotation Framework for Italian. *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, 34–42. 2017c. `http://aclweb.org/anthology/W17-1806`

- Minard A.L., Speranza M., Urizar R., Altuna B., van Erp M., Schoen A., and van Son C. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In Calzolari N., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., and Piperidis. S., editors, *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*. 2016

- Agerri R., Agirre E., Aldabe I., Altuna B., Beloki Z., Laparra E., de Lacalle M.L., Rigau G., Soroa A., and Urizar R. NewsReader project. *Procesamiento del Lenguaje Natural*, 53:155–158, 2014

To conclude this section we present the awards obtained by some of those works.

- SEPLN 2017 best paper award in *XXXIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* (Murcia, Spain, 2017), for the paper *EusHeidelTime: Time Expression Extraction and Normalisation for Basque.*

- Best oral presentation award in Human sciences area in *II. Ikergazte, Nazioarteko ikerketa euskaraz* conference (Iruñea, Spain, 2017), for the paper *Euskarazko ezeztapenaren tratamendu automatikorako azterketa.*

- Best paper award in Human sciences area in *I. Ikergazte, Nazioarteko ikerketa euskaraz* conference (Durango, Spain 2015), for the paper *Euskarazko denbora-egituren tratamendu automatikorako azterketa.*

2 Background and Related Works

More and more complex language analysis have been done in NLP and now we are able to process semantics and discourse level information. Temporal information analysis and processing is part of those, as semantics of the events and time expressions and the relations among those are analysed, and, thus, discourse is analysed.

Temporal information processing has largely improved in the last two decades. Taking into account the theoretical models of the mid 20^{th} century, mark-up schemes and tools for temporal information identification and normalisation have been developed. We will present the most relevant works done for temporal information processing in this section. First, we will define the most important terminology in Section 2.1 and we will describe the theoretical background in Section 2.2. In Section 2.3, we will present the temporal information mark-up languages and the manual annotation tools. Then, tools for automatic temporal information extraction will be presented in Section 2.4. We will analyse the corpora that contain temporal information in Section 2.5. Finally, we will present some of the most relevant works on automatic timeline creation in Section 2.6.

2.1 Terminology

Many terms have been employed to name the temporal information elements. Some, such as *event* and *eventuality* come from the semantics and logic tradition, whereas other have been created to address new concepts. In this section we will describe which terms and with which sense we will use them in our work.

2.1.1 Events

Events have been described in many ways. Some definitions are prominently generic, whereas other are closely related to time and hence, more relevant to our task. One can say the definitions of events have gained complexity over time. Bach (1986) named events *eventualities* and defined them as events and situations that happen, obtain or endure. (Bayer, 1986, p. 28), instead, gave a more generic definition and stated that "events are spatio-temporally located bodies".

In the last years, Wonsever *et al.* (2008) have considered that events are any action or situation denoted by a predicate. Saurí *et al.* (2009), instead, following (Bach, 1986, p. 3), have stated that *event* is a "cover term for situations that happen, occur, hold, or take place and states and circumstances in which something obtains or holds true". Finally, Dowden (2009) related events with time and said that in ordinary discourse, an event was usually understood as a happening lasting a finite duration during which some object changed its properties.

Some definitions are more closely related to the task. (Setzer, 2001, p. 10) considered that "an event is intuitively something that happens, with a defined beginning and end" while a state was "a relation between the entities or the holding of an attribute of an entity which, while capable of change, is ongoing over a time span; often without a defined beginning or end". Nonetheless, both were considered to last in time, be part of the real world and be expressed by propositions.

Finally, we reckon the definition of events offered in (Wonsever *et al.*, 2012, p. 207) is the one that best represents our idea of events:

> *events can be actions (carried out voluntarily by an agent), processes (events spontaneously set off or caused by a force external to the process, which can, in both cases, be punctual or have duration), or states (situations maintained along a period or that are permanent). Generic predications will also be considered as events for they refer to states of things, states about which it is asserted that they take place.*

We have opted for it as it considers generic predications as events.

In what concerns defining the term for *event* in Basque, we have identified *gertakari* (2) and *gertaera* (3) used indistinctly in Altuna *et al.* (1987). We have also observed the same trend in *Sareko Euskal Gramatika*[1] (Online Basque Grammar). For our research we have opted for *gertaera* as the translation for *event*.

(2) ... adizki aspektudunek *gertaera* gertagarriago bat edo adierazten bide baitute.

(3) ... testuinguru berezi batzutan iragan burutua, aspektu perfektua adieraztean, gaurko *gertakari* bati buruz erabil dezakegula.

2.1.2 Time Expressions

Time expressions express the times events happen. (Schilder and Habel, 2001, p. 1) define them as "chunks of text that express some sort of direct or inferred temporal information". Ahn *et al.* (2005) offered a more detailed definition and said that temporal expressions were natural language phrases that directly referred to time points or intervals. Finally (Bittar, 2010, p. 38) only offers a simplistic definition: "the expressions used to express times".

We have noticed that *time expression* and *temporal expression* are used in literature to express the same concept. That is what happens, for example in Ahn *et al.* (2005) and Hacioglu *et al.* (2005). For our research, we have chosen the term *time expression* for it seems to be more concise as we have observed that *temporal expression* has been employed in literature to express all the elements taking part in temporal information. In Basque we use *denbora-adierazpen*.

2.1.3 Signals

Signals are defined in Pustejovsky *et al.* (2003a) as "sections of text, typically function words, that indicate how temporal objects are to be related to each other". Actually, signals can make explicit any kind of relation. For example in the first version of TimeML (Pustejovsky *et al.*, 2003a, p. 3), signals for subordination relations were analysed and causality relations were marked as `C-SIGNAL` in the NewsReader project (Tonelli *et al.*, 2014). Nonetheless, in our work we will only consider signals for temporal relations and we will call them *seinale* in Basque.

[1] `http://www.ehu.eus/seg/aurkezpena`

2.1.4 Relations

In this dissertation we will go through the relations between events and/or time expressions. We have defined three kinds of relations: temporal relations, subordination relations and aspectual relations. First, temporal relations are the ones created between events and/or time expressions, by means of which events and time intervals are anchored in time or the duration of an event is defined. Secondly, subordination relations link an event and a subordinated event which is its argument. Finally, aspectual relations express the phase of an event (beginning, end, continuation, etc.) and are built between the event that indicates the phase and its subordinated event.

2.2 Theoretical Background

Much research has been done on the temporal information area during the last decades and it has settled the basis for temporal information analysis used for automatic processing. Events are the nucleus of the sentence and convey much of the temporal information. Some researchers, namely Vendler (1957) and Smith (1991), have focused on event typology, whereas the aspect and tense have been the target for other scholars (Reichenbach, 1947). Another research area has been the analysis of the relations between time intervals (Allen, 1983). Hence, not only the events but the way humans situate those on a timeline has been studied.

As one may notice, these theoretical issues cover a wide range of topics and, thus, we will be presenting them in the following sections. First, we will describe some of the event classifications that are most closely related to temporal information processing in Section 2.2.1. Secondly, we will give an account of other theoretical studies concerning event expression in Section 2.2.2. We will describe theories on time conceptualisation in Section 2.2.3. Finally, we will present Allen's Interval Theory in Section 2.2.4.

2.2.1 Event Classification

Not all the events happen the same way, they may happen differently with respect to time. Additionally, some authors have mentioned evolution some events go through as some events also express some kind of transition or alteration. Following another point of view, many linguist have classified events or predicates according to their linguistic features. We will discard the classifications based on syntactic features and we will focus on semantic classifications of events, since they are more closely related to our task.

2.2.1.1 Vendler's Classification

Vendler (1957), in his attempt of making the difference between processes and non-processes, classified events in four types: activity, accomplishment, achievement and state. The main distinction between those four groups was the way they happened in time. In the following Table 2.1 one can see the four distinct event types identified, their main features and some examples of each kind.

2.1 Table – Vendler's event typology

Event type	**Features**	**Examples**
State	It captures an aspect of a world at a single point in time. These events have no internal phases or changes.	love, know, desire
Activity	They are events undertaken by a participant that have no particular end point.	run, swim, push, draw
Accomplishment	They are events that have a natural end point and result in a particular state.	draw a circle, read a book, run a mile
Achievement	They happen in an instant and result into a different state.	reach, win, stop

As one can see from Table 2.1, events can be divided according their relation to time: states and activities do not have any explicit begin nor end point, accomplishments seem processes that last a time and then end in a new situation and achievements are events that happen in a single point of time and virtually last no time. Vendler only considered activities, accomplishments and achievements to be events, but admitted some states could be considered events in some contexts. Such is the case for *to sit*, *to stand* or *to lie*. However, it is also to be pointed out that Carlson (1977) added that all events in Vendler's classification could have a non-eventive reading. For

2.2 Table – Croft's event classification

Event type	**Event subtype**		**Examples**
States	transitory		The road is wet.
	permanent	acquired	The window is broken.
		inherent	Marie is French.
	point		It's 5 AM.
achievements	reversible directed		The window is open.
	irreversible directed		The window is smashed.
	cyclic		The mouse squeaked.
activities	directed		The soup cooled.
	undirected		The girls chanted.
performances	incremental accomplishments		I ate an apple.
	non-incremental accomplishments		Harry repaired a computer.

example, *John smokes* is comparable to *John is a smoker*, which is not an event according to Vendler.

Smith (1991) added a fifth event group to Vendler's calssification. She took the slavic denomination and proposed "semelfactive" events to enlarge Vendler's typology. These events are instantaneous and atelic, that is to say, these events happen in an instant but can go indefinitely. Some examples for this fifth type are *cough*, *blink* and *sneeze*.

2.2.1.2 Croft's Classification

Croft (2015) proposed a new classification of events according to their unfolding over time. He proposed four categories: states, achievements (instantaneous changes of state), activities (durative and unbounded processes) and performances (bounded processes). Additionally, a subclassification was also offered. Those classes and their subclasses are described in the following Table 2.2.

2.2.1.3 Setzer's Classification

Setzer discussed Vendler's ranking among others and deduced that the classification of events based on aspect was not appropriate for temporal information processing. According to Setzer, Vendler's aspect based approach required sentence level analysis, which is not possible in journalistic texts because those text are formed by short paragraphs. Additionally, Setzer stated that classifying each event as a point, culmination, culminated process, or process was a big burden for the annotators.

Hence, Setzer created the following event classification:

- **Occurrence events**: events that will appear on the temporal graph.
- **Reporting events**: events that connect the source of information and an event. Usually, they occur after the indicated event and help to place that event in time.
- **Perception events**: they indicate that an event is "felt". Usually, the occurred event and the feeling that the event has occurred happen together.
- **Aspectual events**: they take an event as argument and indicate the start or end of that event.
- Other kind of events:
  - **Attitude events**: they take an event as argument, but they do not guarantee the reality of that event, nor even a clear temporal relation for that event.
  - **Hypothetical events**: they produce hypothetical timelines based on the consequences.

As one can see, this sort of classification of events made explicit relationship between each type of event and time. It is noteworthy that it does not take into account situations, which seemed impossible to be placed in time. It was considered the base of the TimeML event classification.

2.2.1.4 TimeML Classification

During the development of the TimeML mark-up scheme (TimeML Working Group, 2010), categories for event classification were defined. The inspiration on Setzer's classification is undeniable, but there is also a significant difference: in TimeML states were considered events, since they could be placed in time. Thus, the following event classification was proposed for TimeML:

- **Occurrence**: dynamic events that happen or occur.

- **State**: events describing circumstances in which something obtains or holds true and do not vary over time.

- **Reporting**: events that describe the utterance, narration, description, etc. of an event.

- **Aspectual**: events that indicate the beginning, continuity or end of an event.

- **Perception**: events that describe the physical perception of another event.

- **Intensional action**: dynamic events that select for an event-denoting argument which is explicitly in the text.

- **Intensional state**: states that, as intensional actions do, select for an event-denoting argument which is explicitly in the text.

This categorisation fulfilled two major objectives. First it served to determine whether an event could be taken as an argument for another event. Secondly, the difference between dynamic and stative events offered a preliminary view on how events happened in time, that is to say, whether they were punctual or lasted through a period of time.

The categorisation presented in this section was intended to cover all the different event types and gave relevant semantic information of them. The different events described in this section can be represented by means of the expressions described in Section 3.1.1.

2.2.2 Aspect and Tense Theories

For aspect and tense analysis, Reichenbach (1947) presented a thorough study on the verbal tempus and the relation between the verb forms and the time of speech. As he noted, the time of speech (S), the time in which the narrated event took place (E) and the reference time (R)—the time the speaker took as a reference—conditioned the choice of the verb tense. He proposed a scheme for English verb tenses based on that idea (Figure 2.1).

Structure	Reichenbach's name	Traditional English
E-R-S	Anterior past	Past perfect (or pluperfect)
E,R-S	Simple past	Simple past
R-E-S R-S,E R-S-E	Posterior past	-
E-S,R	Anterior present	Present perfect
S,R,E	Simple present	Present
S,R-E	Posterior present	Simple future
S-E-R S,E-R E-S-R	Anterior future	Future perfect
S-R,E	Simple future	Simple future
S-R-E	Posterior future	-

2.1 Figure – Reichenbach's temporal intervals

2.2.3 Theoretical Approaches to Time

There are many formalisms to express time. However, we will compare two points of view: instant-based theories and interval-based theories. In instant-based models of time, the primitive temporal entities are time instants and the basic relationship between them (besides equality) is precedence (Goranko and Galton, 2015). As a consequence, instant-based theories are not suitable for real-world event expressions as it does not admit instant overlapping.

Interval-based temporal models are ontologically richer than instant-based ones, as there are many more possible relationships between time intervals than between time instants (Goranko and Galton, 2015). Intervals have begin and end points and there is no interruption between those. Moreover, that kind of representation allows expressing time points too: time-points are intervals for which begin and end points are the same.

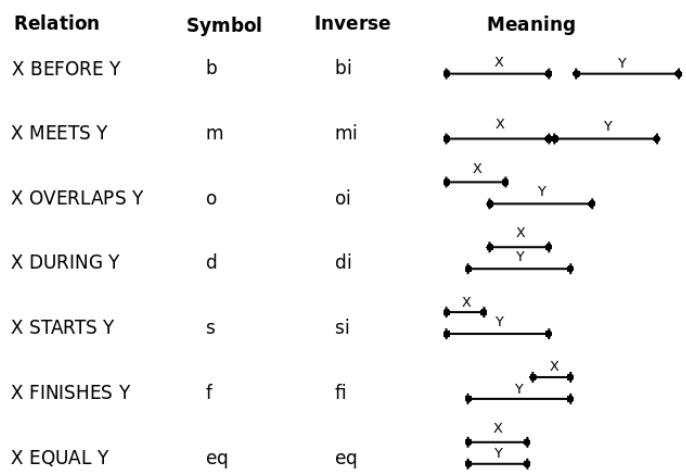
Setzer supports interval-based representation for two reasons:

- Humans are not empirically able to discern between points and very short intervals.
- Using time points is a completely arbitrary decision (Galton, 1990).

Additionally, according to Setzer, a time point can always be represented as the intersection of two intervals. When interpreting time as intervals Allen's interval theory (Allen, 1983) is possible.

2.2.4 Allen's Interval Theory

Allen's Interval Theory (1983) has also been a major theoretical foundation for temporal information analysis. His goal was to characterise the temporal inferences humans do. The model was composed of temporal intervals as humans only perceive time if there is a change in the world. In addition, Allen defended time could be divided into smaller units as well as eventualities could be decomposed in simpler ones. He proposed a set of 13 temporal relations, six and their opposites plus an identity relation (see Figure 2.2), which expressed all the possible relationships between pairs of intervals (and their corresponding eventualities).



2.2 Figure – Allen's temporal intervals[2]

This information on event typology, event temporal features and chronology has been the basis for temporal information processing in the recent decades. Nevertheless, a further step has to be taken in order to be able to process this information automatically: making that information machine-readable. In the following sections we present the main resources created for

[2]http://stackoverflow.com/questions/12069082/allens-interval-algebra-operations-in-sql

that task.

2.3 Mark-up Languages and Manual Annotation Tools

In order to make the temporal information machine-readable, the tokens —the sequences of contiguous characters delimited by spaces or spaces and punctuation marks— that convey temporal information have to be annotated and their features extracted. These tokens may express different temporal structures, namely, time expressions, events, signals and temporal relations, and mark-up languages have been developed for annotating those. In this section we present the mark-up languages that have been created to structure and normalise temporal information (Section 2.3.1), as well as the tools employed in the manual annotation of temporal information (Section 2.3.2).

2.3.1 Mark-up Languages for Temporal Information

Mark-up languages for temporal information assign a tag to those constructions and provide a set of attributes and values to describe their features. In the beginning only time expressions and some of their main features were addressed, but mark-up languages soon grew in complexity and now cover most of the aspects related to temporal information.

A first temporal mark-up language was developed for the MUC (Message Understanding Conferences) (Grishman and Sundheim, 1996) based on SGML[3]. From MUC-1 to MUC-5, entities and events were identified. In MUC-6 (1996), instead, temporal information was annotated, as being able to identify the moment or time span in which an eventuality happened was very important in the automatic comprehension of texts.

Research on temporal information has augmented since then. Mani and Wilson (2000) first proposed the TIMEX tag for those expressions and a temporal normalisation scheme based on the ISO-8601 standard (ISO Commitee, 1997). The tag had some attributes such as `VAL` (value) and `POS` (part-of-speech) that made explicit the features of the time expressions.

TIMEX2 tag (Ferro *et al.*, 2003) was proposed to annotate time expressions at the TIDES (Translingual Information Detection, Extraction and

[3]Standard Generalized Markup Language

Summarisation) programme (Cieri and Liberman, 2002). This tag had different attributes such as `VAL` (value), `MOD` (modificators), `SET`, `PERIODICITY` and `GRANULARITY`. Setzer (2001) also used this tag, but she went further and proposed the first complete annotation scheme for temporal information. She proposed a tag for events and for temporal relations and provided a set of attributes and possible values for those to illustrate the features of the events and relations. Temporal annotation with TIMEX2 was also done for other languages as Korean (Jang *et al.*, 2004) or German (Strötgen and Gertz, 2011).

TimeML (Pustejovsky *et al.*, 2003a) differed from the previous mark-up proposals: it i) anchored events to temporal expressions, ii) ordered the event expressions relative to one another and iii) allowed delayed interpretation to underspecified temporal expressions. It also extended the features of its forerunners: it widened the attribute range of tags and the event classification, allowed the processing of relative time expressions and proposed relations between events and times. TimeML was first proposed along with the TimeBank corpus (Pustejovsky *et al.*, 2003b) for the creation of a question answering system that would address temporal questions.

Since then, TimeML has improved. It has become an ISO International Standard (ISO-TimeML working group, 2008) and the latest version (TimeML Working Group, 2010) is a reliable mark-up language for temporal annotation. We describe TimeML more in depth in Section 4.1.1.

Some other mark-up languages were later created inspired in TimeML. SIBILA (*Sistema automático de respuestas basado en un modelo de discurso*) (Wonsever *et al.*, 2012) was a mark-up language that centred on events and their factuality. TimeML put the strength in resolving when an event took place, whereas SIBILA focused on whether one could confirm the event had happened.

PLIMEX (Kocoń *et al.*, 2015) has also been based on TimeML. More precisely, it is a combination of the adaptations of the TIDES TIMEX2 and TimeML TIMEX3 tags for time expressions for Polish. Time expressions get the same types as in TimeML: date, time, duration and set. The same authors proposed also an event identification and classification scheme (Marcińczuk *et al.*, 2015), that followed the TimeML event classification.

For NewsReader project[4] (Vossen *et al.*, 2014), TimeML was adapted for temporal annotation (Tonelli *et al.*, 2014). Although the main tags were

[4]`http://www.newsreader-project.eu/`

preserved, some of their features were modified for the task and causal and grammatical relations were added. Temporal annotation was also combined with event and entity co-reference and semantic role information.

TEMANTEX (Wonsever *et al.*, 2015) is a temporal annotation scheme that annotates temporal expressions as well as other lexical indicators of temporality, which they call "temporal indicators", "elements that influence the temporal interpretation of the text". Although it is mainly compatible with TimeML, it does not follow the ISO-8601 rule (ISO Commitee, 1997) and it offers a different classification for temporal expressions and attributes.

2.3 Table – Mark-up languages for temporal information

Mark-up language	**Annotated entities**	**First project**
TIMEX (Mani and Wilson, 2000)	Time expressions	MUC conferences
TIMEX2 (Ferro *et al.*, 2003) and (Setzer, 2001)	Time expressions	TIDES (Cieri and Liberman, 2002)
TimeML (Pustejovsky *et al.*, 2003a)	Events, time expressions, signals and temporal, aspectual and subordination relations	TimeBank
SIBILA (Wonsever *et al.*, 2012)	Events and their factuality	SIBILA
PLIMEX (Kocoń *et al.*, 2015)	Time expressions (and events)	
NewsReader (Tonelli *et al.*, 2014)	Events, time expressions, signals and temporal, grammatical, causal and subordination relations	NewsReader (Vossen *et al.*, 2014)
TEMANTEX (Wonsever *et al.*, 2015)	Time expressions and temporal indicators	

A summary of the mark-up languages presented above can be seen in Table 2.3. The corpora presented in Section 2.5 have been annotated following these mark-up schemes.

2.3.2 Manual Annotation Tools

Gold standard corpora are crucial in NLP for they offer high quality annotations for system development, training and evaluation purposes. Temporal information processing is an area in which corpora have a special relevance as events and time expressions can only be fully interpreted according to their contexts. More precisely, the relations between events and time expressions are of utmost importance in this field, as temporal ordering is deduced from these relations. Hence, not only the temporal constructions have to be taken into account but also their contexts. Thus, not only the temporal constructions have to be annotated but also the relations created among them have to be marked.

It is widely reckoned that best quality corpora is achieved by manual annotation in which human annotators follow certain annotation guidelines. Nonetheless, as corpora are understood as text collections in digital format, annotations are added by means of annotation software. These tools offer an interface in which annotators manually add linguistic information to documents. These tools also normalise and make machine-readable those annotations. Commonly a set of tags and attributes are defined within the tool and the annotators select the portion of text they want to annotate and add a certain tag to it. Apart from the basic manual annotation interface in which an annotation scheme is loaded and tags and attribute values are added manually, the many different software offer some specific features.

The first annotation tool employed in the creation of temporally annotated corpora was the Alembic Workbench (Day *et al.*, 1997). It was specifically developed for the task in the contexts of the MUC conferences and it allowed combining automatic and human-made annotation. In fact, it followed a bootstrapping procedure by means of which the first manual annotation was reused to create rules from it and automatically annotate the following documents. As a consequence, the annotation effort gradually switched from manual annotation to revision. This approach hugely accelerated the corpus building effort. Additionally, the tool was engineered to reduce the hand motion so as to make manual annotation fast.

Not only was it a hybrid tool that combined manual and rule-based annotation, but it also contained a scorer that allowed arbitrary SGML mark-up to be selected for scoring. It was freely distributed and it was used beyond the MUC conferences; namely, it was used in the Schilder and Habel (2001) experiment, but it is not available any more[5].

Callisto (Day *et al.*, 2004) is a manual annotation tool that was focused at modularity and configurability. It provided a Model-View-Controller design that ensured the user interface components' independence, while always displaying consistent information. That is to say, any change to the annotation was propagated to the graphical components to update the view displayed to the annotator. Task-specific models were directly implemented in the source code, which allowed the incorporation of task-specific logics and graphical user interfaces.

Callisto was created in the ATLAS[6] context (Bird *et al.*, 2000) that aimed to provide abstraction over the linguistic signals and their annotations. It was built as a multilingual annotation tool since Java handled excellently the Unicode character encodings, so the best font for each was automatically chosen by the tool. Moreover, as Callisto was conceived as a Java tool, tasks could be saved and run independently. In addition, new tasks were not to be created from scratch as similar tasks could be built through little modifications of the existing ones. Finally, Callisto could handle the XML[7]-based Atlas Interchange Format (AIF) as well as other SGML or XML-based markups. Hence, the tool was widely used in temporal information annotation and some related tasks were defined such as TIMEX2 normalisation, TimeML tagging and event annotation. The tool is open-source and currently freely downloadable[8].

Tango (Verhagen *et al.*, 2006) was employed by the creators in temporal relation annotation as well as in the creation of the French TimeBank. Comparing to the existing annotation tools, Tango offered a graph visualisation of temporal information in which events and time expressions were nodes and the temporal relations among them were arcs, whereas the other annotation tools offered a table representation. The nodes could be placed anywhere in the display and could be arranged to create a visual timeline. It also contained an assistant that suggested possible relations in order to augment the manually annotated links and a T-BOX representation, by means of which temporal entities were represented as boxes. In what concerns relation representation, box inclusion means temporal inclusion and stacked boxes imply simultaneity.

[5] `http://annotation.exmaralda.org/index.php?title=Alembic_Workbench`

[6] Architecture and Tools for Linguistic Analysis Systems

[7] Extensible mark-up language

[8] `https://github.com/mitre/callisto`

Tango was run in the Callisto environment as a specific task. While the original task in Callisto served for event, time expression and signal annotation, the Tango task provided the environment for temporal, subordination and aspectual relation annotation too. The Tango task is freely available in `http://timeml.org/tango/tool.html`.

The Brandeis Annotation Tool (BAT) (Verhagen, 2010) was a tool based on annotation layers and task decomposition. These made merging different annotation tasks as well as splitting complex tasks into more simple subtasks easy. For example, it kept the extent and attribute layers separated and, thus, representing markable extents acquired from other sources became easier.

While it did not seem to happen in the previously presented annotation tools, the role of the administrator was crucial in BAT. The administrator could create and edit corpora, layers and tasks, could assign annotators to specific tasks and could monitor the annotators' progress.

BAT was first used for the Tempeval evaluation task since textual annotation was required in a very short period of time. In what concerns technical information, BAT is based on PHP[9] and JavaScript and data is stored in MySQL databases. It was web-based, which made handling updates and upgrades much easier. Unfortunately, it is not available any more.

Content Annotation Tool[10] (CAT) (Lenzi *et al.*, 2012) is a web-based annotation tool for linguistic and semantic annotation. It can deal with multi-layer annotation, which allows the combination of different annotations of a single document. Its main strengths are the intuitiveness and customisation capabilities. Moreover, it provides searching and agreement measuring facilities, as well as a tokeniser, so it accepts both raw text and tokenised text files as input. CAT produces XML files as output and DTDs[11] are automatically generated in order to achieve a consistent annotation. In the interface, the annotations are displayed in text and the relations are represented in table format.

Annotation efforts are often very task-related and integrating an annotation scheme in a tool should be easy. CAT does not require any programming skills and annotation schemes can be easily created through the interface. CAT has been employed in many tasks such as the creation of the Italian TimeBank, the EVENTI task at EVALITA 2014 and for textual annotation

[9]PHP: Hypertext Preprocessor

[10]Previously CELCT Annotation Tool

[11]Document Type Declaration

2.4 Table – Manual tools for temporal information annotations

Annotation tool	**Uses**	**Manual vs. auto**	**Mark-up language**	**Link representation**	**Motor**	**Availability**
Alembic Workbench	MUC conferences Schilder and Habel's work	Manual and rule-based bootstrapping	SGML	Table	Sun	Free Not available (2009)
Callisto	TimeML TIMEX2 Events	Manual	SGML XML AIF	Table	Java	Open source Freely downloadable
Tango	TimeML French TimeBank	Manual + assistance	XML (TimeML)	Graph	Java (Callisto task)	Free
Brandeis Annotation Tool	TempEval	Manual	TimeML XML	Table	PHP Javascript	Web-based Not available
Content Annotation Tool	Ita-TimeBank EVENTI TERENCE Excitement NewsReader	Manual	XML	Table		Web-based Free (under request)
Inforex	KPWr corpus Polish TimeML	Manual	XML		JavaScript	Web-based authorised and public access

in the TERENCE[12], Excitement[13] and NewsReader projects. It is freely available[14], even if a user profile has to be requested in advance.

The Inforex annotation tool (Marcińczuk *et al.*, 2017) is a JavaScript-based web annotation tool that can handle XML documents and XML-based mark-up languages. It is language independent and annotation tasks can be easily created. It has been employed in several tasks for manual annotation of Polish texts. In what concerns temporal information annotation, Inforex has been used for the manual annotation of Polish time expressions (Kocoń *et al.*, 2015) and events (Marcińczuk *et al.*, 2015). In those experiments, temporal information in Polish has been added as an extra layer to the KPWr corpus (Broda *et al.*, 2012) following a TimeML-inspired scheme.

The aforementioned annotation tools are summarised in Table 2.4.

2.4 Temporal Information Processing Systems

Automatic processing systems use available linguistic information in corpora and algorithms to extract the temporal information of texts. Automatic processing tools can be classified in four groups according to their nature: i) some are rule-based, ii) others are statistical and iii) there is a third group

[12]http://vittorini.univaq.it/?page_id=869
[13]https://sites.google.com/site/excitementproject/home
[14]http://celct.fbk.eu:8080/CAT_WEB_APP/checkUsers

formed by the hybrid tools. Finally, iv) there have also been some experiments based on deep learning methods. These automatic tools can also be sorted according to their role on the temporal information extraction process. Automatic tools can be used for e. g. temporal expression and event extraction, temporal annotation and temporal relation extraction.

2.4.1 Rule-based Tools

There are several rule-based systems for temporal information extraction. These kind of systems were specially developed in the early stages of temporal information processing due to the scarcity of annotated data at the time. First experiments were conducted on time stamping and later on resolution of the relative time expressions such as *now*, *today* or *in 5 minutes*. Mani and Wilson (2000) developed TempEX, a tagger capable of extracting and annotating what they called dates, times and temporal expressions. The tool first solved self-contained expressions and secondly the context-dependent ones. The software used syntactic clues to assign a value to each expression.

Schilder and Habel (2001) used Finite State Transducers (FST) to extract temporal information. First, the documents went through a pre-processing stage in which they were syntactically annotated and their lemmas extracted. Then, FSTs specialised in different temporal expressions tagged the resulting chunks as time expressions, events (verbs) and nouns expressing events. Time expressions were also given a normalised value based on an indexical time (document creation time).

FSTs were also employed at the TERQAS project to extract events and time expressions, while temporal relations were extracted using a small set of heuristics. The Alembic Workbench (Day *et al.*, 2004) was used for the task as it allows combining automatic and human-made annotation.

KTX (Korean Temporal eXpression tagger) (Jang *et al.*, 2004) used a rote learning method based on the induction of a dictionary from training data, which was later upgraded with human-made patterns; a total of 460. Temporal expressions were then found using morphological analysis and a stop word list. KTX had also a rule module for the disambiguation of absolute and relative temporal expressions from durations.

CTEMP (Wu *et al.*, 2005) was a temporal parser which consisted of two modules, one for temporal extraction and a second one for temporal normalisation. It was based on chart parsing, context-free grammar and constraint rules. The parsing provided each identified time expression with some attributes that were employed afterwards for the normalisation of time expressions.

The two-module architecture for time expression extraction and normalisation was also employed in DANTE (Mazur and Dale, 2007). This rule-based system performed time expression recognition based on a JAPE Java-based grammar that contained 31 gazetteers, 80 macros and 250 rules. Rules were traditional pattern-action rules in which the first part of the rule designed the textual pattern to be matched and the second the normalisation pattern to be assigned. DANTE identified and normalised time expressions following the TIMEX2 standard.

Other three rule-based temporal parsers are the ones created by Bittar (2009), Robaldo *et al.* (2011) and Edinburgh-LTG (Grover *et al.*, 2010). These three were used to extract events and their information. The first one was developed for French and combined rules with specific hand-made dictionaries for event detection and classification. The second one used dependencies and an event list to automatically identify and classify events in Italian. Edinburgh-LTG merged lists extracted from previously annotated corpora with WordNet information.

HeidelTime (Strötgen and Gertz, 2013) is a rule-based system for time expression extraction and normalisation. The rules, patterns and normalisation information are language dependent, while the source code is common to all languages. HeidelTime has been built as an UIMA component so it can be integrated in UIMA pipelines. HeidelTime identifies the elements in the time expressions matching the tokens with the elements in the regular expressions defined (weekdays, months, numbers, parts of teh day, etc.). Then the time expressions are identified by the rules and are given a type value (date, time, duration or set) a normalised value. Time expressions are annotated using the TimeML `TIMEX3` tag. Apart from English, HeidelTime has been used for temporal expression extraction and normalisation in German (Strötgen and Gertz, 2011), Dutch (van de Camp and Christiansen, 2013), French (Moriceau and Tannier, 2014) and Croatian (Skukan *et al.*, 2014) among others.

Finally, ParsTime (Mansouri *et al.*, 2018) is a rule-based time expression extraction and normalisation system for Farsi that detects date patterns on previously temporally annotated texts. That is to say, tokens in texts are given one of the 12 predefined tags (day, month, season, number, etc.) and rules to identify those patterns are created. 346 rules have been defined and these assign a TimeML `TIMEX3` tag to each time expression as well as a type value (date, time, duration or set) a normalised value. A special feature of ParsTime is its ability to identify and normalise not only time expressions based on the Gregorian calendar, but also on the Hijri and Jalali calendars.

Rule-based temporal information annotation tools are summarised in Table 2.5.

2.5 Table – Rule-based tools for temporal information processing

Tool	**Developers**	**Task**	**Kind**
TempEX	Mani and Wilson (2000)	Extract and annotate "dates, times and temporal expressions"	Rule-based (syntactic clues)
	Schilder and Habel (2001)	Temporal information extraction	FST (Finite State Transducers)
Alembic Workbench	Day *et al.* (2004)	Extract events and time expressions	FST
KTX	Jang *et al.* (2004)	Temporal extraction and normalisation	rote learning + patterns + rules
CTEMP	Wu *et al.* (2005)	Extraction and normalisation	Rule-based
DANTE	Mazur and Dale (2007)	Time expression extraction and normalisation	Rule-based
	Bittar (2009)	Event extraction and classification	Rule-based + handmade dictionaries
	Robaldo *et al.* (2011)	Event extraction and classification	Rule-based. Dependencies and event lists
Edinburgh-LTG	Grover *et al.* (2010)	Event extraction and classification	Rule-based. Lists and WordNet
HeidelTime	Strötgen and Gertz (2013)	Time expression extraction and normalisation	Rule-based
ParsTime	Mansouri *et al.* (2018)	Time expression extraction and normalisation	Rule-based

2.4.2 Statistical Tools

Other automatic tools apply statistical methods. Bethard and Martin (2006) propose a multi-class classification method that gives every token an "inside", "outside" or "begin" tag depending on whether they are events. March and

Baldwin (2008), instead, propose an event extraction and classification algorithm based on Support Vector Machines (Boser *et al.*, 1992).

TIPSem (Temporal Information Processing based on Semantic information) (Llorens *et al.*, 2010) is a system for temporal structure recognition, classification, normalization and link-categorization and its main feature is the use of semantic information to fulfil those four tasks. The system is based on Conditional Random Fields (Lafferty *et al.*, 2001), a machine-learning technique.

ClearTK-TimeML (Bethard, 2013) follows the work done on Bethard and Martin (2006). It is based on the ClearTK framework and provides machine-learning classification for the identification of events, time expressions and temporal relations. It applies Conditional Random Fields, Support Vector Machines and Logistic Regression.

Kocoń and Marcińczuk (2017) have developed a time expression extraction and normalisation tool for Polish. It is based on the Liner2 tool (Marcińczuk *et al.*, 2013) which employs the CRF++ toolkit. It follows a wrapper approach to select the most suitable linguistic features.

Murat *et al.* (2017) present a CRF-based time expression recognition system for Uyghur. They first conduct lemmatisation and word-segmentation and then apply five feature sets sequentially: token, POS, number (to identify numerical time expressions), character (to identify time expressions containing special characters) and temporal lexical triggers.

TOMN (Zhong and Cambria, 2018) is a machine-learning method that models time expressions. It identifies time expressions under CRFs, giving TOMN tags to the tokens in texts. In fact, TOMN identifies time tokens (T), modifiers (M), numerals (N) and assigns an *out* (O) tag to the rest of the tokens. Time tokens, modifiers and numerals are gathered in TmnRegex regular expressions that are matched during the preprocessing feature extraction stage.

We have summarised the machine-learning-based temporal annotation tools in Table 2.6.

2.6 Table – Automatic tools for temporal information processing

Tool	**Developers**	**Task**	**Kind**
	Bethard and Martin (2006)	Event extraction and classification	Multi-class classification
	March and Baldwin (2008)	Event extraction and classification	Support Vector Machines

TIPSem	Llorens *et al.* (2010)	recognition, classification, normalization and link-categorization	Data-driven system. CRF (Conditional Random Fields) models
ClearTK-TimeML	(Bethard, 2013)	Temporal information extraction	Conditional Random Fields, Support Vector Machines and Logistic Regression
	Kocoń and Marcińczuk (2017)	Time expression extraction	Conditional Random Fields
	Murat *et al.* (2017)	Time expression extraction	Conditional Random Fields
TOMN	Zhong and Cambria (2018)	Time expression extraction	Conditional Random Fields

2.4.3 Hybrid Tools

As one can see the rule-based vs statistical division is a suitable one. However, hybrid methods have also been developed. Verhagen *et al.* (2005) present in their work some partial parsers like GUTime, GUTenLINK or Evita as parts of the TARSQI[15] Toolkit Verhagen and Pustejovsky (2008). These parsers act like modules and solve different annotation steps. GUTime is a time expression extraction and normalisation tool, Evita (Events in Text Analyzer) is an event recognition tool that recognizes verbs, nouns and adjectives expressing events and extracts their grammatical features. Verb annotation is based on lexical look-up, noun annotation, on machine learning and adjectives are only annotated if they are events on TimeBank.

In the TARSQI system, for the annotation of relations GUTenLINK, Slinket and Sputlink have been developed. The first is a robust temporal relation tagger based on FSTs to create relations between events and time expressions and rules to order the events. The second creates links between the main verb and its subordinate event based on linguistic-based knowledge and rules which are implemented in FSTs. The third allows making explicit implicit temporal relations following some constraints and Allen's interval algebra (Allen, 1983).

[15]Temporal Awareness and Reasoning Systems for Question Interpretation

TRIOS (UzZaman and Allen, 2010) first apply some rules for event detection and the uses Markov Logic Network (MLN) for classification in order to improve the results.

Language Independent Feature Extractor (LIFE) (Jeong *et al.*, 2015) has also been used to annotate temporal structures and their features. LIFE uses CRF models and is based on letters rather than words and analyses texts at letter-level. It uses a set of algorithms and its output can be enhanced with hand-written rules in order to get a fuller temporal information analysis and feature extraction. This system has recently been improved in ExoTime (Jeong *et al.*, 2017), which employs and external knowledge-base to improve time expression normalisation.

2.4.4 Deep Learning Tools

Recently, there have been attempts on temporal information extraction and classification using deep learning techniques. More precisely, Etcheverry and Wonsever (2017) present a time expression recognition system based on distributed representations and artificial neural networks. They used word representations inferred from Spanish Wikipedia through GloVe (Pennington *et al.*, 2014). They followed TimeML classification for time expressions. For the modelling, they considered feedforward and recurrent models with different numbers, size of hidden layers and regularization techniques and included local contextual information concatenating vector representation of fixed size window to both sides from the word to label. A similar approach has been followed in Gupta *et al.* (2015) for time expression recognition in English clinical texts.

The hybrid systems and the neural network experiments have been summarised in Table 2.7.

2.7 Table – Automatic tools for temporal information processing

Tool	**Developers**	**Task**	**Kind**
GUTime	Verhagen *et al.* (2005)	Time expression extraction and normalisation	
GUTenLINK, Slinket and Sputlink	Verhagen *et al.* (2005)	Temporal and subordination link creation	FST
Evita	Verhagen *et al.* (2005)	Event recognition and feature extraction	lexical look-up + machine learning

TRIOS	UzZaman and Allen (2010)	Event extraction and classification	Combined: rules + Markov Logic Network
LIFE and ExoTime	Jeong *et al.* (2015, 2017)	Temporal information extraction and classification	Conditional Random Fields + hand-written rules
	Etcheverry and Wonsever (2017)	Time expression extraction and classification	Distributed representations and neural networks

2.5 Corpora

A corpus for Natural Language Processing (NLP) is nowadays understood as "a body of text which exists in electronic form and which can be processed by a computer, used as part of linguistic research and language processing" (Setzer, 2001, p. 64), that is to say, a text collection that contains some kind of linguistic annotation and is machine-readable. Corpora have been created for many languages and contain different kinds of texts. In the case of corpora that contain temporal information, the elements that express temporal information are annotated and their features extracted.

The corpora that contain temporal information are mainly used as gold standards to train and evaluate temporal information processing tools for automatic textual comprehension. Adding temporal information to the existing information in the corpora is also a general goal; such is the case of Catalan TimeBank (Saurí and Badia, 2012), Estonian TimeBank (Orasmaa, 2016) and the Chinese Temporal Annotation Corpus (Cheng *et al.*, 2007), which add temporal information to the previously existing morphosyntactic and lexical-semantic information, in order to have a corpus annotated in many levels.

Some corpora annotated with temporal information will be presented in this section along with the myriad of mark-up languages used to create temporal information corpora and the features of the corpora; for example, what kind of documents it contains or which annotation tool has been employed to annotate them. First, we present the early temporally annotated corpora in Subsection 2.5.1, the news corpora that contain temporal information in Subsection 2.5.2, and the corpora for other domains in Subsection 2.5.3.

2.5.1 First Temporally Annotated Corpora

As mentioned before, the first attempts of temporal information processing where made in the MUC conferences. These conferences proposed several information extraction tasks and corpora were created to evaluate the tools which took part in those tasks. For MUC-6 a corpus of 318 annotated Wall Street Journal articles was released for training and evaluation (Chinchor and Sundheim, 2003). This corpus used in the MUC-6 conference as well as some other early temporally annotated corpora are summarised in Table 2.8.

In the following years there were many experiments on temporal information extraction. Mani and Wilson (2000) developed a rule-based time expression identification algorithm and they evaluated its performance using a manually annotated news gold standard corpus. Their test corpus consisted of 22 written news articles and 199 TV voice transcripts from the TDT-2 collection(Fiscus *et al.*, 1999).

Following the MUC conferences promoted by DARPA (Defense Advanced Research Projects Agency), the TIDES project was developed. The "TIMEX2 standard" (Ferro *et al.*, 2003) was created within this temporal expression annotation project. This standard conferred the possibility of annotating time expressions and their attributes with the SGML[16] mark-up language. Along with this annotation scheme, the TIDES temporal corpus (Ferro *et al.*, 2002) was also created. This corpus contained time expressions as well as temporal relations. A modification of the Alembic Workbench (Day *et al.*, 1997) was employed for the task and annotation of the corpus was made manually. TIDES corpus was formed by 95 English-Spanish parallel conversation documents and 193 documents from TDT-2. The corpus was employed to evaluate the tools developed within the TIDES programme.

At the same period, Setzer (2001) made a proposal for the annotation of temporal expressions alongside eventualities and temporal relations using SGML. Setzer presented the tags and their attributes to make explicit the temporal information in texts and developed an annotation tool based on Perl programming language. In order to evaluate the adequacy of the guidelines, the inter-annotator agreement on temporal structures and the annotation effort, a corpus of 6 MUC-7 news texts was created. Despite the small size of the corpus, many decisions taken in Setzer's research were crucial in the development of TimeML (Pustejovsky *et al.*, 2003a).

[16]Standard Generalized Markup Language

2.8 Table – First temporally annotated corpora

Corpus	Language	Composition	Mark-up language	Annotation tool	Manual vs. auto
MUC-6 (Chinchor and Sundheim, 2003)	English	318 Wall Street journal articles	SGML		Auto
(Mani and Wilson, 2000)	English	22 written articles and 199 voice transcripts from TDT-2	TIMEX		Manual
Tides Temporal Corpus (Ferro *et al.*, 2002)	English Spanish	95 conversations (Spanish) + translations (English) and 193 documents of the TDT-2 corpus	TIMEX2	Alembic Workbench	Manual
(Setzer, 2001)	English	6 MUC-7 articles	SGML	Perl-based annotation tool	Manual

2.5.2 Temporally Annotated News Corpora

In the last decades, temporal information processing has strongly focused on news documents and, as a consequence, various news corpora have been annotated with temporal information. TimeML became the most used temporal mark-up language and it has been considered a standard for temporal annotation. The corpora TimeBank 1.1 (Pustejovsky *et al.*, 2003b) and TimeBank 1.2 (Pustejovsky *et al.*, 2006) were annotated with this mark-up language.

The TimeBank corpora were created in the TERQAS[17] programme for the training and evaluation of question answering (QA) systems, but they have been used for different purposes since then. They have been used in the first TempEval (Verhagen *et al.*, 2007) tasks and TimeBank 1.2 has become the basis for English temporal structure analysis as well as for the development of corpora in other languages. Its Spanish version is considered its twin corpus, but it has also been translated or has served as a reference for the creation of corpora for French (Bittar, 2010), Italian (Caselli *et al.*, 2011),

[17]Temporal and Event Recognition for QA Systems

Romanian (Forăscu and Tufiş, 2012), Portuguese (Costa and Branco, 2012), Catalan (Saurí and Badia, 2012), Farsi (Yaghoobzadeh *et al.*, 2012), Japanese (Asahara *et al.*, 2013) and Korean (Jeong *et al.*, 2015) until date.

A version that contains denser relation annotation also exists for TimeBank: TimeBank-Dense (Cassidy *et al.*, 2014). This corpus contained 36 randomly chosen TimeBank documents that were already annotated following TimeML. The annotation was manually enriched by the TimeML Dense annotation and the annotated documents contained much more describing relation graphs.

French TimeBank (Bittar, 2010) was intended to be comparable to TimeBank 1.2 corpus and that is the reason for it being formed of 109 news documents in French and annotated with TimeML. Documents were chosen to be balanced and covered many sub-genres of journalistic literature. Due to the big amount of data, a semi-automatic pre-annotation strategy was followed. In the automatic stage, text was preprocessed and temporal expressions annotated with TempEx tagger (Mani and Wilson, 2000). Then, some trained annotators supervised and corrected the automatic work using the Callisto (Day *et al.*, 2004) and the Tango (Verhagen *et al.*, 2006) annotation tools.

Italian TimeBank (Ita-TimeBank) (Caselli *et al.*, 2011) was built from the fusion of the CELCT corpus[18] and the ILC corpus (Bindi *et al.*, 1989). These news corpora were annotated following TimeML specifications for Italian (Caselli, 2010) and they were created to be gold standards for temporal annotation in Italian. The CELCT corpus was annotated using the Brandeis Annotation Tool (BAT) (Verhagen, 2010) as well as the CELCT Annotation Tool (CAT) (Lenzi *et al.*, 2012), whereas only BAT was used to annotate the ILC corpus. Both corpora were annotated manually.

The Romanian (Forăscu and Tufiş, 2012) and Portuguese (Costa and Branco, 2012) TimeBanks were created by translating the English version. The annotation in each language, though, was carried out differently. For the Romanian TimeBank the TimeML tags were transferred after token alignment was done, while TimeBankPT was manually annotated.

The Catalan TimeBank contained 210 documents (over 75,800 tokens) extracted from the Ancora-Ca corpus (Taulé *et al.*, 2008). It was automatically annotated with TIPSem (Llorens *et al.*, 2010) following the Catalan version of TimeML.

The Persian TimeBank (Yaghoobzadeh *et al.*, 2012) is a 43 document

[18] `http://www.livememories.org/Home.aspx`

corpus for Farsi. The documents were extracted from de Peykareh corpus (Bijankhan *et al.*, 2011) and semi-automatically annotated. The events were annotated with the Persian Event Tagger (PET) (Yaghoobzadeh *et al.*, 2012) and the manual supervision was done with the Multipurpose Annotation Environment (MAE)[19].

The BCCWJ-TimeBank (Asahara *et al.*, 2013) is a 1.3 million-word corpus for Japanese annotated following TimeML. The corpus was annotated manually with XML Editor oXygen[20] which provided help with XML validation.

The TimeBank-styled corpora have been summarised in Table 2.9.

2.9 Table – TimeBank-styled temporally annotated news corpora

Corpus	**Language**	**Composition**	**Mark-up language**	**Annotation tool**	**Manual vs. auto**
TimeBank 1.1, TimeBank 1.2 and TimeBank QA Corpus (Pustejovsky *et al.*, 2003b) and (Pustejovsky *et al.*, 2006)	English	News texts (183 documents)	TimeML	Alembic Workbench (Day *et al.*, 1997)	Manual
TimeBank-Dense (Cassidy *et al.*, 2014)	English	183 documents from TimeBank	TimeML-Dense		Manual
French TimeBank (Bittar, 2010)	French	109 news documents	TimeML	TempEx tagger (Mani and Wilson, 2000), Callisto (Day *et al.*, 2004) and Tango (Verhagen *et al.*, 2005)	Semi-automatic

[19] `https://code.google.com/p/mae-annotation/`
[20] http://www.oxygenxml.com/

Ita-TimeBank (Caselli *et al.*, 2011)	Italian	CELCT Corpus + ILC Corpus	TimeML		Manual
CELCT Corpus[1]	Italian	News from Italian Content Annotation Bank corpus	TimeML	Brandeis Annotation Tool (BAT) (Verhagen, 2010) and CELCT Annotation Tool (CAT) (Lenzi *et al.*, 2012)	Manual
ILC Corpus (Bindi *et al.*, 1989)	Italian	171 newspaper stories	TimeML	Brandeis Annotation Tool (BAT) (Verhagen, 2010)	Manual
Romanian TimeBank (Forăscu and Tufiş, 2012)	Romanian (translated from English)	183 news documents	ISO-TimeML		Automatic
TimeBankPT (Costa and Branco, 2012)	Portuguese (translated TempEval corpus)	182 documents	TimeML		Manual
Catalan TimeBank (Saurí and Badia, 2012)	Catalan	210 documents (news and fiction) from the Ancora-Ca corpus (Taulé *et al.*, 2008)	TimeML	TIPSem (Llorens *et al.*, 2010)	Automatic
PersTimeBank (Yaghoobzadeh *et al.*, 2012)	Farsi	43 texts from Peykareh corpus (news of many subjects)	ISO-TimeML	Persian Event Tagger (PET) and Multipurpose Annotation Environement (MAE)[2]	Semi-automatic

BCCWJ-TimeBank (Asahara *et al.*, 2013)	Japanese	1.3 million words, many genres	ISO-TimeML	XML Editor oXygen[3]	Semi-automatic

Alongside those corpora, a corpus for Chinese, the Chinese Temporal Annotation Corpus (Cheng *et al.*, 2007), was created. It contained temporal information, but the main difference with the previously mentioned corpora is that the annotation effort was backed by dependency information in order to reduce the effort of the annotators. The annotated corpus was the 10% of the Penn Chinese TreeBank corpus (Xue *et al.*, 2005), which was translated to dependency structures and annotated following TimeML.

Li *et al.* (2014) also annotated 105 documents of the NIST[21] open MT 2008 evaluation corpus following TimeML. The corpus contained 1,357 pairs of Chinese-English sentences. It was named Chinese TimeBank, although it differs from most of the TimeBanks for other languages. It contains annotations for events and time expressions as well as for the relations among those. It has been annotated using SUTime (Chang and Manning, 2012) and verified by a human annotator.

The Chronolines project (Bittar *et al.*, 2012) went a step further and presented a French and English news corpus which would be used for the creation of temporal information visualization tools. The corpus was annotated following a mark-up language based on and compatible with TimeML. The annotation process was human-made and the Glozz annotation tool (Widlöcher and Mathet, 2009) was employed for the task. The result of the annotation effort was a temporal annotation gold standard.

The TIDES project continued over the years and the ACE TERN corpus (Ferro *et al.*, 2010) was created in it. It was formed of 127 broadcast news and 65 news documents in English and it contained temporal expressions marked with TIMEX2 tags.

Orasmaa (2014) used a 80 news articles (22,000 tokens) collection from Estonian newspapers to evaluate different event annotation models. These

[1] `http://www.livememories.org/Home.aspx`
[2] `https://code.google.com/p/mae-annotation/`
[3] `http://www.oxygenxml.com/`
[21] National Institute of Standards and Technology

articles were semi-automatically annotated by three annotators and a judge using Brandeis Annotation Tool (BAT) (Verhagen, 2010) and formed a gold standard for experimentation. The corpus contained morphological and dependency information used to extract relations between events. Temporal information was annotated following Estonian TimeML.

A corpus (Minard *et al.*, 2016) of 120 news documents from WikiNews was created as part of the NewsReader project (Agerri *et al.*, 2014). The main aim of the project was to build a multilingual system able to create storylines from news texts. As a consequence, those documents had event and temporal information annotated in order to be useful for the evaluation of storyline creation tools. Two main features make this corpus interesting: it is aligned at markable level for four languages (English, Spanish, Italian and Dutch) and it provides intertextual temporal and event annotation. It was manually annotated for each language using CAT (CELCT Annotation Tool) and CROMER (CROss-document Main Events and entities Recognition) (Girardi *et al.*, 2014).

The Event StoryLine Corpus (ESC) (Caselli and Vossen, 2017) is a 258 document set from the ECB+ corpus (Cybulska and Vossen, 2014) concerning calamity events. It contains event and time expression annotations as well as temporal relations among those and causal relations between events. The annotation scheme employed was TimeML and the causality information was added. It also contains event co-reference information. The corpus was annotated by two experts using CAT (Lenzi *et al.*, 2012).

These news corpora annotated with temporal information are summarised in Table 2.10.

2.10 Table – News corpora carrying temporal information

Corpus	**Language**	**Composition**	**Mark-up language**	**Annotation tool**	**Manual vs. auto**
Chinese Temporal Annotation Corpus (Cheng *et al.*, 2007)	Chinese	10% of Penn Chinese Tree-Bank (Xue *et al.*, 2005)	TimeML + dependencies		Automatic (dependencies) + manual

Chinese TimeBank (Li *et al.*, 2014)	Chinese-English	105 documents of the NIST open MT 2008 corpus. 1,357 pairs of Chinese-English sentences.	TimeML	SUTime (Chang and Manning, 2012)	Semi-automatic
Chronolines (Bittar *et al.*, 2012)	French	Written texts from AFP[4]	Based on TimeML	Glozz annotation tool (Widlöcher and Mathet, 2009)	Manual
ACE TERN (Ferro *et al.*, 2010)	English	127 broadcast news and 65 news text documents	TIMEX2[5]		Manual
(Orasmaa, 2014)	Estonian	80 news documents, 22,000 tokens	TimeML	Brandeis Annotation Tool (BAT) (Verhagen, 2010)	Semi-automatic
NewsReader (Minard *et al.*, 2016)	English, Spanish, Italian, Dutch	120 news documents	TimeML based mark-up language	CELCT Annotation Tool (CAT) (Lenzi *et al.*, 2012) and CROMER (Girardi *et al.*, 2014)	Manual
Event StoryLine (Caselli and Vossen, 2017)	English	258 news documents	TimeML and causality	CELCT Annotation Tool (CAT) (Lenzi *et al.*, 2012)	Manual

[4]Agence France Presse
[5]Only time expressions

2.5.3 Temporally Annotated Corpora for Other Domains

Even if most of the early research on temporal information has been conducted on the news domain, in the last decade the domain range has been widely enlarged. As a consequence, relevant corpora to each domain have been built. Corpora of different textual genres will be now described and will be summarised in Table 2.11.

WikiWars (Mazur and Dale, 2010) contains 22 Wikipedia documents that narrate historical events like wars. The authors aimed at demonstrating that temporal relations in historical texts are more complex, as longer texts lead to transformations in narrative style. A German version of the corpus has also been created: WikiWarsDE (Strötgen and Gertz, 2011), which contains the articles on the German Wikipedia related to the same historical events in the English WikiWars. Both corpora have been annotated following the TIMEX2 scheme, and only contain the temporal information of time expressions. WikiWars corpora have been semi-automatically annotated and human annotators have only corrected the temporal tagging. Wikiwars was annotated using DANTE (Mazur and Dale, 2007) and Callisto[22], while WikiWarsDE was annotated using HeidelTime (Strötgen and Gertz, 2013).

There are also other WikiWars versions for other languages. WikiWarsHr (Skukan *et al.*, 2014) is the Croatian version of WikiWars and it contains 22 articles on wars from Wikipedia, most corresponding with those in WikiWars. The corpus is formed of almost 60,000 non-punctuation tokens and 1,440 time expressions. The main difference is that this corpus has been annotated with the TimeML TIMEX3 tags for temporal expressions which represent richer information.

Another historical text corpus is ModeS TimeBank (Guerrero Nieto *et al.*, 2011). This corpus is formed by 102 historical documents on navigation and the spreading of information in the 18th century. The texts are written in Modern Spanish, which requires a previous pre-processing and normalisation stage before temporal annotation can be done. Temporal information has then been annotated manually following TimeML using the Brandeis Annotation Tool (BAT).

As part of the THYME project (Styler *et al.*, 2014) and using a markup language based on ISO-TimeML, a gold standard that gathers electronic medical histories has been created. This is called the THYME corpus and it

[22] `https://www.mitre.org/research/technology-transfer/open-source-software/callisto-0`

contains 1,254 notes on two oncology fields from Mayo clinic practice. The notes contain the interaction between doctors and patients and have some characteristic features: i) they are quite standardised among institutions, ii) each of them describe a single interaction between the patient and the doctor and iii) they tend to document events in a certain timeframe. They are rich in temporal references and contain many context dependent expressions, abbreviations and highly specialised terminology. The corpus has been annotated semi-automatically. A section of 600 documents has been used for the Clinical TempEval task (Bethard *et al.*, 2015) in 2015.

The i2b2 temporal relation corpus (Sun *et al.*, 2013) is another temporally annotated clinical domain corpus. It contains 310 de-identified discharge summaries and circa 178,000 tokens. It was manually annotated following an adaptation of TimeML with events, time expressions and their normalisations and temporal relations. They consulted the THYME annotation guidelines, but they followed a simpler approach: they did not annotate signals and aspectual and subordination links. Eight annotators took part in the annotation effort and the Multi-purpose Annotation Environment (MAE) toolkit and the Multi-document Adjudication Interface (Stubbs, 2011) were employed.

The corpus used in the SIBILA project (Wonsever *et al.*, 2012) combines news and historical texts in Spanish. It is formed by 11,986 tokens and 408 sentences, the ones form news texts being extracted from TempEval2 corpus (Verhagen *et al.*, 2007). It has been manually annotated and it has later been used for machine learning methods (Conditional Random Fields and Support Vector Machine (Boser *et al.*, 1992) training and evaluation.

Korean TimeBank (Jeong *et al.*, 2015) is a corpus formed of Wikipedia pages of the personage, music, university and history domains and hundreds of manually-generated question answer pairs formed by more than 3,700 sentences. It has been manually annotated and it follows TimeML guidelines, although it has been modified to accommodate the features of the Korean temporal expressions. It has been used to train machine-learning temporal structure extractors using O-LIFE (online version of LIFE (Jeong and Choi, 2015)).

The SoNaR-corpus (Reynaert *et al.*, 2010) is a 500 million word corpus for Dutch from which a 1 million word core corpus has been created. This corpus contains named entities co-reference relations, semantic roles and spatial and temporal expressions, as well as PoS tagging, lemmatisation and syntactic analysis. The Temporal information was annotated using STEx annotation scheme (Schuurman *et al.*, 2010).

The KPWr corpus (Broda *et al.*, 2012) is a 1,637 document corpus that contains written and spoken documents in Polish. They aimed at building a representative corpus of the Polish language and, hence, they gathered formal and informal texts, old and contemporary and general and technical documents. It has a multilayer annotation that covers morphosyntactic and semantic levels. In what concerns temporal information, it contains time expressions and events as well as their classification (Kocoń and Marcińczuk, 2017), manually annotated following the PLIMEX (Kocoń *et al.*, 2015) markup scheme and using the Inforex annotation tool (Marcińczuk *et al.*, 2017).

In the PHEME project[23] (Derczynski and Bontcheva, 2014) a 400 tweet corpus was created for the training of rumour identification tools. Tweets were in English and contained microblogs about shootings or accidents. The corpus contained ISO-TimeML annotation for events and temporal expressions and ISO Space (Pustejovsky *et al.*, 2011) for locations and spatial entities. The resulting dataset contained 605 events, 122 time expressions, as well as 139 spatial entities and 223 locations. This corpus was later combined with the TempEval2 data for temporal entity annotation.

Mostafazadeh *et al.* (2016) built a common-sense story corpus. The ROCStories corpus contained 40,000 five-sentence stories written by the Amazon Mechanical Turk workers. Events and the temporal and causal relations among them were annotated. Temporal information was manually annotated by experts following a simplified TimeML scheme.

RussianFlu-DE (Canh *et al.*, 2017) is a corpus formed by news in the ANNO repository[24] containing information about the flu epidemics in Austria and Germany between 1889 and 1893. The corpus is formed by 639 documents in German which preserve the original spelling but have also been normalised to modern German. Documents were converted to SGML format and the document creation date was set to the publication date of the news. Time expressions have been semi-automatically annotated using HeidelTime and have been assigned TIMEX2 tags. In total, the corpus contains over 453,000 tokens and 7,492 time expressions.

[23] `https://www.pheme.eu/`
[24] `http://anno.onb.ac.at/`

2.11 Table – No-news corpora carrying temporal information

Corpus	**Language**	**Composition**	**Mark-up language**	**Annotation tool**	**Manual vs. auto**
WikiWars (Mazur and Dale, 2010)	English	22 Wikipedia historical texts	TIMEX2	DANTE (Mazur and Dale, 2007) + Callisto	Semi-automatic
WikiWarsDE (Strötgen and Gertz, 2011)	German	22 Wikipedia historical texts	TIMEX2	HeidelTime (Strötgen and Gertz, 2013)	Semi-automatic
WikiWarsHR (Skukan *et al.*, 2014)	Croatian	22 Wikipedia historical texts (ca. 60,000 non punctuation tokens, 1,440 time expressions)	TimeML TIMEX3	HeidelTime (Strötgen and Gertz, 2013)	Semi-automatic
ModeS TimeBank (Guerrero Nieto *et al.*, 2011)	Modern Spanish	102 documents of 18th century	TimeML	Brandeis Annotation Tool (BAT)	Manual
THYME corpus (Styler *et al.*, 2014)	English	1,254 notes on oncology (600 documents used in Clinical TempEval 2015)	ISO-TimeML	Rule-based system	Semi-automatic
i2b2 corpus (Sun *et al.*, 2013)	English	310 discharge summaries (ca. 178,000 tokens)	TimeML	Multi-purpose Annotation Environment and Multi-document Adjudication Interface (Stubbs, 2011)	Manual

SIBILA (Wonsever *et al.*, 2012)	Spanish	11,986 tokens and 408 sentences in news (from TempEval2 corpus) and historical texts	SIBILA		Manual
Korean TimeBank (Jeong *et al.*, 2015)	Korean	Wikipedia pages and question-answer pairs (3,700 sentences)	TimeML		Manual
SoNar corpus (Reynaert *et al.*, 2010)	Dutch	1 million words (out of 500 million)	STEx for temporal expression (+ spatial and linguistic information)		
KPWr corpus (Broda *et al.*, 2012)	Polish	1,637 documents	PLIMEX (Kocoń *et al.*, 2015)	Inforex (Marcińczuk *et al.*, 2017)	Manual
PHEME tweet corpus (Derczynski and Bontcheva, 2014)	English	400 tweets (605 events, 122 time expressions, 139 spatial entities and 223 locations)	ISO-TimeML and ISO-Space		
ROCStories (Mostafazadeh *et al.*, 2016)	English	40,000 5-sentence stories	Simplified TimeML		Manual
RussianFlu-DE (Canh *et al.*, 2017)	German	639 documents 453,000 tokens and 7,492 time expressions	TIMEX2	HeidelTime	Semi-automatic

As the development of tools and the correction and precision continuously grow, corpora will enlarge faster and language processing will reach unexpected goals. For example, they might now be useful in automatic timeline creation.

2.6 Works Done on Timeline Creation

The large quantities of structured temporal information gathered in corpora and the temporal information extraction tools can be employed in tasks that build or produce new information based on existing temporal information. Some of the most relevant works on temporal information usage are described in this section.

First, it can be used for time-based information clustering. More precisely, Alonso *et al.* (2009) propose using time expressions in documents to build document temporal profiles. Those profiles, a sort of timelines that list the time expressions in the document, are then used to arrange documents along a chronological axis. The authors assume that adding temporal information to the information retrieval task will help users get more accurate hits for their queries.

Temporal information was also taken into account when building InZeit (Setty *et al.*, 2010), a system which assists users to obtain quick insights on the temporal milestones of the topic of the query by determining insightful time points. As a matter of fact, query hits are commonly arranged according to their estimated relevance or to their chronological date, but the temporal evolution of the topic cannot be observed easily. InZeit reassigns the relevance value of a document according to the insightfulness of the time boundaries of the document, which is obtained by measuring the insightfulness of the other ranked document times.

InZeit was proved on the New York Times corpus (Sandhaus, 2008), which contains circa 1.8 million daily articles published in the New York Times newspaper between 1987 and 2007. Through a Lucene-based[25] query interface, InZeit showed the hits for a query placed in a timeline and, hence, the evolution of the topic. Additionally, InZeit displayed the key real-world events relevant to the topic.

Some other efforts are focused on building timelines with the times and events present in text. That is the case in news summarisation or building timelines from historical texts.

[25] `https://lucene.apache.org/`

One of the first attempts on cross-document topic summarisation was performed by Allan *et al.* (2001). They aimed at creating methods that would help the monitoring of changes in event through time. They assumed that a news topic was formed by several events and that documents on a certain topic would deal with some of the events related to the topic. Additionally, the order of the events within the topic and the order of the sentences in document combined made up the total order of the sentences.

The task of their system was to assign a score to every sentence that indicated the importance of that sentence in the summary, considering a summary consisted of all sentences scoring over some threshold. They made some experimentation based on language model representation of topics and events and they also took into account the "usefulness" of a sentence (its relation to the topic) and the novelty of the event introduced by the sentence. For their experimentation the selected 22 medium-sized topics from the TDT-2 corpus. The results on the test corpus (11 topics) showed that the language models that assigned the highest score to the first sentences of the documents got the best results in summary effectiveness.

Chieu and Lee (2004) attempted to summarize a large document collection obtained from a query-based search, by placing sentences that reported "important" events related to the query along a timeline. They considered important events were those often repeated in many news articles over a period of time surrounding the date of occurrence of the event. In order to build the timeline they first identified the sentences relevant to the topic and resolved the dates of the events in those sentences. For this they considered the first temporal expression in the sentence was the anchor of the event and built a simple rule-based system to normalise those time expression. Then, sentences were ranked and duplicated sentences removed, clustering events according to their anchor dates.

System development was done on the Reuters Volume 1 corpus (Rose *et al.*, 2002) using the queries "earthquake" and "quake" and two major peaks of relevant events were shown related to the two biggest earthquakes in the period the news in the corpus comprehended. For system evaluation, they performed a small experiment to compare their timelines against manually constructed timelines. They used articles from the English Gigaword corpus[26] and used person names for the queries. The evaluation showed that human evaluators found automatically created timelines representative of

[26] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05`

media coverage, and sometimes judged them to be better than human built timelines.

Evolutionary Timeline Summarization (ETS) (Yan *et al.*, 2011) consisted of outputing a timeline with items of component summaries which represented evolutionary trajectories on specific dates from the returned collection of a user query. In order to build the timeline, they obtained a collection of sentences related to a query and associated to their publishing dates. The output of the system was an evolutionary timeline which consisted of a series of individual but correlated summary items for which the relevance, coverage and coherence and the diversity of novel ideas was taken into account. As a consequence, they got a *utility* measure for each summary item that would decide which information will be introduced in the timeline.

In order to evaluate the system, they gathered a dataset containing manually built timeline gold standards. They tried different systems on their corpus in order to assess the convenience of the ETS. They measured the different systems using ROUGE toolkit (Lin, 2004) for summarisation performance evaluation, and they discovered that both the ETS system that built timelines with a steady set of documents and the one that took an augmenting corpus outperformed all the systems tried, including the system developed by Chieu and Lee (2004) which was ranked third.

Tran *et al.* (2015) focused on date selection for timeline building. That is to say, which dates should be included in a timeline. Their task consisted in determining the dates of the subevents in the time span of a main event. They built a date reference graph including all the dates in texts and they modelled the connections between the dates according to their frequency, their temporal influence and their topical influence. Then, they performed random walks to rank the collection of dates.

Their dataset[27] contained 21 ground truth timelines on four events as well as a corpus of news articles covering each event. They implemented a system that built graphs of all the dates in the corpus of an event and then implemented five unsupervised strategies that combined frequency, temporal influence and topical influence. Results proved that the linear combination of the influence factors greatly improved the performance. Their work showed that the frequency in the corpus of a certain date, as well as its co-appearance with other relevant information to the topic and the fact that the date was mentioned over a long time were crucial for the task.

[27] http://l3s.de/~gtran/timeline/

2.7 Summary

In this section, we have analysed the most important works in the analysis and processing of temporal information, in order to determine the path in the treatment of the temporal information in Basque. First, we have identified the most important concepts that will be used in our work and we have decided to how to name them. Second, we have analysed the theoretical bases of temporal information processing. Then, we have analysed the resources created for temporal information processing in other languages, such as mark-up schemes, corpora and mark-up tools and automatic extraction tools for temporal information.

The study of previous works has helped us to decide how to structure our work. For example, we have determined which elements of temporal information will be given the biggest relevance, and we have decided that those elements will be normalised through the Basque version of the TimeML mark-up scheme. In fact, TimeML is a standard language that has been used in many languages, and hence, using TimeML for temporal annotation in Basque, will make the annotation comparable to other languages.

ANNOTATION IN BASQUE AND CORPUS CREATION

3

Events and Time Expressions and the relations among those

Temporal information can be summarised in what happens and when does it happen. What happens are events and when do those happen is expressed in text by time expressions. However, not all the events in text happen at the same time; they are placed in time ordered relative to one another or can be anchored to a time point. Thus, relations are created among events and time expressions. In this chapter events (Section 3.1), time expressions (Section 3.2) and the relations created among those (Section 3.3) will be described.

In this chapter we will offer many examples in order to illustrate Basque temporal constructions. In those examples relevant constructions will be highlighted in bold and other interesting information will be presented in italics.

3.1 Events

As mentioned in the previous chapter, we will consider events actions processes and states, as well as generic predications. In the following sections we will see how these are expressed in Basque and what kind of linguistic information they convey.

3.1.1 Event expressions in Basque

Events can be expressed by more than one grammatical category. As in many other languages, mainly verbs (4), nouns (5), adjectives (6) and adverbs (7) (in bold) can express events in Basque:

(4) *Hor* *ez* ***dira sartuko*** *Edesako* *langileak.*
There no **AUX enter.FUT** Edesa.REL workers.ABS
'Edesa workers **will** not **enter** in there'.

(5) *Fagor Etxetresnak* *enpresak* ***konkurtsora*** *joko* *du.*
Fagor Etxetresnak company.ERG **tender.ALL** go.FUT AUX
'Fagor Etxetresnak company will go out to a **tender**'.

(6) *Sartu* *den* *emakumeak* ***gaztea*** *dirudi.*
Come.in AUX.REL woman.ERG **young** looks
'The woman who has come in looks **young**'.

(7) ***Txaloka*** *egin* *dute* *ibilaldi* *guztia.*
Clapping do AUX walk all.DET
'They have done all the walk **clapping**'.

The events in bold in (4–7) express a single event: a single action or state. We have analysed these event expressions and they linguistic information they give based on Basque grammars(Altuna *et al.* (1985), Altuna *et al.* (1987) and Hualde and de Urbina (2003)), a classification of complex predicates Jędrzejko (2011) and the decisions taken for temporal annotation in other languages (TimeML Working Group (2010), Caselli *et al.* (2011)). These expressions are described more in depth and examples for each are given below.

3.1.1.1 Events Expressed by Verbs

Events in Basque are mainly expressed by verbs. We will now present the verb forms that may express an event.

Synthetic Forms Synthetic forms (8) are one-word units. The lexical root conveys the semantic information and a compound of morphemes add aspect, tense, person and mood information. Only a handful of verbs possess synthetic forms and their extension in the verbal paradigm is also restricted to some tenses. As a consequence, the majority of verbal events in Basque will be expressed by means of periphrastic forms.

(8) *Zientzialariek urteak* ***daramatzate*** *fusiozko energia*
Scientists.ERG years.ABS **have.been** fussion.INS power
merkearen bila.
cheap.GEN looking.for.
'Scientists **have** long **been** looking for cheap fusion power'.

Periphrastic Forms For periphrastic forms (also called analytical) we stick to the traditional Basque definition. These forms are formed by a lexical head (*atera*) which bears aspectual information and a mood, person and tempus carrying auxiliary (*dira*) (9). All verbs in Basque have periphrastic forms. The main formal variation happens in auxiliaries, which drastically reduces the mechanisms of morphological creation since lexical heads do not vary largely. This phenomenon makes the creation of periphrastic forms an easy language resource and leads to the reduction of synthetic form use.

(9) *8etan* ***atera dira*** *mendizaleak mendi tontorrerantza.*
8.PL.LOC **leave AUX** hikers.ABS mountain summit.DIR
'Hikers **have left** at 8 towards the summit'.

Non-finite Forms Verbal expressions may also appear in Basque texts as non-finite forms (radical, participles (10) and verbal nouns). These forms are used on their own as sentence heads in contexts such as fossilised expressions, exclamatory sentences and questions.

(10) *Akordiorik* ***lortu*** *ezean, grebari eutsiko diote.*
Agreement.PART **reach** NEG, strike.DAT continue AUX
'If no agreement **is reached**, (they) will continue on strike'.

3.1.1.2 Events Expressed by Nouns

Some nouns may express events in Basque. These can be verb nouns (11), common nouns (12) or proper nouns denoting a particular event (13).

(11) *Derrigorrezkoa da Greziari zorraren zati bat*
Compulsory is Greece.DAT debt.SG.GEN part.ABS a
barkatzea.
condone.ABS

'It is compulsory to **condone** a part of the Greek debt'.

(12) *Lau* ***eskaera*** *nagusi egin dituzte.*
Four **request** major.ABS do AUX
'(They) have done four major **requests**'.

(13) *2016an* ***Olinpiar Jokoak*** *Rio de Janeiron ospatuko dira.*
2016.LOC **Olympic Games**.ABS Rio de Janeiro.INE held.FUT AUX
'In 2016 **Olympic Games** will be held in Rio de Janeiro'.

We follow a test proposed by the TimeML Working Group (2010) to decide whether a noun refers to an event. A noun may express an event if it fits at least in two of the presented settings:

- NOUN lasted for several minutes/days/years/...
- NOUN was very fast/immediate/...
- NOUN took/takes/will take place *in temporal expression.*
- NOUN began/continued/ended *in temporal expression.*

3.1.1.3 Events Expressed by Adjectives

Adjectives express the qualities of the entity they refer to. Although they may appear in many contexts, we only consider events the adjectives acting as predicate adjectives (14).

(14) *Zer egin behar da enpresa* ***bideragarria*** *egiteko?*
What do have.to AUX business **profitable** make.FIN
'What has to be done to make the business **profitable**?'

3.1.1.4 Events Expressed by Adverbs

Adverbs will be considered event expressions when they accompany verbs (15) to create a more complex event construction (Section 3.1.1.6). These will mainly be adverbs of manner.

(15) *Arrakasta* ***harro*** *egoteko modukoa da*
Success.ABS **proud** be.FIN likely.ABS is
'Success is big enough to be **proud** of it'.

3.1.1.5 Events Expressed by Pronouns

Pronouns themselves do not express events, but may have a deictic value when they co-refer with another event in the text (16).

(16) *Bihar* *egingo* *da* *mozorro* *desfilea.* ***Horretarako***
Tomorrow do.FUT AUX.3.SG costume parade.ABS .PUN
erdigunea *itxiko* *dute.*
That.FIN centre.ABS close.FUT AUX
'Costume parade will be done tomorrow. For **that** the centre will be closed'.

3.1.1.6 Complex Structures

Complex predicates conform a non-homogeneous yet gradated group (Jędrzejko, 2011). We will now present Basque complex predicate structures, based on Jedzrejko's (2011) classification for Polish complex predicates. We have adapted this list to accommodate Basque predicate features as can be seen in the following lines:

- **Standard nominal predicates** are constructions with a basic auxiliary verb (*izan* (to be) (17) or *ukan* (to have)). The verbs in these constructions are semantically or referentially empty and the nominal or adjectival predicate carries all the predicative information.

  (17) *Ordenagailua* ***geldoa da.***
  Computer.ABS **slow.ABS is**
  'The computer **is slow**'.

- **Modal predicates** are complex constructions formed by a lexical verb in its participle form and a conjugated modal verb (*nahi* (to want), *behar* (to have to, must) (18) and *ahal* (to can)) or derived nouns expressing modality (*nahi* (wish), *behar* (need, obligation), *ahal* (possibility)) and a participle.

  (18) *Jende* *nagusiak* *noizbehinka* ***jesarri beharra du.***
  People elderly.ERG sometimes **sit.down need.ABS has**
  'Elderly people **has the need to sit down** sometimes'.

- **Aspectual predicates** are formed by an aspectual verb (*amaitu*) or noun and a verbal or nominal predicate (*beherapenak*) (19). The aspectual expression in the construction marks the phase of its argument.

(19) ***Beherapenak amaitu dira***
Sales.ABS end AUX
'**Sales have ended**'.

- **Generic verb constructions**. Generic verbs are those which are used to give predicative properties to nouns. Therefore, generic verb constructions are formed by a noun that carries the lexical meaning of the event (*musu*) and a verb that provides the syntactic information (*eman*) (20).

(20) *Ona da lagunek elkarri* ***musu ematea***.
Good is friends.ERG each.other.DAT **kiss give.ABS**
' It is good for friend to **kiss** each other'.

- **Metaphors**. In these constructions a fully predicating verb is used next to a noun phrase with a meaning other than its main meaning (21).

(21) *Entzuleak* ***barrez lehertu ziren***.
Hearers.ABS **laugh.INS explode AUX**
'Hearers **laughed** a lot'.

- **Idiomatic expressions** are formed by a noun (*hautsak*) and a verb (*harrotu*) which carries the grammatical information of the construction (22). Nevertheless, idiomatic constructions cannot be seen as a simple sum of the meanings of its parts and can only be understood as a single meaning unit.

(22) *Adierazpenek* ***hautsak harrotu zituzten***.
Statements.ERG **dusts.ABS raise AUX**
'The statements **caused a commotion**'.

Although these complex constructions express a single complex event, they may contain more than one event expressing form.

3.1.2 Event features

We have analysed the event features that offer the most relevant information to temporal information processing.

3.1.2.1 Event classification

Events can be classified according to their semantics. We followed the event classification in TimeML for the Basque events; a classification that has an interesting feature: it divides the events according to their semantic features, as well as whether they can take an event as an argument.

- **Occurrence**: these are dynamic events that happen or occur, e.g. *salto egin* "to jump", *dantzatu* "to dance" or *ibili* "to walk".
- **State**: these are events describing circumstances in which something obtains or holds true and do not vary over time, e.g. *egon* "to be" or *geratu* "to remain".
- **Reporting**: reporting events describe the utterance, narration, description, etc. of an event, e.g. *esan* "to say", *azaldu* "to explain" or *iragarri* "to announce".
- **Aspectual**: aspectual events indicate the beginning, continuity or end of an event: e.g. *hasi* "to begin", *jarraitu* "to continue" or *amaitu* "to end".
- **Perception**: these events describe the physical perception of another event, e.g. *ikusi* "to see", *entzun* "to hear" or *sumatu* "to perceive".
- **Intensional action**: these are dynamic events that select for an event-denoting argument which is explicit in the text, e.g. *saiatu* "to try", *agindu* "to order" or *aztertu* "to analyse".
- **Intensional state**: these are states that, as intensional actions do, select for an event-denoting argument which is explicitly in the text, e.g. *pentsatu* "to think", *gorrotatu* "to hate" or *prest egon* "to be ready".

This classification is of an utmost importance, since the semantics of each event will condition the relations that are created between events. For example, reporting events will introduce a subordinated event for which they will give evidence and aspectual events will determine the status (e.g. start, continuation, end) of the event they take as an argument.

3.1.2.2 Tense

Verbal tense indicates that the event that occurred in the past (23), in the present (24), or in a virtual time (25) (Altuna *et al.*, 1985). This information helps to place verbal events in the timeline. In fact, tense contributes place events with respect to the time of utterance, to the document creation time or to another event.

(23)

Orain	*arteko*	*terminal*	*guztiak*	*prest*	***daude***	*A380a*	
Now	until.REL	terminal	all	ready	**are**	A380.ABS	
hartzeko.							
take.							

'All the present terminals **are** ready to take the A380'.

(24)

Gehitu	***zuen***	*lurreratzea*	*gertaera*	*gogoangarria*	*zela.*
Add	**AUX**	landing	event	memorable	was.

'(S)he **added** the landing was a memorable event'.

(25)

Lanak	*garaiz*	***amaituko***	***balira****,*	*3.*	*terminalak*	*datorren*
Works	on.time	**end**	**were**,	3^{rd}	terminal.ERG	coming
urteko	*urtarrilean*	***zabalduko***	***lituzke***	*ateak.*		
year.REL	January.INE	**open**	**would**	doors.ABS		

'If works **ended** on time, the 3^{rd} terminal **would open** January next year'.

3.1.2.3 Aspect

Verbal aspect provides information about the perfectness of the event. Aspect expresses whether i) the event is finished (26), ii) it is an ongoing or habitual event (27) or iii) it is a future event (28).

(26)

Gehitu	***zuen***	*lurreratzea*	*gertaera*	*gogoangarria*	*zela.*
Add	**AUX**	landing	event	memorable	was.

'(S)he **added** the landing was a memorable event'.

(27) *Orain arteko terminal guztiak prest* ***daude*** *A380a*
Now until.REL terminal all ready **are** A380.ABS
hartzeko.
take.

'All the present terminals **are** ready to take the A380'.

(28) *3. terminalak datorren urteko urtarrilean* ***zabalduko***
3rd terminal.ERG coming year.REL January.INE **open**
ditu *ateak.*
would doors.ABS

'the 3rd terminal **will open** January next year'.

3.1.2.4 Modality

We also paid special attention to modality expression in Basque. We determined that there are three main modality values that are expressed by verbs and nouns. The analyzed expressions and their meanings are shown in Table 3.1.

3.1 Table – Modality expressions and their meaning.

Verb form	Noun	Meaning
Ahal izan (ezin izan)	ahal (ezin)	can, may, might, be able to (cannot, may not, might not, not be able)
Behar izan	behar	need, ought to, must, have to
Nahi izan	nahi	want, wish

3.1.2.5 Polarity

Event polarity expresses whether the event is affirmed (29) or negated (30). Negation is a linguistic phenomenon that inverts the truth value of the proposition it is applied to (Martí *et al.*, 2016). Negation is usually expressed by lexical and syntactic elements that are called negation cues (*ez* "no " (30)). In general, every negation cue is associated with its scope and focus. More precisely, scope is the extent of the text affected by the cue (Morante *et al.*, 2011) and the focus is the most directly negated part of the sentence. As a consequence, one can say that if the event is affected by the negation, it will be negated, whereas if there is no negation cue in the sentence, the events in it will be of positive polarity.

(29) *Aljeriarrek egozketa adarra eta etxe-tresna txikiena **erosi***
Algerians.ERG expulsion branch and appliance smallest **buy**
***nahi dute**.*
want AUX.
'The algerians **want to buy** the expulsion and small appliances branches'.

(30) *Garraio-ministroak hegazkin europarrak ez **erosteko** eskatu*
Transportation.minister plane European *no* **buy** ask
die aire-lineei.
AUX airlines.DAT.
'The transportation minister has asked the airlines *not* to **buy** European planes'.

3.1.3 Additional event information

As the research has progressed, we have felt the need to treat additional information that is useful in temporal information processing. Initially, when analysing temporal information, we only examined the relation between events and chronology. However, during that analysis, we have seen that the relation between textual events and events in real world should also be studied. For that we have analysed the forthcoming event features.

3.1.3.1 Certainty

Certainty expresses the commitment of the source with the information expressed. We have considered that, unless there is an explicit uncertainty marker or it is impossible to give a certainty value, we will consider the events certain (31) (Altuna *et al.*, 2018b). The uncertainty particle *ote* in example (32) marks the uncommitment of the utterer for the certainty of *galdu* ("to loose").

(31) *Boeing-ek 11.000 milioi dolar **lortu ditu***
Boeing.ERG 11,000 million dollar **obtained has**
akordioetan.
agreement.PL.INE.

3.2 Table – Proposed factuality annotation scheme for Basque

Factuality related attributes		**Factuality value**
Certainty	**Special cases**	
Certain Uncertain Underspecified	Cond. Condition Cond. Main clause Generic Statement None	Factual Counterfactual Non factual No factuality value Underspecified

'Boeing **has obtained** 11,000 million dollars in agreements.'

(32) *Hegazkin-merkatuaren kontrola galdu* ***ote*** *zuen*
Airplane-market.GEN control.ABS loose **UNCERT.PART** AUX
eztabaida piztu zen.
discussion light AUX.

'Discussion on **whether** (it) had lost control over the aeroplane market was started.'

3.1.3.2 Factuality

Event factuality is described in Saurí (2008) as the level of information expressing the factual nature of eventualities mentioned in text, that is, whether events correspond to a fact in the world, a possibility or a situation that does not hold.

After analysing the state of the art proposals for factuality annotation, we opted for a simple scheme in order to ease the burden of manual annotation (Altuna *et al.*, 2018b). However, we did not want to sacrifice much information. Thus, we defined the scheme shown in Table 3.2.

As can be seen in the table, we represent factuality through five factuality values: `FACTUAL` for events that have happened, `COUNTERFACTUAL` for events that have not happened in the past, `NON_ FACTUAL` for future events, `UNDERSPECIFIED` for those events of which the factuality value cannot be assessed and `NO_FACTUALITY_ VALUE` for the events that do not express any specific event. These values are conditioned by the values of the factuality related attributes.

Certainty, as well as *polarity* and temporality (verb *tense* and *aspect*), are widely considered factuality features as they convey the majority of factual information, while identifying *special cases* adds relevant information to factuality resolution. That is the reason for adopting it from Minard *et al.* (2014).

3.1.3.3 Special cases

In what concerns the *special cases*, we wanted to emphasize the effects of conditionals and generic statements (Altuna *et al.*, 2018b). For example, when using the hypothetical tense like in "If only I had come. . . ", although the verb has a positive polarity, humans know that the utterer has not come. In (33) *ematen badu* ("if it gives") in the protasis expresses the preliminary condition while *bilatuko du* ("will look for") in the apodosis is expresses the consequence. The specific mark for generic statements express that those events do not refer to a specific event in a specific time and place. Such is the case of *da* ("is") in (34).

(33) *Bilaketak fruiturik ematen ez badu, BEAk*
Search.ERG results.PART *bring* no *AUX*, ANR.ERG
jarraitzeko dirua ***bilatuko du.***
continue.FIN money.ABS **look.for AUX**.
'If the search *brings* no results, the ANR **will look for** money to continue.'

(34) *Airbus A320a korridore bakarreko hegazkina* ***da.***
Airbus A320.ABS aisle single.REL aeroplane **is**.
'The Airbus A320 **is** a single aisle aeroplane.'

3.2 Time Expressions

Time expressions are textual constructions that refer to points and intervals in time. We will now see how they are expressed in Basque and which is the most relevant information they convey.

3.2.1 Time expressions in Basque

Time expressions in Basque can be expressed by noun phrases (35), adjective phrases (36), adverb phrases (37), noun phrases that contain a free postposition (38) and a number of expressions such as dates and hours (39).

(35) ***Igande goizean*** *ibiltzera joan ginen.*
Sunday morning.INE walk.ADL go AUX
'We went for a walk on **Sunday morning**.'

(36) *The Black Dwarf Londreseko* ***astekako*** *argitalpena izan zen.*
The Black Dwarf London.REL **weekly.REL** publication be AUX
'The Black Dwarf was a London **weekly** publication.'

(37) *Garaipena* ***berandu*** *iritsi zen.*
Victory **late** arrive AUX
'Victory arrived **late**.'

(38) *Moskurako trena* ***bostak aldera*** *abiatuko da.*
Moscow.ADL.REL train **five.DET side.ADL** leave.FUT AUX
'The train to Moscow will leave by five.'

(39) *Bidaia* ***2016-06-20an*** *hasi zen.*
Trip.DET **2016-06-20.INE** start AUX
'The trip started on **06/20/2016**.'

3.2.1.1 Lexical triggers

Lexical triggers are crucial in time expression recognition as they explicitly express time and give evidence on the presence of a temporal expression. In the examples (40–41), *gaur* (today) and *ordu* (hour), as well as the numeric expression *23:55*, explicitly express time.

(40) ***Gaur*** *uda hasi da.*
Today Summer.DET started has.
'Summer has started **today**'.

(41) *Filmak bi* ***ordu*** *iraun du.*
Film.ERG two **hour** lasted has.
'The film lasted two **hours**'.

(42) *Azken trena:* ***23:55***
Last train.DET: **23:55**
'Last train: **23:55**'.

3.2.2 Linguistic features

We have analysed the features of time expressions that offer the most relevant information to temporal information processing.

3.2.2.1 Classification

According to the type of time they express four different temporal expressions (TimeML Working Group, 2010) can be distinguished:

- **Dates**: expressions referring to a particular period based on the Gregorian calendar, e.g. *martxoaren 8a* "8th of March".
- **Times**: expressions that refer to a particular subdivision of the day, e.g. *bostak* "five o'clock".
- **Durations**: these expressions refer to an extended period of time e.g. *hiru aste* "three weeks".
- **Sets**: expressions that consist in the quantification of a temporal unit e.g. *egunean 8 ordu* "8 hours a day" or express the regularity or re-occurrence of an event e.g. *egunero* "every day".

3.2.2.2 Value

The time points and durations expressed by time expressions have to be assigned a normalised value in temporal information processing. That is to say, it is to be stated which point in chronology or which duration they refer to.

In some cases, it is easy to identify which point in the chronology the time expression refers to (43), for it is explicit in text. Some others, instead are relative time expressions (44) or vague (45). The first can be normalised if a reference time point is defined. The vague time expressions, instead, cannot be assigned a defined normalised value, but an approximated normalised value.

(43) *iTunes Music Store onlineko musika-denda merkaturatu zuen*
iTunes Music Store online music.shop launched AUX
2003ko apirilaren 28an.
2003.REL April.GEN 28.INE.

'(It) launched the online iTunes Music Store music shop the 28th of April 2003.'.

(44) *3. terminalak* ***datorren urteko urtarrilean*** *zabalduko*
3rd terminal.ERG **coming year.REL January.INE** open.FUT
ditu ateak.
AUX doors.

'The 3rd terminal will open on **January next year**'.

(45) *Hegazkinari 400 bat gonbidatuk egin zioten ongietorria* ***laster***
Plane.DAT 400 one guests.ERG did AUX welcome **soon**
zabalduko den 3. terminalean.
open.FUT AUX 3rd terminal.INE

'About 400 guests welcomed the plane in the 3rd terminal, which will **soon** be opened'.

3.2.2.3 Function in document

Not all the time expression play the same role in text. Most of them can only be understood as being part of the text (46). Apart from those expressions that are explicit in texts, we have also considered the document creation time (DCT). This is normally a time point that has a day, minute or second granularity and expresses the moment the document has been created. It is commonly presented as a numerical expression (*2017-04-11*, "11/04/2017") or a full date (*2017ko apirilaren 11*, "11th of April 2017"), although it is not always explicit in the text. The creation time is used for the resolution of relative time expressions in the text such as the one in (46).

(46) *3. terminalak* ***datorren urteko urtarrilean*** *zabalduko*
3rd terminal.ERG **coming year.REL**
ditu ateak.
January.INE open.FUT AUX doors.

'The 3rd terminal will open on **January next year**'.

3.3 Relations

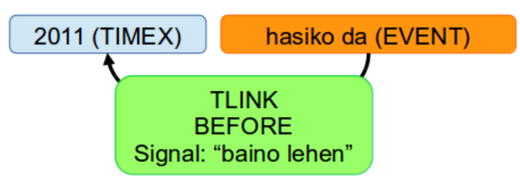
For our purposes, relations are links created between events and/or time expressions. We have defined three types of relations:

- **Temporal relations**: they represent the temporal relation holding between two events, two time expressions or between an event and a time expression.
- **Subordination relations**: relate an event and its subordinated event.
- **Aspectual relations**: represent the relationship between an aspectual event and its argument event.

3.4 Temporal relations

Finally, temporal relations take the `TLINK` tag and the `relType` and `Signal` attributes (Figure 3.1 shows the relation in (47)). The first expresses the temporal relation between the main event or temporal expression and the subordinated event or time expression. The second is employed to express which element in the sentence (signal) makes the temporal relation explicit. All three links and their corresponding attributes have been included in the Basque TimeML guidelines with no changes with respect to English.

(47) *Superjumboa* *2011* *baino lehen* ***hasiko*** ***da*** *zerbitzuan.*
Superjumbo.ABS *2011* than before **start.FUT** **AUX** service.INE.
'The Superjumbo will start working before 2011.'

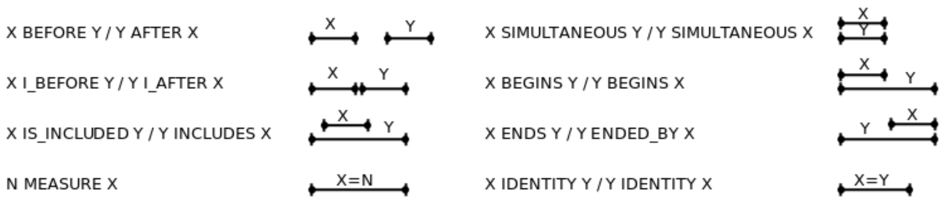


3.1 Figure – Representation of a temporal link in (47).

3.4.1 Classification

We have classified those temporal relations following the classification in TimeML which was based on Allen's Interval Theory (1983).

As can be seen from Figure 3.2, in TimeML's classification there are eight temporal relations that represent the possible relations between two



3.2 Figure – Representation of the temporal relations in TimeML.

time intervals (or events). Except the measure relation (n measures x) , all the relations have a complimentary relation that allows a two-direction equivalent relation.

3.4.2 Signals

Signals indicate how temporal objects are to be related to each other. In our case, we only analysed signals that express temporal relations, that is to say, those that help ordering and placing events and time expressions in the timeline (Altuna *et al.*, 2014a). In Basque, these are mainly expressed by free postpositions (48), temporal relation suffixes (49) temporal relation words (50), time particles (51) and special characters such as "-" or "/" (52).

(48) *Bigarren singlea entzungai izango da urriaren* ***6tik***
Second single audible be.FUT AUX October.GEN 6.**ALA**
aurrera*.*
onwards

'The second single will be available for listening from the 6th of October.'

(49) *Partida amaitu* ***zenean*** *dimisioa eman zuen*
Match end **AUX.TEMP** resignation give AUX
entrenatzaileak.
coach .

'The coach resigned when the match ended.'

(50) *Bira amaitu* ***orduko*** *kantu berriak lantzen hasi ginen.*
Tour finish **hour.REL** song news working start AUX

'As soon as we finished the tour, we started working on new songs.'

(51) *Bira amaitu* ***berri*** *da.*
Tour finish **just** AUX

'The tour has just finished.'

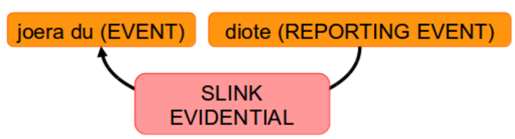
(52) 2017-2020.

In what concerns temporal relation suffixes (49), we decided to only treat those attached to the auxiliary verbs. The reason for this is that we did not want to overlap the event tag and the signal tag, as we annotate on token level. In other contexts, we only take into account postpositions that express a temporal relation when they are free postpositions. As a consequence, we noticed that the number of signals in our corpus is remarkably reduced.

3.5 Subordination relations

Subordination relations introduce the relation between two events in which one is the head and the other is its subordinated (Figure 3.3 shows the subordination relation in (53)). The kind of subordinated relation is conditioned by the class of the main event.

(53) *Historiak errepikatzeko joera duela* ***diote*** *zenbaitek.*
History.ERG repeat.FIN tendency *has* **say** some.ERG .
'Some say that history tends to repeat itself.'



3.3 Figure – Representation of the subordination link in (53).

For EusTimeML we have followed the classification in TimeML and we have defined the following subordination relation types:

- **Modal**: the head event of the relation creates the opportunity for a posible universe or reality.

  (54) *Endesak* ***uste du*** *Garoña berriro zabaltzeko* gai *izango dela.*
  Endesa.ERG **think AUX** Garoña again open *able* *be.FUT AUX.*
  'Endesa thinks it will be ableto open Garoña again'.

- **Factive** and **counter-factive**: the head event of the relation entails the veracity or non-veracity of the subordinated event. That is to say, in factive relations the head entails that the subordinated event has happened and in counter-factives, that it has not happened.

(55) *Kapitainak bera ere inoiz marinel* izan *zela* ***ahaztu***
Capitan.ERG him too ever sailor *be AUX* **forget**
zuen.
AUX.
'The captain forgot he too was once a sailor'.

- **Evidential** (56) and **negative evidential** (57): the head event conveys the evidence or bsence of evidence of the subordinated event.

(56) *Kapitainak bera ere inoiz marinel* izan *zela* ***esan zuen.***
Capitan.ERG him too ever sailor *be AUX* **say AUX**
'The captain said he too was once a sailor'.

(57) *Kapitainak bera ere inoiz marinel* izana ***ukatu zuen.***
Capitan.ERG him too ever sailor *be*.DET **deny AUX**.
'The captain denied he too was once a sailor'.

- **Conditional**: relation between the protasis and the apodosis of a conditional sentence.

(58) *Zuei horrelakorik* ***gertatzen bazaizue****, ez zenukete*
You.DAT similar.PAR **happen AUX**, no AUX
atarramentu onik aterako.
profit good.PAR *get.*
'If the same happened to you, you would not get any profit'.

3.6 Aspectual relations

Aspectual relations express the phase of the subordinated event (the aspectual link in (59) is displayed in Figure 3.4), since aspectual events express the status of the event they take as an argument. In the following examples the source of the relation is in bold and the target is represented in italics.

(59) *Apple multimedia produktu batzuk saltzen **hasi zen**.*
Apple multimedia product some *sell* **start** AUX .
'Apple started to sell some multimedia products.'

3.4 Figure – Representation of an aspectual link in (59).

Aspectual relations can be classified as follows according to EusTimeML:

- Initiation: the aspectual event expresses the start of the subordinated event.

  (60) *1956. urtearen bueltan,* irakasle ***hasi zen**.*
  1956 year.GEN round.INE, *teacher* **begin AUX**
  'Around the year 1956, she **started** to *teach*'.

- Culmination: the aspectual event expresses the natural ending of the subordinated event.

  (61) *Mikelek liburua* irakurtzen ***amaitu du**.*
  Mikel.ERG book.ABS *read* **end AUX**.
  'Mikel **has finished** *reading* the book'.

- Termination: the aspectual event expresses the interruption or termination of the subordinated event.

  (62) *Gobernuak azalpenak* emateari ***utzi dio**.*
  Government.ERG explanations.ABS *give* **stop AUX**.
  'The government **stopped** *giving* explanations'.

- Continuation: the aspectual event expresses the continuation of the subordinated event.

  (63) *Autobide berriaren* lanek *aurrera* ***diraute**.*
  Motorway new.GEN *works.ERG* forward **continue**
  'The new motorway *works* **continue**'.

- Reinitiation: the aspectual event expresses the reinitiation of the subordinated event.

(64) *Isunak* jartzeari ***berrekingo dio*** *udalak*
Fines.ABS *put.DAT* **restart.FUT AUX** council.ERG
datorren hilabetetik aurrera.
coming month.ABL forward.
'The council **will restart** *fining* from the next month'.

3.7 Summary

In this chapter we have seen which event and time expressions express temporal information in Basque and we have also described their most relevant features. We have also seen what kind of relations are created among events and time expressions. After this analysis, we will present how the information is normalised in the next chapter, for information has to be coded and made machine-readable in order to be useful for linguistic processing.

4 Manual Annotation

4.1 EusTimeML

In order to provide Basque with the necessary resources for the processing of temporal information, we have first developed EusTimeML (Altuna *et al.*, 2016b), a mark-up scheme for Basque temporal information based on TimeML. We aimed at accommodating Basque temporal information in a TimeML scheme, for which we first analysed how temporal information is expressed in Basque. We described temporal constructions, that is to say, events, time expressions and signals, as well as the relations between events and time expressions, following the classification proposed in TimeML.

Basque is a highly agglutinative language isolate that needs special attention when adapting resources created for other languages due to the linguistic distance between Basque and languages such as English or Spanish, which are normally used as a reference. One of the main features of Basque is the fact that the information expressed by prepositions in the neighbouring languages is expressed by postpositions in Basque. These postpositions appear attached to the roots by morphotactics. Even if the transmission of the categories and attributes from other languages to Basque has been relatively straight forward, some innovative decisions have been taken to address specific Basque constructions.

In this section we present the annotation decisions and the differences that exist between the TimeML annotation guidelines for English and the EusTimeML annotation guidelines for Basque. We provide examples to illustrate the aforementioned.

4.1.1 TimeML

TimeML is a surface-based annotation language (Im *et al.*, 2009) and, as a consequence, instead of encoding the interpretation of the annotated (event) constructions, only their grammatical features are marked up. Nonetheless, event categorisation is based on their semantic features.

TimeML was first discussed at a NRRC (National Regional Research Center) summer workshop as a resource to "address the problem of how to answer temporally-based questions about the events and entities in text" (Pustejovsky *et al.*, 2002, p. 1). TimeML differed from the previous mark-up languages as it i) anchored events to temporal expressions, ii) ordered the event expressions relative to one another and iii) allowed delayed interpretation to underspecified temporal expressions (Pustejovsky *et al.*, 2003a) as mentioned in Section 2.3.1.

TimeML extended the features of its forerunners: it widened the attribute range of the TIMEX2 tag, creating the TIMEX3 tag, and the event classification and introduced new temporal functions (which allowed the processing of relative time expressions like *last week* or *two days ago*) and created relations between events and times. Through different versions, TimeML has improved and the latest version (TimeML Working Group, 2010) is a reliable mark-up language for temporal annotation. Moreover, TimeML has become an ISO International Standard (ISO-TimeML working group, 2008), which confirms its high quality and wide use.

TimeML has been adapted to many languages: French (Bittar, 2010), Italian (Caselli *et al.*, 2011), Portuguese (Costa and Branco, 2012), Romanian (Forăscu and Tufiş, 2012) and Korean (Jeong *et al.*, 2015) for example. These versions keep the structure of the TimeML scheme, but present little variations in order to accommodate some linguistically dependent features.

Some other changes and versions like TimeML-strict (Derczynski *et al.*, 2013) have also been proposed. TimeML-strict is a "valid, unambiguous and easy-to-process" subset of TimeML (Derczynski *et al.*, 2013, p. 1). A more restricted scheme is presumably easier to process, making temporal information processing less time consuming. Main changes are i) the clear difference between the metadata and the text body of the document, ii) the impossibility of creating a relation with a phantom ID and iii) a reduced set of temporal relations.

TimeML has also been adapted to specific contexts. In the last decade, the automatic interpretation of medical histories has become a major subject of interest. This is a restricted area and the language used is very specific: i) sentences are not always of natural language, ii) many temporal relations are implicit, iii) the name of the medicines implies the treatment, etc. These features make health an interesting language processing field, as the medical texts differ from the texts analysed before. THYME-TimeML (Styler *et al.*, 2014) has been adapted to the medical domain in order to annotate temporal information in clinical narrative and offer clinical timelines. Some attributes of ISO-TimeML `<EVENT>` tag have been adjusted to the needs of the domain. Temporal expressions that express frequencies are more common than in general domain and time references like *preoperative* and *postoperative* are also frequent. Thus, `<TIMEX3>` tag attributes have also been adapted. Temporal relations have also been reduced and the events are ordered according to the time they were observed.

4.1.2 EusTimeML creation methodology

Following the trend of adapting TimeML to different languages, we have created EusTimeML, the temporal information mark-up language for Basque. For that, we have analysed the way temporal information is expressed in Basque (see Chapter 3) and we have defined the tags to make that information explicit and their attributes.

Manual annotation and guideline creation have been conducted simultaneously. Corpora give evidence on linguistic phenomena and how to annotate them has to be decided. Then, it has to be assessed whether annotation decisions represent temporal information properly. We have defined EusTimeML mark-up language in the EusTimeML guidelines. In the first version (Altuna *et al.*, 2014b), we addressed events, time expressions and signals. In the second version (Altuna *et al.*, 2016a), we presented the improved version of the first annotation decisions and added the mark-up guidelines for relations. The steps for guideline definition and corpus annotation have been the following:

1. We have built the first annotation guidelines based on Basque grammar analysis.

2. We have decided how to assign tags and attributes and the contexts they have to be applied in.

3. Annotators have partially annotated the corpus following the guidelines and doubts and disagreement issues have been discused.

4. After the annotation evaluation, we have created the final version of the guidelines.

5. We have manually annotated the gold standard corpus following the final version of the guidelines.

When creating EusTimeML we have taken some annotation decisions. We list them here:

- In the creation of EusTimeML, it is first remarkable that we adopted the single token policy for event annotation and only the lexical head of the event was assigned the tag. However, all morpho-syntactic information contained in the phrase (auxiliaries, demonstratives, etc.)

- In the first version of the guidelines (Altuna *et al.*, 2014b), two attributes for each aspect and tense were created for Basque verbs. This was due to the fact that these attributes have been historically considered bidimensional in Basque grammars Altuna *et al.* (1987). `aspect1` expresses whether the verb is perfective and `aspect2` whether the verb is in its future form, since the future is expressed in Basque as a verb aspect. For tense information, `tense1` expresses if the verb is in its present form and `tense2` if the verb is in a past form. Finally, the negation of both present and past tenses generates a hypothetical tense. But then we decided to merge those attributes in just tense and aspect.

- We have also defined three values for the modality attribute that concord with the three main modal verbs in Basque (see Table 3.1): `behar`, `nahi` and `ahal`. We considered the modality information was easy to express because of the little variation of the modal verbs and nouns. More details and the values that each attribute can take can be consulted in the Basque TimeML annotation guidelines[1].

- We have added factuality information annotation guidelines to the EusTimeML guidelines.

[1] `https://addi.ehu.es/handle/10810/17305`

- As Basque is an agglutinative language the postpositions that express a temporal relation are attached to the time expression itself. As a consequence, in the cases a single token expresses a time expression and a signal, only the time expression has been annotated.

4.2 EusTimeML mark-up language description

EusTimeML is an adaptation for Basque of TimeML and we have followed the precursor quite thoroughly. Most of the attributes and their values have been proved optimal to normalise Basque linguistic features, but some others have been created *ex profeso* in order to accommodate those. Some of these attributes correspond to the features we have presented in Chapter 3. Nonetheless, it should be noticed that those attributes are the most relevant, but there are more attributes that offer many other relevant information.

In the following subsections, we present the attributes and values for each tag in the Backus-Naur Form. It is to be taken into account that some of the attributes express linguistic information, while others are automatically created and refer to identification information or to relations to other tokens or markables. Additionally it needs to be pointed out that not all the attributes are mandatory. We present the optional ones between brackets.

4.2.1 Events

Events in EusTimeML get the `EVENT` tag. In Chapter 3, we have gone through the most describing event features which now get the following attributes:

- Part-of-speech – `pos`
- Class – `class`
- Tense – `tense`
- Aspect – `aspect`
- Modality – `modality`
- Polarity – `polarity`
- Certainty – `certainty`

- Factuality – `factuality`
- Special cases – `specialCases`

The full list of attributes and values events get is presented below.

```
attributes ::= eid eiid class tense aspect pos polarity
modality
eid ::= e<integer>
eiid ::= ei<integer>
pos ::=  'VERB' | 'NOUN' | 'ADJECTIVE' | 'ADVERB' | 'PRONOUN' |
 'OTHER'
class ::= 'REPORTING' | 'PERCEPTION' | 'ASPECTUAL' | 'I_ACTION'
 | 'I_STATE' | 'STATE' | 'OCCURRENCE'
tense ::= 'PRESENT' | 'PAST' | 'ALEGIAZKOA' | 'NONE'
aspect ::= 'PERFECT' | '-PERFECT' | 'FUTURE' | 'NONE'
modality ::= 'AHAL' | 'NAHI' | 'BEHAR' | 'NONE'
polarity ::= 'NEG' | 'POS' {default, if absent, is 'POS'}
certainty ::= 'CERTAIN' | 'UNCERTAIN' | 'UNDERSPECIFIED'
factuality ::= 'FACTUAL' | 'COUNTERFACTUAL' | 'NON_FACTUAL' |
   'NONE' | 'UNDERSPECIFIED'
specialCases ::= 'CONDITIONAL_CONDITION' | 'CONDITIONAL_MAIN' |
 'GENERIC_STATEMENT' | 'NONE'
```

4.2.2 Time Expressions

Time expressions are expressed by the `TIMEX3` tag, which can get the following attributes. As we can remember, in the previous chapter which have highlighted the most informing time expression features. Those are expressed by these attributes as follows:

- Part-of-speech – `pos`
- Type – `type`
- Value – `value`
- Function in document – `functionInDocument`

These features and some more are expressed in EusTimeML as presented now:

```
attributes ::= tid type [functionInDocument][beginPoint]
[endPoint][quant][freq][temporalFunction]
(value | valueFromFunction)[mod][anchorTimeID]
tid ::= ID
tid ::= TimeID
TimeID ::= t<integer>
type ::= 'DATE' | 'TIME' | 'DURATION' | 'SET'
beginPoint ::= IDREF
beginPoint ::= TimeID
endPoint ::= IDREF
endPoint ::= TimeID
quant ::= CDATA
freq ::= CDATA
functionInDocument ::= 'CREATION_TIME' | 'EXPIRATION_TIME' |
'MODIFICATION_TIME' | 'PUBLICATION_TIME' |
'RELEASE_TIME' | 'RECEPTION_TIME' |
'NONE' (default, if absent, is 'NONE')
temporalFunction ::= 'true' | 'false'
(default, if absent, is 'false')
{temporalFunction ::= boolean}
value ::= CDATA
valueFromFunction ::= IDREF
{valueFromFunction ::= TemporalFunctionID
TemporalFunctionID ::= tf<integer>}
mod ::= 'BEFORE' | 'AFTER' | 'ON_OR_BEFORE' | 'ON_OR_AFTER' |
'LESS_THAN' | 'MORE_THAN' | 'EQUAL_OR_LESS' | 'EQUAL_OR_MORE'
| 'START' | 'MID' | 'END' | 'APPROX' |
'NONE' (default, if absent, is 'NONE')
anchorTimeID ::= IDREF
anchorTimeID ::= TimeID
```

4.2.3 Signals

Signals get the `SIGNAL` tag. As can be seen, its attributes only express identification information.

```
attributes ::= sid
sid ::= s<integer>
```

4.2.4 Temporal relations

Temporal relations are expressed by the `TLINK` tag. In Chapter 3 we have presented the classification which is expressed by the `relType` attribute. Nonetheless, in total, temporal relations can get the following attributes.

```
attributes ::= [lid] [origin] (eventInstanceID | timeID)
[signalID] [syntax]
(relatedToEventInstance | relatedToTime) relType
lid ::= ID
{lid ::= LinkID
LinkID ::= l<integer>}
origin ::= CDATA
eventInstanceID ::= IDREF
{eventInstanceID ::= EventInstanceID}
timeID ::= IDREF
{timeID ::= TimeID}
signalID ::= IDREF
{signalID ::= SignalID}
relatedToEventInstance ::= IDREF
{relatedToEventInstance ::= EventInstanceID}
relatedToTime ::= IDREF
{relatedToTime ::= TimeID}
relType ::= 'BEFORE' | 'AFTER' | 'IBEFORE' | 'IAFTER' |
'INCLUDES' | 'IS_INCLUDED' | 'MEASURE' |
'SIMULTANEOUS' | 'BEGINS' | 'BEGUN_BY' |
            'ENDS' | 'ENDED_BY' | 'IDENTITY'
syntax ::= CDATA
```

4.2.5 Subordination relations

Subordination relations are normalised by the `SLINK` tag. The classification is expressed by the `relType` attribute, but it can also take the following attributes:

```
attributes ::= [lid] eventInstanceID
[signalID] subordinatedEventInstance relType [syntax]
lid ::= ID
{lid ::= LinkID
```

```
LinkID ::= l<integer>}
eventInstanceID ::= IDREF
{eventInstanceID ::= EventInstanceID}
subordinatedEventInstance ::= IDREF
{subordinatedEventInstance ::= EventInstanceID}
signalID ::= IDREF
{signalID ::= SignalID}
relType ::= 'MODAL' | 'EVIDENTIAL' | 'NEG_EVIDENTIAL'
| 'FACTIVE' | 'COUNTER_FACTIVE' | 'CONDITIONAL'
syntax ::= CDATA
```

4.2.6 Aspectual relations

Aspectual relations are expressed by the `ALINK` tag and get these attributes, along with the `relType` attribute which expresses the classification.

```
attributes ::= [lid] eventInstanceID [signalID]
relatedToEventInstance relType [syntax]
lid ::= ID
{lid ::= LinkID
LinkID ::= l<integer>}
eventInstanceID ::= ID
{eventInstanceID ::= EventInstanceID}
signalID ::= IDREF
{signalID ::= SignalID}
relatedToEventInstance ::= IDREF
{relatedToEventInstance ::= EventInstanceID}
relType ::= 'INITIATES' | 'CULMINATES' | 'TERMINATES'
| 'CONTINUES' | 'REINITIATES'
syntax ::= CDATA
```

4.3 EusTimeBank corpus

EusTimeBank is a corpus that contains temporal information. It is composed by three subcorpora:

- **FaCor**: a 25 news document corpus on the closure of a company written originally in Basque.

4.1 Table – EusTimeBank subcorpora

	Technical data	**Origin**	**Uses**
FaCor	25 documents 6,016 tokens Originally in Basque	Berria and Argia journals	Guideline definition 9 doc. in EusTimeBank gold
EusMEANTIME	120 documents 31,223 tokens translations	MEANTIME	Guideline definition 51 doc. in EusTimeBank gold
WikiWarsEU	19 documents 35,866 tokens Originally in Basque	Basque Wikipedia	EusHeidelTime development

- **WikiWarsEU**: this corpus contains the corresponding Basque Wikipedia articles on 17 of the 20 wars in WikiWars (Mazur and Dale, 2010). The documents are historical texts and have been written by non professional authors or translators.

- **EusMEANTIME**: it is the translation to Basque of the MEANTIME Corpus (Minard *et al.*, 2016), which contains 120 economy news documents. More precisely, it contains 30 documents for each of the 4 topics (Apple, Boeing, American motor companies and stock market).

The information of those three subcorpora is presented in Table 4.1.

In order to train and test automatic tools, we have gathered the Basque TimeBank gold standard corpus of 60 news documents in Basque annotated following EusTimeML. 30 documents have been used for training purposes, 15 documents for development and 15 documents for evaluation. In what concerns the source of these documents, 51 proceed from EusMEANTIME, the Basque version of the MEANTIME corpus Minard *et al.* (2016), and 9 documents from the FaCor corpus, a corpus we gathered from two Basque newspapers. All the documents have been annotated following EusTimeML and the information they dontain is summarised in Table 4.2.

All the documents in the Basque TimeBank are NAF formatted files Fokkens *et al.* (2014) (NLP Annotation Format) that contain stand-off annotations. The files were annotated using CAT (Content Annotation Tool) Lenzi *et al.* (2012) which accepts raw or tokenized texts as input and allows the direct exportation of labelled texts.

4.2 Table – Token and annotation amount from the EusTimeBank gold standard corpus

	Train (+ Development)	**Test**
Token amount	13.321	3.329
TIMEX3	448	137
SIGNAL	98	26
EVENT	2583	672
ALINK	50	10
SLINK	398	142
TLINK	3310	749

4.4 Corpus building methodology

As mentioned in Section 4.1.2, annotation guidelines and corpora have been annotated simultaneously. The development of the Basque TimeBank corpus was a two-step procedure:

- First, a limited number of events, time expressions, signals and links were annotated following the first version of the EusTimeML guidelines. In order to evaluate the annotation decisions, we have conducted a series of manual annotation experiments as described in Section 4.5.

- Second, the guidelines were updated based on the grammatical reanalysis of the annotations in the first step, and then the 60 documents that constitute the corpus were fully annotated. Although most of the annotation was done manually, we used EusHeidelTime (Altuna *et al.*, 2017a) for semiautomatic time expression annotation in order to hasten the creation of the corpus.

4.5 Experiments

We have conducted a series of experiments in order to evaluate the annotation guidelines. Additionally, these experiments have been useful for the creation of the manually annotated corpus.

4.5.1 Time expressions and signals

First a time expression and signal annotation experiment has been performed (Altuna *et al.*, 2014a). Following the methodology of the NewsReader project (Agerri *et al.*, 2014), three annotators (A, B and C) have annotated time expressions and signals in four documents (56 sentences). Texts are from the FaCor corpus.

Annotations have been done following the first version of the EusTimeML guidelines (Altuna *et al.*, 2014b) and time expressions and signals have been fully annotated; that is to say, all the attributes have been given a value. In order to measure the quality of the guidelines, inter-annotator agreement has been measured and individual annotations have been discussed.

4.5.1.1 Inter-annotator agreement

We have measured inter-annotator agreement considering the amount of tags in text and the agreement in the extent of the tags. For that, we have employed Dice's coefficient (Dice, 1945) which measures whether the same tokens have been chosen for the same tag. The agreement has been measured in pairs to assess whether agreement is balanced among all annotators.

Annotator pairs	**Micro-average (Markable)**	**Micro-average (Token)**	**Macro-average (Markable)**	**Macro-average (Token)**
A – B	0,96	0,976	0,969	0,977
A – C	0,943	0,965	0,923	0,965
B – C	0,902	0,94	0,892	0,942
Total	0,935	0,96	0,928	0,961

4.3 Table – Inter-annotator agreement on time expressions (`TIMEX3`)

Agreement on tags can be seen in Tables 4.3 and 4.4. *Markable* shows the agreement on the tag length and *token* shows agreement on individual tokens. Results are high for time expressions, which shows that time expression recognition has been performed correctly. In what refers signals, since they are commonly single-token structures, *markable* and *token* values are equal.

Annotator pairs	Micro-average (Markable)	Micro-average (Token)	Macro-average (Markable)	Macro-average (Token)
A – B	0,75	0,75	0,79	0,79
A – C	0,556	0,556	0,479	0,479
B – C	0,444	0,444	0,393	0,393
Total	0,583	0,583	0,554	0,554

4.4 Table – Inter-annotator agreement on signals (`SIGNAL`)

Nonetheless, it cannot be forgotten that results for signals are considerably lower.

Apart from the `TIMEX3` tags in text, five other non-consuming tags have been created for the annotation of begin and end points of durations. The agreement on those has been low for annotator A has tagged five, annotator B has annotated one and annotator C has not created any non-consuming tags. Taking into account all the annotations, the agreement between the three annotators for time expressions has been 66.7 %. Regarding signals, the agreement of the three annotators has only reached 31.2 %.

We have also measured the inter-annotator agreement for the most relevant attributes of time expressions. Agreement for `type` and `value` attributes is shown in Table 4.5.

Annotator pairs	A–B	A–C	B–C	General
`type`	0,76	0,64	0,75	0,55
`value`	0,76	0,68	0,46	0,45

4.5 Table – Inter-annotator agreement on `type` and `value` attributes

After analysing the results we have conveyed that the main reasons for disagreement have been the following:

- not identifying a time expression or a signal or not creating a non-consuming tag.

- annotating wrong tokens as time expressions.
- assigning the incorrect tag to a properly identified time expression.
- assigning different values to time-expression attributes.

The assessment of results has given us the clues for guideline improvement and we have conducted a discussion on them. The new decisions have been presented in the new version of the guidelines (Altuna *et al.*, 2016a).

4.5.2 Events

In order to prove the correctness and universality of the EusTimeML markup language and annotation guidelines for events, we have conducted a two-round annotation experiment on event identification and feature extraction (Altuna *et al.*, 2018a). The first was a preliminary experiment to evaluate and discuss the guidelines (Altuna *et al.*, 2014b). There was a guideline tuning period following this first round in which the annotating team added or corrected annotation features. Once the new guidelines (Altuna *et al.*, 2016a) were finished, a second annotation round was used to annotate a gold standard corpus of verbal event expression in Basque.

Both annotation efforts have been done using the CELCT Annotation Tool (Lenzi *et al.*, 2012), which is easily customizable and offers a range of interesting features for textual annotation such as inter-annotator agreement metrics.

4.5.2.1 First Annotation Round

For this first experiment about 172 events[2] were annotated. The annotated documents are part of FaCor. The events were annotated according to the EusTimeML guidelines (Altuna *et al.*, 2014b). Three annotators (A, B and C) took part in this annotation effort.

In this annotation round the agreement on event identification and extension were evaluated. The annotations of the three annotators were evaluated in pairs. Agreement levels ranged between 0.864 and 0.947 in weighted Dice's coefficient (Dice, 1945) depending on the annotator pair. The agreement level on the part of speech category, modality and whether events were aspectual were also evaluated.

[2]The amount of events varies among the annotations.

We found that events expressed by a single token were unanimously annotated in most of the cases. We also discovered an unexpectedly high agreement on events expressed by nouns and adjectives. However, although agreement in general was high, some annotation features were troublesome; we list them below:

- Some tokens were incorrectly considered events; mainly verbs taking part in time expressions and discourse markers.
- Some events on complex structures were neglected.
- Event expressions derived from verbs were not consistently given the same part of speech category.

In order to overcome these disagreement issues in the forthcoming annotation experiment, discussion on the annotation and guidelines among annotators was crucial; mainly in what referred to obscure annotation guidelines and ambiguous categories (namely, grammatical categories). Then we revisited Basque grammars and we updated the annotation guidelines adding more accurate information.

4.5.2.2 Second Annotation Round

After the grammatical reanalysis, a second annotation round was conducted. This second time, four annotators took part; three of them were familiar with EusTimeML and the CAT annotation tool and the fourth one had a deep knowledge on temporal annotation as well as the guidelines and the annotation tool. The annotation was done on 15 documents of EusMEANTIME. The first three annotators have annotated 115 sentences and their annotations have been compared to the fourth annotator's.

The number annotations for each annotator (A, B, C) and super-annotator (fourth annotator) and a counting of unanimously annotated events is given in Table 4.6. The numbers already show a relatively high agreement.

The main reason for disagreement was the difficulty to class some entities as events. In example (65) there is a linguistic form (*ekoizpena*) which expresses an event in the EusMEANTIME corpus and in example (66) there is the same form not expressing any event. This phenomenon was more pronounced in the cases in which a form refers both to a process and the final product of that process.

4.6 Table – Annotated events by each annotator and agreed events

Document sets	Annotator	Super-annotator	Agreed events
First (Ann. A)	96	74	69
Second (Ann. B)	394	418	358
Third (Ann. C)	95	99	84

(65) ***Ekoizpena*** *AEBra ekartzeko asmoa du.*
Production.DET.ABS USA.ALL bring.FIN intention.ABS has
'(He/She) intends to bring the **production** to the USA'.

(66) *Nekazariek euren* ***ekoizpena*** *salgai jarriko dute.*
Farmers.ERG their **production.ABS** to.be.sold put.FUT AUX
'Farmers will put their **production** on sale'.

State denoting events (*desengainatuta* (67)) were also a disagreement point. It was sometimes difficult to decide whether they were events as they do not always express an ongoing state but a very generic situation.

(67) *Oso* ***desengainatuta*** *gaude*
Very **disappointed** are
'(We) are very **disappointed**'.

In addition, in Basque the verb *egin* (to do) is used to focus events expressed by verbs. This verb, when working as a focaliser, does not offer any event information and, although it was stated not to annotate it, it has sometimes been annotated.

In what concerns event extension identification, results in Table 4.7 show a high agreement (Dice, 1945) on markable extent between annotators (the first three and the super-annotator). Markable extent agreement refers to the perfect overlap of the tags of two different annotators. Token extent agreement, instead, refers to the markable extent considering only the overlapping tokens. In our case both, markable and token extent, agreement results get the same values as markables have always a single-token extent. One may consequently deduce that all annotators have respected the single-token rule for event annotation in EustimeML guidelines.

4.7 Table – Event extent agreement results

Annotator pairs	**Micro-average (Markable)**	**Micro-average (Token)**	**Macro-average (Markable)**	**Macro-average (Token)**
A – SA	0.812	0.812	0.819	0.819
B – SA	0.877	0.877	0.875	0.875
C – SA	0.866	0.866	0.883	0.883

4.8 Table – Unanimously annotated POS

Event annotation	**A-SA**	**B-SA**	**C-SA**
Verbs	45	242	51
Nouns	14	58	33
Adjectives	1	3	1
Advebrs	0	9	2
Pronouns	0	1	0
Other	0	0	0
TOTAL	60 (87%)	311 (87%)	77 (92%)

As shown in Table 4.8, a rather high agreement on the grammatical category of events has been reached. Most of the disagreement is due to one of the annotators not giving any value to an event or forgetting to change the default value. However, some other disagreement is due to grammatical reasons:

- Verbal nouns ended with *-tea/-tzea* have been annotated as nouns and verbs.
- Participles with a relative mark *-tako/-riko* have been annotated as adjectives and verbs.
- Some adverbs have been considered part of the verb form and have been given a verb value or an "other" value.

The modal verbs unanimously annotated by the first three annotators and the gold standard can be seen in Table 4.9. Modal events have been easy to

4.9 Table – Modality agreement results

Modal event annotation	**A-SA**	**B-SA**	**C-SA**
Behar	0	5	2
Nahi	0	3	1
Ahal	3	4	0
TOTAL	3	12	3

identify, since there is little variation on the modality expressing forms. Moreover, there is virtually no possibility of confusedly giving a wrong value to a modal event expression as they have very distant meanings. Although the number of modal events is low, the result analysis has shown that mistakes in the annotation were due to annotators' mistakes during the annotation; not to wrong perceptions of those events.

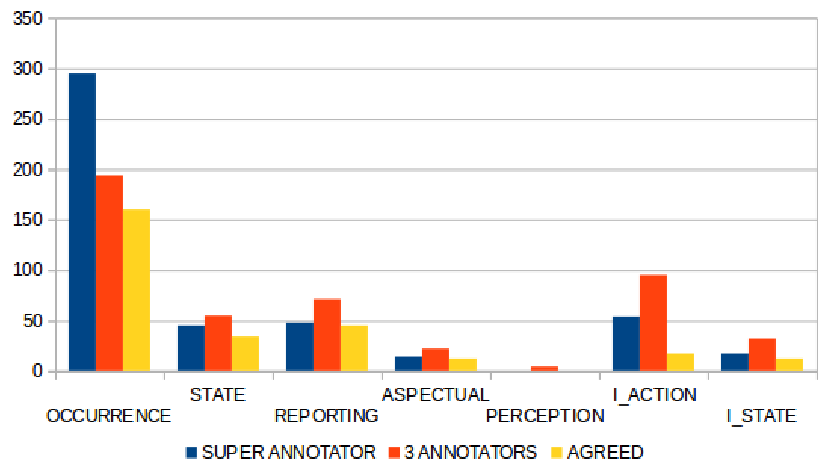
Finally we have measured the agreement on event category. The results are not as high as expected (A-SA: 58%, B-SA: 56% and C-SA: 49%), however, it is to mention that the agreement strongly varies between categories. Reporting and aspectual events have been easily identifiable, despite the fact that some have been incorrectly annotated presumably by mistake in many of the cases. Occurrence and intensional actions, instead, have been a major matter for disagreement. The agreement can be seen in Figure 4.1.

From a thorough analysis of the agreement, we have noticed that the documents that have been annotated later get higher agreement in event categorisation. Therefore, one may deduce that the more the training the better results in categorisation.

4.5.2.3 Final Guideline Tuning

Once we have analysed the annotation results, we have dropped some conclusions and have made some decisions:

- The more trained and familiar with the task an annotator is, the less mistakes will make and the higher agreement will achieve.
- In order not to forget filling or saving the attribute values, means for it will be designed.



4.1 Figure – Agreement on event category

- Event identification and part-of-speech categorisation do not seem difficult to master.
- Although modal events have been correctly annotated in general, we expect further discussion and training on them to improve the results.
- Event categorisation agreement has been lower than expected. Although some categories seem easier to assign, we will set a new analysis guideline tuning period for the most conflictive.

After the corrections to the 15 annotated documents are done, the trained annotators will continue enlarging the gold standard corpus, as well as annotating more temporal structures such as time expressions, temporal linking constructions and temporal relations.

4.5.2.4 Factuality annotation

Factuality information is very useful in temporal information processing since factuality helps identifying whether events in texts have happened in reality or not. That feature is specially useful in timeline creation as only events that have happened should appear in those.

After defining the annotation guidelines for factuality, we conducted an annotation experiment to measure the quality of the annotation decisions (Altuna *et al.*, 2018b). Two annotators took part in this experiment. They were asked to fully annotate the events so as to use the EusTimeML information to determine the factuality value. They were also asked to use world knowledge to resolve factuality. In total 734 events (out of 787 or 818) were annotated by both annotators and the factuality referring attributes in the agreed ones were analysed.

Table 4.10 shows the accuracy and κ values for the attributes that convey factuality information. As one can see, accuracy is rather high for most of the attributes and κ shows also a high agreement. The lower κ values are a consequence of a large quantity of certain categories. In fact, some values, such as the *certain* or *factual* values for certainty and factuality are very frequent in our corpus since news narratives tend to represent facts and that conditions the κ values.

4.10 Table – Inter-annotator agreement results for factuality annotation

	Polarity	**Certainty**	**Special Cases**	**Factuality**
Accuracy	0.98	0.89	0.95	0.77
κ	0.68	0.24	0.29	0.53

Analysing the disagreement has given us better knowledge about factuality annotation. Most of the mistakes were due to too loose definitions of the guidelines and were corrected in a guideline discussion session. In addition, we expect that i) redefining the `UNCERTAIN` and `UNDERSPECIFIED` values for certainty, ii) defining the boundaries of the generic statement and iii) better analysing the focus of the negation will help us define more accurate guidelines.

(68) *20 milioi dolar arteko laguntza* ***emateko*** *prest dago.*
20 million dollar until.REL help **give** ready is.
'(It) is ready to **give** up to 20 million dollar help.'

To illustrate this, *emateko* ("to give") in example (68) has been assigned `UNCERTAIN` and `UNDERSPECIFIED` by the annotators. It is stated in the guidelines that the events that express an aim will condition the certainty value of the subordinated event (`UNCERTAIN`). Nonetheless, *prest dago* ("is ready") is not a clear volition expression and was wrongly annotated by one of the annotators.

4.5.3 Relations

We have conducted two experiments for relation annotation guideline evaluation. In the first we have assessed the guidelines and in the second we have evaluated whether the changes have been effective.

Three annotators have taken part in the first annotation round. Annotator A has created 102 realtions, annotator B has created 96 and annotator C, 87. From these, 43 relations have been annotated by all the annotators and 39 more have been annotated by at least two annotators. The agreement on those 43 relations is displayed in Table 4.11.

4.11 Table – Inter-annotator agreement in relation annotation (1st round)

Annotator pairs	**Relation agreement**	**Relation classification agreement**
A – B	7	2
A – C	18	10
B – C	14	13
ABC	43	17

As can be seen in Table 4.11, relations have been given the same category by the three annotators 17 times out of 43. In what concerns, two-annotator agreement, the same classification has been given to the relations 25 times out of 39. It is to be pointed out that annotators B and C share more annotations than the ones they share with A. A more in-deep analysis of the annotations has shown that:

- 12 annotator A relations have no class, presumably for problems when saving data.
- For subordination links we have identified confusion between evidential and factive classes.
- For temporal links, semantically close categories have been a main reason for different classification. *E.g.* `BEFORE` vs `IBEFORE` or `SIMULTANEOUS` vw `INCLUDES`.

A guideline discussion period has been opened to confront the issues arisen in the first annotation effort. Then, in order to validate the new decisions, we have conducted a second annotation experiment.

In the second annotation round, two annotators have re-assessed the guidelines. For that they have annotated the relations in a document. The first annotator has created 71 links while the second has created 61. Those 61 have been annotated by both annotators and the category agreement in those has been measured. The relations created have been summarised in Table 4.12.

4.12 Table – Inter-annotator agreement in relation annotation (2nd round)

	TLINK		SLINK	
	Annotated	Agreed	Annotated	Agreed
Annotator A	68	59	3	2
Annotator B	59	59	2	2

In what concerns category agreement, the subordination relations annotated by both annotators have been given the same category (2 out of 2) and in the case of temporal relations agreement reaches 92% (54 out of 59). For the relations between events and the document creation time (DCT), agreement reaches 96%.

When analysing individual responses, we have seen that the identity class for temporal relations has been the major disagreement point. Nonetheless, the visible improvement in inter-annotator agreement results shows the guideline improvement has been effective.

4.6 Summary

In this cahpter we have gone through the steps for the creation of an annotated corpus. First, we have defined the EusTimeML mark-up language to normalise the temporal constructions and relations identified in Chapter 3. We have assessed the annotation decisions by means of some experimentation in which human annotators have annotated a series of documents. Inter-annotator agreement has been measured and disagreement has been assessed. Once the annotation guidelines have reached their final version, the EusTimeBank corpus has been annotated manually. This corpus has been used as a gold standard for the development and evaluation of the processing tools presented in the next chapter.

TOOLS AND APPLICATIONS

5 Temporal information precessing tools: EusHeidelTime and KroniXa

After creating the basic resources for temporal information extraction and normalisation, we have built advanced tools that take advantage of that information for temporal information processing. First we present EusHeidelTime, the rule-based tool for time expression extraction and normalisation. Secondly, we present KroniXa, the advanced tool for automatic timeline creation. These two tools have been trained and evaluated using EusTimeBank and deal with EusTimeML annotated data.

5.1 EusHeidelTime

The EusHeidelTime tool for time expression extraction and normalisation in Basque was presented in the *XXXIII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* (SEPLN 2017) and then published in the *Procesamiento del Lenguaje Natural* journal (vol. 59). The paper describes the process of the creation of the tool, its features and the evaluation effort.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 15-22 recibido 30-03-2017 revisado 08-05-2017 aceptado 18-05-2017

EusHeidelTime: Time Expression Extraction and Normalisation for Basque

EusHeidelTime: extracción y normalización de expresiones temporales para el euskera

Begoña Altuna, María Jesús Aranzabe, Arantza Díaz de Ilarraza
Universidad del País Vasco/Euskal Herriko Unibertsitatea
Manuel Lardizabal, 1, 20018 Donostia
{begona.altuna,maxux.aranzabe,a.diazdeilarraza}@ehu.eus

Abstract: Temporal information helps to organise the information in texts placing the actions and states in time. It is therefore important to identify the time points and intervals in the text, as well as what times they refer to. We developed EusHeidelTime for Basque time expression extraction and normalisation. For it, we analysed time expressions in Basque, we created the rules and resources for the tool and we built corpora for development and testing. We finally ran an experiment to evaluate EusHeidelTime's performance. We achieved satisfactory results in a morphologically rich language.
Keywords: Time expressions, information extraction, normalisation

Resumen: La información temporal ayuda a organizar la información textual situando las acciones y los estados en el tiempo. Por eso, es importante identificar los puntos e intervalos temporales en el texto, así como los tiempos a los que estos se refieren. Hemos desarrollado EusHeidelTime para la extracción y normalización de expresiones temporales para el euskera. Para ello, hemos analizado las expresiones temporales en euskera, hemos creado las reglas y recursos para la herramienta y hemos construido un corpus para el desarrollo y la evaluación. Finalmente, hemos realizado un experimento para evaluar el rendimiento de EusHeidelTime. Hemos conseguido resultados satisfactorios en una lengua con morfología rica.
Palabras clave: Expresiones temporales, extracción de información, normalización

1 Introduction

Temporal information is a core resource for textual organisation as it structures the discourse along a temporal axis. Its extraction and normalisation is useful and relevant in text comprehension and generation for tasks such as text summarisation (Aramaki et al., 2009), chronology creation (Bauer, Clark, and Graepel, 2015), event prediction (Radinsky and Horvitz, 2013) and event forecasting (Kawai et al., 2010).

Temporal information is composed by the events that happen or occur, the times those events happen in and the relations among those events and times. However, in this work we focus on time expression processing. Time expressions refer to a point in time in which an event takes place, starts or ends, or the duration of an event. For time expression processing, time expressions in texts must be marked and normalised and their features are extracted following a mark-up scheme. The corpora annotated with temporal information can be used for training machine-learning systems or as a gold standard to evaluate the performance of the tools.

Many tools and resources were developed to fulfill the task of identifying and normalising temporal information. On one hand, mark-up languages for temporal information annotation and annotated corpora were created, *e.g.* TimeML (TimeML Working Group, 2010), which was taken as an annotation standard and the TimeBank corpus (Pustejovsky et al., 2006). On the other hand, systems for temporal information extraction and normalisation were developed employing: i) machine-learning methods, e.g. GUTime (Verhagen and Pustejovsky, 2008) and TIPSem (Llorens, Saquete, and Navarro, 2010) ii) rule-based approaches such as CTEMP (Wu et al., 2005) and HeidelTime (Strötgen and Gertz, 2013) and iii) hy-

ISSN 1135-5948

brid tools, for example, TempEX (Mani and Wilson, 2000) and KTX (Jang, Baldwin, and Mani, 2004).

For our experimentation, we analysed Basque time expressions (Section 2), we created the EusTimeBank annotated gold standard corpus (Section 3), we integrated the HeidelTime parser in the Basque processing pipeline and we adapted and created the linguistic resources the system needs (Section 4). Finally, we conducted an annotation experiment (Section 5) and an error analysis for the evaluation of our tool's performance (Section 6). Some final remarks are given in Section 7.

2 *Time expressions in Basque*

We analysed Basque time expressions following (Bittar, 2010) and we have identified five different time expression types:

- dates: expressions referring to a particular period based on the Gregorian calendar, *e.g.* *martxoaren 8a* (8^{th} of March).
- times: expressions that refer to a particular subdivision of the day, *e.g.* *bostak* (five o'clock).
- durations: these expressions refer to an extended period of time *e.g.* *hiru aste* (three weeks).
- frequencies: these constructions express the regularity or re-occurrence of an event *e.g.* *egunero* (every day).
- temporal quantifications: expressions that consist in the quantification of a temporal unit *e.g.* *egunean 8 ordu* (8 hours a day).

These time expressions are classified in TimeML in four categories: date, time, duration and set (for frequencies and temporal quantifications). All time expressions are annotated with `TIMEX3` tag in TimeML and its features are normalised by means of a `DATE`, `TIME`, `DURATION` or `SET` type attribute, an ISO-8601 normalised value, as well as other attributes.

We annotated the time expressions in Basque following the EusTimeML guidelines[1], the adaptation of TimeML for Basque, which were used for the annotation of the sentence in (1) as can be seen in Figure 1.

[1] `https://addi.ehu.es/handle/10810/17305`

The time expression (*Iaz*, Last year) appears along with its class (`DATE`) and normalised value (`2016`). An event (*fakturatu zituzten*, turned over) is also displayed as well as the relation between the time expression and the event: the event is included in the time point the time expression refers to.

(1) Iaz 1.167 milioi euro
Last.year 1,167 million euro
fakturatu zituzten.
turn.over 3.PL.PAST

Last year they turned over 1,167 million euros.

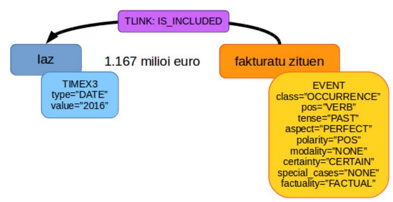


Figure 1: Annotation of example (1) following EusTimeML

An extended description of an annotation process for time expressions is described in Altuna, Aranzabe and Díaz de Ilarraza (2014).

3 *EusTimeBank*

EusTimeBank is a corpus that contains temporal information. It is composed by three subcorpora:

- **FaCor**: a 25 news document corpus on the closure of a company written originally in Basque.
- **WikiWarsEU**: this corpus contains the corresponding Basque Wikipedia articles on 17 of the 20 wars in WikiWars (Mazur and Dale, 2010). The documents are historical texts and have been written by non professional authors or translators.
- **EusMEANTIME**: it is the translation to Basque of the MEANTIME Corpus (Minard et al., 2016), which contains 120 economy news documents.

The documents were manually annotated using the CELCT Annotation Tool (Bartalesi

Lenzi, Moretti, and Sprugnoli, 2012) and following the EusTimeML mark-up scheme. A selection of 67 documents was used for development and evaluation purposes of the temporal information processing tools we created: 25 from FaCor, 17 from WikiWarsEU and 25 from EusMEANTIME. We provide the amount of `TIMEX3` tags, time expression tag, and the size of the annotated corpora for the experiment in Table 1.

Corpora	Size	
	Development (words/TIMEX3)	Test (words/TIMEX3)
FaCor	4,503/142	1,513/59
EusMEANTIME	5,247/200	1,258/53
WikiWarsEU	22,299/701	7,399/343
TOTAL	**32,049/1043**	**10,170/455**

Table 1: Size of the annotated corpora

4 The EusHeidelTime tool

We adapted HeidelTime for Basque time expression extraction and normalisation due to the re-usability of the source code and the easiness for linguistic resource creation, as well as the lack of large annotated corpora in Basque. The rules, patterns and normalisation information are language dependent, while the source code is common to all languages. This allows an easy adaptation to new languages. Apart from English, HeidelTime was used for time expression extraction and normalisation in German (Strötgen and Gertz, 2011), Dutch (van de Camp and Christiansen, 2013), French (Moriceau and Tannier, 2014) and Croatian (Skukan, Glavaš, and Šnajder, 2014) among others.

4.1 Integration of EusHeidelTime in the Basque pipeline

HeidelTime was originally developed as a UIMA (Unstructured Information Management Architecture) (Ferrucci and Lally, 2004) component and integrated as a document processing pipeline. As explained in Strötgen and Gertz (2010), for English, the UIMA pipeline contains a sentence splitter and tokenizer and an OpenNLP PoS tagger to be used by the temporal tagger. For Basque, instead, we defined and integrated the temporal tagger in a document processing pipeline, ixa-pipe-pos-eu, following the Otegi et al. (2016) approach. More specifically, our pipeline (Figure 2) includes, a tokeniser, a robust and wide-coverage morphological analyser and a PoS tagger for Basque and the EusHeidelTime temporal tagger. Ixa-pipe-pos-eu is part of ixaKat[2], a modular chain of NLP tools for Basque where all the modules read and write NAF (Fokkens et al., 2014), a linguistic annotation format designed for complex NLP pipelines. The temporal tagger has these features too, but the core of the module is based on HeidelTime. Thus, the integration of the temporal tagger in a UIMA pipeline would be quite straightforward. In addition, we parametrised the temporal tagger so that it is possible to obtain the temporal information in NAF or TimeML format (Figure 3), which was used for the evaluation of the tool. TimeML format implies XML documents containing XML `TIMEX3` tags that mark time expressions and offer information about their type, normalised value and modifier information if any.

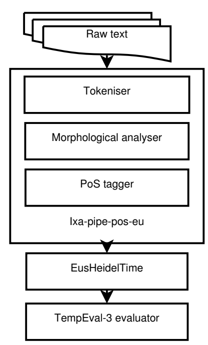


Figure 2: Diagram for time expression extraction in Basque

```xml
<?xml version="1.0" ?>
<TimeML xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:noNamespaceSchemaLocation="http://
timeml.org/timeMLdocs/TimeML_1.2.1.xsd">
<DOCID>1380-World_largest_passenger_
airliner_makes_first_flight.txt.xml</DOCID>

<DCT><TIMEX3 tid="t0" type="TIME"
value="2005-04-27" temporalFunction="false"
functionInDocument="CREATION_TIME">2005-04-27
</TIMEX3></DCT>

<TEXT>
Munduko bidaiari-hegazkinik handienak estreinako
hegaldia egin du . <TIMEX3 type="DATE"
value="2005-04-27" tid="t1">2005eko apirilaren 27a</
TIMEX3> . A380 hegazkina <TIMEX3 type="DATE"
value="2005-01" tid="t2">2005eko urtarrilean</
TIMEX3> aurkeztu zuten .
```

Figure 3: An EusHeidelTime annotation example

[2]http://ixa2.si.ehu.es/ixakat/

4.2 Adapting language dependent resources

As mentioned before, we adapted HeidelTime to Basque. For this, we created three resource sets:

- **Rules**: rules contain the patterns to be extracted and their normalisation, as well as value modifiers and constraints, *e.g.* part-of-speech (PoS) constraint of a token in the pattern. Figure 4 shows a rule for patterns as "Datorren urteko urtarrilean" (On January next year). The rule contains a name (`RULENAME`), the pattern to match (`EXTRACTION`) and the normalisation pattern (`NORM_VALUE`) that will turn the text segment into a TimeML normalised value. There are four rule sets (dates, durations, sets and times), which correspond to the different types of time expressions in EusTimeML[3].
- **Patterns**: pattern resources are regular expressions that gather together patterns of the same kind, *e.g.* months, weekdays etc.
- **Normalisation files**: these contain normalised values of the time expressions. Figure 5 shows weekdays and the normalised value for each string.

For the development of resources, two main features of Basque were taken into account. First, as Basque is agglutinative, the rich morphology as well as the morphotactics were added. Second, since it is a head-final language, many acquired patterns were reversed to accommodate its syntax. As a consequence, some resources, namely a significant quantity of rules, were created from scratch to accommodate specific Basque temporal constructions. Nonetheless, some rules and patterns for Basque (mainly numeric expressions) were directly transferred from other languages and most of the patterns (*e.g.* month names, weekday names) were translated.

Apart from the relevant linguistic features, the internal architecture of HeidelTime was also taken into account. HeidelTime applies the rules sequentially and when more than one rule matches a time expression,

[3]Rules for intervals were disregarded as intervals are not defined in EusTimeML.

```
//Adibidea: datorren urteko urtarrilean

RULENAME="Data_erl_datorren_year_month",
EXTRACTION="%reDatorren urte%Singularra
%reMonth%reSingularra",
NORM_VALUE="UNDEF-next-year-%normMonthFull(group(4))"
```

Figure 4: An EusHeidelTime rule

```
"[Aa]stelehen","1"
"[Aa]stearte","2"
"[Aa]steazken","3"
"[Oo]stegun","4"
"[Oo]stiral","5"
"[Ll]arunbat","6"
"[Ii]gande","7"
```

Figure 5: Weekday pattern normalisation values

it chooses one following this order: dates, times, durations, sets and intervals[4]. The rules in each category are also ordered and read sequentially.

In Table 2 one can see the amount of resources created for EusHeidelTime. The quantity of rules is due to i) the intention to avoid optional elements in the rules, and ii) grammatical aspects of Basque as word order restrictions with the numeral determiner *bat* (one). This led to defining two different rules for strings containing numerals.

Resource type	**Quantity**
Rules	
DATE	142
TIME	64
DURATION	101
SET	6
Pattern files	58
Normalisation files	29

Table 2: EusHeidelTime resources

5 *Experimentation*

We processed a 17 document set of the test corpora (Section 3) and we evaluated the output against our gold standard annotation to evaluate the developed resources[5]. We followed the TempEval-3 (UzZaman et al., 2013) criteria to evaluate the performance of our tool. In Table 3 we present the results for each corpus in these four fields:

- Strict match: the extent of the obtained temporal expression and the correspond-

[4]We do not apply rules for intervals in Basque since they are not a category in EusTimeML.

[5]EusHeidelTime resources and corpora for replication can be downloaded from http://ixa2.si.ehu.es/eusheideltime/

	FaCor			**EusMEANTIME**			**WikiWarsEU**		
	P	R	F1	P	R	F1	P	R	F1
Strict match	79.39	83.64	81.42	81.4	74.47	77.78	77.98	87.8	82.6
Relaxed match	87.93	92.73	90.27	93.02	85.11	88.89	82.67	93.09	87.57
Value			58.41			64.44			74.57
Type			83.19			82.22			86.81

Table 3: Evaluation results for EusHeidelTime

ing one in the gold standard overlap perfectly.

- Relaxed match: partial overlap between the automatically obtained expression and the corresponding one in the gold standard.
- Value: the normalised value of the automatically obtained and the gold standard match.
- Type: the type of the automatically obtained and the gold standard match.

For strict and relaxed matches, precision (P), recall (R) and F-measure (F1) were calculated and for value and type the F-measure was given, in order to be comparable to the TempEval-3 results.

The performance of our tool is in the same range of the best systems for English in TempEval-3. We achieved a F1 of 81.42 for strict match in FaCor and 82.6 in WikiWarsEU, which are close to the best performing tool in TempEval-3, ClearTK-1,2 (Bethard and Martin, 2013) (82.71) and HeidelTime for English (81.34). In what concerns the relaxed match, for which we achieved a F1 score of 90.27 in FaCor corpus, we also get close to the best performing tools, NavyTime-1,2 and SUTime (90.32) and HeidelTime (90.30).

We also got similar results for news (FaCor and EusMEANTIME) and for historical texts (WikiWarsEU). Nevertheles, a high rise on the F1 for value (74.57) can be seen for historical texts, presumably because of the large amount of the absolute dates.

6 Error analysis

We conducted an analysis to identify the nature of the different errors. We classified manually the errors in 8 categories (Table 4) and we tried to solve them.

As one can see from Table 4, the errors identified are quite heterogeneous, but can be divided in human-made and processing errors. The first group is formed by i) the absence of rules for certain time expressions. *E.g.*, “hondarrean” in (2) is not a common term to express the end and we did not consider it when creating the rules; and ii) the too general rules led to false positives. For example, we created restricting rules to treat polysemy as in “urri” (October/scarce), “hil” (month/dead) and “lehen” (past/first) among others, but they proved not to be sufficient. Finally, iii) the errors in the gold standard, mainly typos. These rules can be fixed by adding or correcting the rules and the errors in the gold standard. However, we are aware that we will not be able to address all the possible time expressions in Basque.

Error	**Quantity**
Absence rule	24
Too general rules	21
Wrong gold standard	6
Wrong tokenisation	11
Wrong rule selection	18
Wrong resolution of relative date	21
Rule not performing well	18
Ambiguous reference	6

Table 4: Classification of errors

(2)
```
gold annotation: <TIMEX3
type="DATE" value="2014-07"
tid="t15">uztailaren
hondarrean</TIMEX3>
system annotation: <TIMEX3
type="DATE" value="2014-07"
tid="t15">uztailaren</TIMEX3>
-- relaxed match
```

In what concerns errors due to processing, we first noticed the errors due to wrong tokenisation. In example (3), the initials “(UTC)” were wrongly tokenised and this impaired the time expression from being identified although a rule for times containing “(UTC)” existed.

(3)
```
gold annotation not found in
system: <TIMEX3 type="TIME"
```

```
value="2008-09-18T08:00Z"
tid="t2">8:00etan ( UTC
)</TIMEX3>
```

In what refers to rule selection, as mentioned in section 4.2, the rules are applied sequentially and there is a hierarchy between categories. This has led to the wrong rule selection. In example (4) we got a partial match since the system privileged a date rule instead of a duration rule.

(4)
```
system annotation: <TIMEX3
type="DATE" value="PAST_REF"
tid="t7">lehen</TIMEX3>
gold annotation: <TIMEX3
type="DURATION" value="PT90M"
tid="t7">lehen 90 minutuetan
</TIMEX3> -- relaxed match
```

We also identified some rules not performing well. "Gaur" in example (5) is annotated as a generic present reference, although it refers to the exact date of "today". Both interpretations are possible, but HeidelTime systematically chooses the generic interpretation although the exact one is higher in hierarchy. This may be due to a mistake in the rule and needs further analysis.

(5)
```
system wrong value: <TIMEX3
type="DATE" value="PRESENT_REF"
tid="t5">gaur</TIMEX3>
```

Some annotation errors are much more difficult to correct. Those are the ones that i) involve relative time expressions or ii) ambiguous constructions that can only be resolved through world knowledge or a deep contextual comprehension. For the first, HeidelTime sets the last time expression annotated as a temporal anchor for the next. In example (6) the value for "bihar bertan" (tomorrow) is not well resolved as the temporal anchor is not the right one. The solution for the second is much more complicated because of the difficulty of adding world or contextual knowledge to automatic systems. It is virtually impossible to decide the real duration of "Epe laburreko" (short period) (7), since a short period can be considered hours, days or months in different contexts.

(6)
```
gold value: <TIMEX3
type="DATE" value="2014-10-31"
tid="t8">bihar bertan</TIMEX3>
system wrong value: <TIMEX3
type="DATE" value="2014-10-28"
tid="t8">bihar bertan</TIMEX3>
```

(7)
```
system wrong value:
<TIMEX3 type="DURATION"
value="PXD" tid="t3">epe
laburreko</TIMEX3>
```

After the error analysis, we will improve the rules correcting the problems identified.

7 Conclusions and future work

In this paper we presented an experiment on temporal expression annotation in Basque with EusHeidelTime, a rule-based tool based on HeidelTime. Considering Basque is a highly agglutinative and head-final language, we proved that HeidelTime can be used for languages with complex morphology. We also profited the modularity of HeidelTime and we added it to our Basque pipeline.

EusHeidelTime achieved results comparable to those obtained for English. Having reached F1 measures of circa 80% in strict match, we consider the resources created for this experiment are already adequate for the automatic annotation of temporal expressions, although that annotation will have to be supervised by a human annotator. We will also proceed to a final tuning of the resources to correct the flaws identified during the error analysis.

We achieved similar results both in news and historical documents. Therefore, we presume our tool can annotate documents of different domains. In the future, we aim to perform temporal annotation of clinical texts, due to the relevance of the temporal ordering of events in that field.

Temporal expression extraction and normalisation is only a part of a more extended work on temporal information annotation. It will be combined with event information processing and temporal relation processing for the creation of a system able to treat temporal information in its entirety.

Acknowledgments

This work was financed by the Basque Government scholarship PRE_2016_2_294.

References

Altuna, B., M. J. Aranzabe, and A. Díaz de Ilarraza. 2014. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. *Linguamática*, 6(2):13–24, Dezembro.

Aramaki, E., Y. Miura, M. Tonoike, T. Ohkuma, H. Mashuichi, and K. Ohe. 2009. Text2table: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 185–192, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bartalesi Lenzi, V., G. Moretti, and R. Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 333–338, Istanbul, Turkey. European Language Resources Association (ELRA).

Bauer, S., S. Clark, and T. Graepel. 2015. Learning to Identify Historical Figures for Timeline Creation from Wikipedia Articles. In L. Aiello and D. E. McFarland, editors, *SocInfo 2014 International Workshops, Barcelona, Spain, November 11, 2014, Revised Selected Papers*, volume 8852 of *Lecture Notes in Computer Science*, pages 234–243, Barcelona, Spain.

Bethard, S. and J. H. Martin. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In S. Manandhar and D. Yuret, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM) 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bittar, A. 2010. *Building a TimeBank for French: a Reference Corpus Annotated According to the ISO-TimeML Standard*. Ph.D. thesis, Université Paris Diderot, Paris.

Ferrucci, D. and A. Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348.

Fokkens, A., A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W. R. van Hage, and P. Vossen. 2014. NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 9, Reykjavik, Iceland.

Jang, S. B., J. Baldwin, and I. Mani. 2004. Automatic TIMEX2 Tagging of Korean News. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(1):51–65, March.

Kawai, H., A. Jatowt, K. Tanaka, K. Kunieda, and K. Yamada. 2010. Chronoseeker: Search engine for future and past events. In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*, ICUIMC '10, pages 25:1–25:10.

Llorens, H., E. Saquete, and B. Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 284–291, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mani, I. and G. Wilson. 2000. Robust Temporal Processing of News. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 69–76, Stroudsburg, PA, USA.

Mazur, P. and R. Dale. 2010. WikiWars: A New Corpus for Research on Temporal Expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 913–922, Stroudsburg, PA, USA. Association for Computational Linguistics.

Minard, A.-L., M. Speranza, R. Urizar, B. na Altuna, M. van Erp, A. Schoen, and C. van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Moriceau, V. and X. Tannier. 2014. French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Otegi, A., N. Ezeiza, I. Goenaga, and G. Labaka. 2016. A Modular Chain of NLP Tools for Basque. In P. Sojka, A. Horák, I. Kopeček, and K. Pala, editors, *Proceedings of the 19th International Conference on Text, Speech and Dialogue — TSD 2016, Brno, Czech Republic*, volume 9924 of *Lecture Notes in Artificial Intelligence*, pages 93–100. Springer International Publishing.

Pustejovsky, J., M. Verhagen, R. Saurí, J. Littman, R. Gaizauskas, G. Katz, I. Mani, R. Knippen, and A. Setzer. 2006. TimeBank 1.2. Technical report, Linguistic Data Consortium.

Radinsky, K. and E. Horvitz. 2013. Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 255–264. ACM.

Skukan, L., G. Glavaš, and J. Šnajder. 2014. HeidelTime.Hr: Extracting and Normalizing Temporal Expressions in Croatian. In *Proceedings of the Nineth Language Technologies Conference*, pages 99–103. Information Society.

Strötgen, J. and M. Gertz. 2010. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 321–324, Stroudsburg, PA, USA. Association for Computational Linguistics.

Strötgen, J. and M. Gertz. 2011. WikiWarsDE: a German Corpus of Narratives Annotated with Temporal Expressions. In H. Hedeland, T. Schmidt, and K. Wörner, editors, *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*, pages 129–134, Hamburg University.

Strötgen, J. and M. Gertz. 2013. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2):269–298.

TimeML Working Group. 2010. TimeML Annotation Guidelines version 1.3. Manuscript. Technical report, Brandeis University.

UzZaman, N., H. Llorens, J. F. Allen, L. Derczynski, M. Verhagen, and J. Pustejovsky. 2013. TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. In S. Manandhar and D. Yuret, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM) 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

van de Camp, M. and H. Christiansen. 2013. Resolving relative time expressions in Dutch text with Constraint Handling Rules. In *Revised Selected Papers of the 7th International Workshop on Constraint Solving and Language Processing - Volume 8114*, CSLP 2012, pages 166–177, New York, NY, USA. Springer-Verlag New York, Inc.

Verhagen, M. and J. Pustejovsky. 2008. Temporal Processing with the TARSQI Toolkit. In *22n d International Conference on on Computational Linguistics: Demonstration Papers*, COLING '08, pages 189–192, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wu, M., W. Li, Q. Lu, and B. Li. 2005. CTEMP: A Chinese Temporal Parser for Extracting and Normalizing Temporal Information. In R. Dale, K.-F. Wong, J. Su, and O. Y. Kwong, editors, *Natural Language Processing – IJCNLP 2005: Second International Joint Conference, Jeju Island, Korea, October 11-13, 2005. Proceedings*, pages 694–706, Berlin, Heidelberg. Springer.

5.2 KroniXa

KroniXa is a system for automatic creation of timelines from Basque texts. The system creates timelines from temporally annotated documents and uses temporal relation information and dependency relations to anchor events to the chronology. The work has been submitted to the *Journal of Information Processing and Management* for the special issue on narrative extraction from texts.

KroniXa: Timeline Creation from Basque Texts

Begoña Altuna, Ander Soraluze, María Jesús Aranzabe, Olatz Arregi, Arantza Díaz de Ilarraza

Ixa Group, University of the Basque Country
Manuel Lardizabal 1, 20018 Donostia
begona.altuna@ehu.eus

Abstract

Timelines extracted from texts allow humans to arrange events in chronological order. KroniXa creates timelines from texts written in Basque based on the information provided by EusHeidelTime, which identifies time expressions, and bTime, a system which captures events and temporal relations. An improvement of the KroniXa system is also proposed: temporal relations inferred from syntactic dependencies are added to the automatically extracted timelines. This system obtains F-scores of 24.36 and 31.47 when applied on raw texts and temporally annotated texts respectively.

Keywords: Timelines, Temporal information, Basque language, event ordering
2010 MSC: 00-01, 99-00

1. Introduction

Temporal information processing has been a subject of major interest among the Natural Language Processing (NLP) scholars for the last three decades. Since TimeML (Pustejovsky et al., 2003) became the *de facto* standard for temporal information annotation, many temporal information extraction and annotation tools as well as large corpora have been built. The availability of highly comprehensive and representative corpora has now open the possibility of developing tools that make use of the information in those corpora.

A first application of temporal information is improving information retrieval systems. For this task, temporal information in documents is taken into account for the enhancement of the relevance measuring of a query answer, in order to provide more relevant search hits. For example, Alonso et al. (2009) proposed using time expressions in documents to build document temporal profiles. InZeit (Setty et al., 2010), instead, reassigned the relevance value of a document according to the insightfulness of the time boundaries of the document.

A second application of temporal information is timeline creation. A timeline is a historical account of events ranged in chronological sequences. Hence, in order to build a timeline, events and the times they happen in have to be identified. Timelines have traditionally been created manually, which is very time consuming, as the creation of timelines require consulting and merging information from different sources. In the last too decades though, the attempts of automatically building timelines have multiplied.

Allan et al. (2001) built a system that chose the most relevant sentences in the corpus. They gave a relevance score to each sentence and built the summary with the top-scoring sentences. Chieu & Lee (2004) also attempted to summarize a large document collection obtained from a query-based search by placing sentences that reported "important" events related to the query along a timeline.

Evolutionary Timeline Summarization (ETS) (Yan et al., 2011) took into account temporal information in documents. In order to build the timeline, they obtained a collection of sentences related to a query and associated to their publishing dates. Tran et al. (2015), instead, focused on the date selection for timeline building. Their task consisted of determining the dates of the subevents in the time span of a main event.

In 2015 a cross-document event ordering task was launched in SemEval-2015 (Minard et al., 2015). In that task, events involving a certain entity were to be ordered chronologically based on cross-document event coreference and temporal relation extraction. The dataset consisted of four subsets of 30 documents on four different topics annotated following the NewsReader annotation guidelines (Tonelli et al., 2014) and containing also entity annotations.

For timeline evaluation, two approaches have been proposed. On one hand, in the SemEval-2015 task they evaluated the relations following the TempEval-3 metrics (UzZaman et al., 2012). On the other hand, Bauer & Teufel (2015) proposed a method which attempted to measure to which degree an automatically generated timeline contained semantic units found in gold-standard timelines. They proposed to evaluate Historical Content Units in which an event and its agent, patient, time and location were addressed.

In the case of Basque, temporal information extraction and processing has been a matter of interest in the last five years. In this period, the EusTimeML mark-up language (Altuna et al., 2016) has been developed and several tools for temporal information processing, such as EusHeidelTime (Altuna et al., 2017) and bTime (Salaberri Izko, 2017), have been created for temporal information in Basque.

Nevertheless, to the best of our knowledge, there has been no attempt on automatic extraction of timelines from Basque texts. In this paper, we present a first attempt on automatic timeline creation from Basque texts: the KroniXa tool.

The paper is organized as follows: in Section 2 we describe EusTimeML, the temporal information mark-up language for Basque, and EusTimeBank, the temporally annotated corpus for Basque. In Section 3 we introduce EusHeidelTime and bTime, the modules that obtain time expressions, events and temporal relations. Section 4 presents KroniXa, our timeline creation tool. The experimentation and results on automatic timeline building are detailed in Section 5. This section describes also the KroniXa enhancement effort by applying syntactic dependencies information. Finally, we describe and we discuss our approach in Section 6, and we conclude our work in Section 7.

2. EusTimeML and EusTimeBank

Temporal information has to be encoded and normalized in order to be accessible for temporal information processing tools. In NLP, information in text is made explicit using mark-up schemes, by means of which information in texts is tagged. EustimeML is a temporal information mark-up scheme for Basque that follows the TimeML (TimeML Working Group, 2010) philosophy and provides a mark-up scheme for events, time expressions, signals and temporal relations. These are the categories and their associated tags in EusTimeML:

- **events** (`<EVENT>`): situations that happen, occur, hold, or take place and states and circumstances in which something obtains or holds true Saurí et al. (2009);
- **time expressions** (`<TIMEX3>`): natural language phrases that directly refer to time points or intervals Ahn et al. (2005);
- **signals** (`<SIGNAL>`): any function words that suggest a particular temporal relationship TimeML Working Group (2010);
- **temporal relations** (`<TLINK>`): temporal relations created between two events, two time expressions or an event and a time expression Saurí et al. (2006).

EusTimeML also provides a set of attributes and attribute values to make the temporal information conveyed by those different elements explicit. Additionally, EusTimeML offers tags for subordination and aspectual relations, `<SLINK>` and `<ALINK>` respectively.

EusTimeML follows closely the TimeML scheme. It is XML-based and preserves most of the TimeML features. Events, time expressions and temporal relations follow the same classification and attributes and their values are directly transferred in most of the situations. Additionally, EusTimeML shares the single token annotation policy for events proposed in TimeML. Figure 1 shows the EusTimeML annotation for the *Iaz 1.167 milioi euro fakturatu zituen* ("Last year they turned over 1,167 million euros") sentence.

Nonetheless, there are some differences between EusTimeML and TimeML: i) factuality annotation has been added directly to EusTimeML (Altuna et al., 2018b), ii) signal annotation has been reanalyzed, since Basque is a highly inflectional language and temporal relation markers commonly appear attached

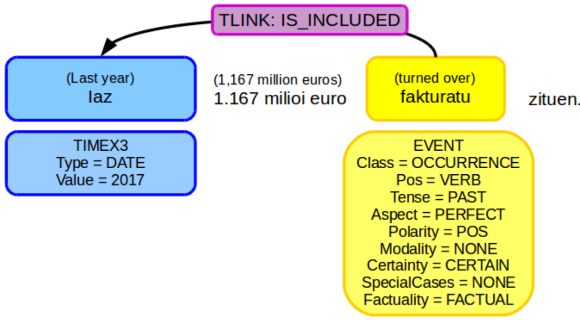


Figure 1: Annotation of *Iaz 1.167 milioi euro fakturatu zituen.* ("Last year they turned over 1,167 million euros")

to an event or a time expression, and iii) some attribute values (such as aspect and tense for verbal events) have been modified to be consistent with Basque grammar rules.

That third issue is of special relevance in our timeline creation effort as only events expressed by tensed verbs are to be linked to the document creation time (DCT). In Basque tensed verbs can be synthetic (1) or periphrastic (2). The events expressed by synthetic tensed verbs should not be a special hassle, since all the information is expressed by a single token. On the contrary, special attention has to be put on periphrastic verbs, for the semantic and aspectual information is expressed by the head while the tense—which places the event in time in relation to time—is expressed by the auxiliary. Due to the single token annotation policy for events, the information of unannotated tokens has also to be taken into account.

(1) NASAk urteak **daramatza** Martitzen ur bila.
NASA.ERG years.ABS **takes** Mars.INE water look.for
'NASA **has been** years looking for water in Mars.'

(2) UEFA 2016 Europar Txapelketa Frantzian **ospatuko** **da**.
UEFA 2016 European Championship France.INE **celebrate.FUT** **AUX**
'The UEFA 2016 European Championship **will be held** in France'

EusTimeML was employed in the annotation of EusTimeBank, a corpus for training and evaluation purposes. The EusTimeBank gold standard corpus contains 60 news documents in Basque annotated following EusTimeML. 30 documents have been used for training purposes, 15 documents for development and 15 documents for evaluation. The documents in the EusTimeBank are NAF[1] formatted files (Fokkens et al., 2014) that contain stand-off annotations and were annotated using CAT (Content Annotation Tool) (Bartalesi Lenzi et al., 2012).

In Table 1 we present the amount of documents, sentences, tokens, events, time expressions (TIMEX) and temporal links (TLINK) between tensed verbs and the document creation time (DCT) in the gold standard corpus. Although the corpus may seem small at first glance, the training corpus was inspired in the SemEval-2015 Task 4 trial corpus (Minard et al., 2015), which contains 30 documents, 464 sentences and 10,373 tokens.

Table 1: Basque TimeBank Gold Standard corpus

	Doc.	Sentences	Tokens	Events	TIMEX	TLINKs
TRAIN	30	509	8,794	1,903	332	940
DEVEL	15	201	3,424	680	116	358
TEST	15	174	3,329	672	137	347

3. Temporal information processing

The temporal information taken as input by KroniXa is captured and normalized by the integration and combination of two tools created for temporal information processing in basque. EusHeidelTime (Section 3.1) extracts and normalizes time expressions, while bTime (Section 3.2) identifies and classifies events and creates temporal relations between those and time expressions—more precisely document creation times (DCT).

[1]NLP Annotation Format

3.1. EusHeidelTime: Time expression processing

The annotation of time expressions is performed using the EusHeidelTime annotator (Altuna et al., 2017)—the Basque version of HeidelTime (Strötgen & Gertz, 2013). EusHeidelTime is a rule-based time expression tagger that has an absolute separation between the processing module and the linguistic resources. The HeidelTime processing module has easily been integrated in the Basque processing chain (Otegi et al., 2016) and only the language-dependent resources have been created *ex profeso*:

- **Rules** contain the patterns to be extracted and their normalization, as well as value modifiers and constraints, *e.g.* part-of-speech (PoS) constraint of a token in the pattern. The rule contains a name, the pattern to match and the normalization pattern that will turn the text segment into a TimeML normalized value. There are four rule sets (dates, durations, sets and times), which correspond to the different types of time expressions in EusTimeML.
- **Patterns** are regular expressions that gather together patterns of the same kind, *e.g.* months, weekdays, etc.
- **Normalization files** contain normalized values of the time expressions.

The main features of Basque taken into account when building the linguistic resources were i) the rich inflection and morpho-phonological system of Basque; and ii) the fact that Basque is a head-final language. As a consequence, morphotactics were considered when building patterns and many rules acquired from other language rule sets had to be reversed to accommodate Basque syntax. Some rules and patterns for Basque (mainly numeric expressions) instead could be directly transferred from other languages and most of the patterns (*e.g.* month names, weekday names) were translated. In total, EusHeidelTime employs 313 rules, 58 patterns lists and 29 normalization files.

EusHeidelTime achieves good results in temporal information extraction and normalization. As can be seen in Table 2, the time expression recognition module scores 82.4 points for F_1 when strict match evaluation is used and 90.99 points when partial or relaxed match is considered.

Table 2: Results of EusHeideltime

Element	Feature	Precision	Recall	F_1
TIMEX	extension (strict match)	87.27	78.05	82.4
	extension (relaxed match)	96.36	86.18	90.99
	type			84.98
	value			64.38

3.2. Temporal information processing with bTime

bTime is an end-to-end temporal annotator for Basque that captures events (Section 3.2.1) and, taking as input temporal expressions provided by EusHeidelTime, obtains the temporal relations between events expressed by tensed verbs and the document creation time (DCT) (Section 3.2.2). bTime employs machine learning methods which have been trained using the EusTimeBank corpus (Section 2). Figure 2 shows the system architecture.

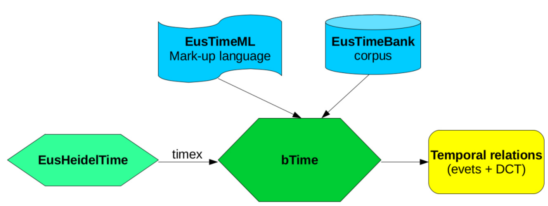


Figure 2: bTime system architecture.

3.2.1. Event information processing

The event processing module in bTime comprehends nine classifiers. One model addresses event extent identification, while the other eight have been built for event attribute classification. The identification of event extents is a binary classification task, for only the heads of the events are annotated in the corpus. The annotation of event attributes, instead, is a binary or a multi-class classification issue, depending on the number of classes for each attribute. The event classifiers use 24 features organized in four groups according to their type: lexical, Part-of-Speech related, syntactic and semantic (Salaberri Izko, 2017).

Most of the features have also been employed in similar tasks for other languages (Jung & Stent, 2013), whereas the remaining, such as the PoS subcategory, have been added for Basque event processing. In fact, PoS subcategory plays a crucial role in Basque event detection since the right disambiguation between proper and common nouns conditions the final results.

3.2.2. Temporal relation processing

The identification and classification of the temporal relations between events and the DCT is approached as a multiclass classification problem in which the temporal relation type (`relType`) is predicted. A preliminar temporal relation is created between all the events and the DCT and a class is assigned to each pair. More precisely, a relation type is assigned to the relations comprehending a tensed verb and a *fake* `NORELATION` relation type is assigned to the event-DCT pairs that do not contain tensed verbs. The set of features used to predict the type of temporal relation is based on the adaptation to Basque of the features used for English by Bethard (2013). Those features correspond to the `form` and the EusTimeML attributes for events: `class`, `tense1`, `tense2`, `aspect1`, `aspect2`, `polarity`, `modality` and `pos`.

For event capturing, we have used the best configuration achieved by bTime that reaches 76.59 F_1 points in event extent identification. For temporal relations between events and the document creation time (DCT), the system reaches 66.57 F_1 points in Temporal Awareness Score (UzZaman et al., 2012). These results are summarized in Table 3.

Table 3: Results of bTime's best configuration in the EusTimeBank test corpus

Element	**Feature**	**Precision**	**Recall**	F_1
EVENT	extension	86.22	68.9	76.59
	class			56.71
	tense1			65.51
	tense2			67.0
	aspect1			64.78
	aspect2			63.89
	polarity			71.48
	pos			69.15
	modality			73.78
TLINK	Event-DCT	67.76	65.42	66.57

4. KroniXa: timeline creation from temporal information

KroniXa takes the previously extracted temporal information for building timelines. We will illustrate the timeline creation process using the text presented below. This piece of news relates the Apple Inc.'s sales doubling during Christmas period in 2006. It is not our intention to describe thoroughly the contents of the text, but in order to highlight the relevant information to our research, events are presented in bold while time expressions are represented in italics. The numbers added as subindexes to events represent token position (identification). Next to the Basque word-form we present the corresponding English translation (in case the word-form is not a verb, the POS has been added).

Apple Inc.-ek bikoiztu egin ditu irabaziak

2007ko urtarrilaren 18

Apple-ren irabaziak ehuneko 78 **hazi**$_{17}$ ziren *azkeneko hiruhilekoan*, *zeina 2006ko abenduaren 30ean* **amaitu**$_{26}$ baitzen. **Igoera**$_{29}$ horren arrazoia *Gabonetako* eta *Urte Berriko* **salmentetako**$_{36}$ Apple-ren iPod musika irakurgailu digitalaren **eskari**$_{42}$ handia **izan**$_{44}$ zen.

Gabonetako **salmenta**$_{48}$ arrakastatsuen ondorioz, *urte ekonomikoaren hasieran* **bikoiztu**$_{55}$ egin ziren Apple-ren **salmentak**$_{59}$, *iaz tarte berean* **irabazitako**$_{64}$ 565 milioi dolarrekin **konparatuz**$_{68}$ gero. Konpainiak, 1.000 milioi dolar **irabazita**$_{76}$, errekorra **hautsi**$_{79}$ zuen. Konpainiaren *urteko* diru-sarrera orokorrak 7.100 milioi dolarrera **igo**$_{89}$ ziren *2006an*, *2005eko* 5.800 milioi dolarretatik. Konpainiaren diru-sarrerak **hazi**$_{100}$ izana Apple-k **ekoiztutako**$_{103}$ ordenagailu eramangarriak salduenen artean **egon**$_{108}$ izanari **zor**$_{110}$ zaio nolabait.

Wall Street-eko adituek ere ez zuten hain emaitza harrigarririk **iragarri**$_{123}$ Apple ordenagailu-konpainiarentzat. Apple-ren emaitzak merkatuko iragarpenak baino askoz **hobeak**$_{133}$ **izan**$_{134}$ ziren. Irabazi garbiak akzioko 1,14 dolarrekoak **izan**$_{142}$ ziren; Wall Street-ek, berriz, soilik 77 zentabokoak **izango**$_{153}$ zirela **iragarria**$_{155}$ **zuen**$_{156}$.

Urte ekonomikoko lehen hiruhilekoan, Apple Inc.-ek 21 milioi iPod irakurgailu baino gehiago eta 1,6 milioi ordenagailu inguru **saldu**$_{179}$ zituen; **horrek**$_{182}$ **esan**$_{183}$ **nahi**$_{184}$ du **salmentak**$_{186}$ %28 eta %50 **hazi**$_{190}$ direla beste behin *aurreko urtekoekin* **konparatuz**$_{196}$, hurrenez hurren. New Yorkeko Burtsan emaitzak **jakin**$_{205}$ bezain laster, Estatu Batuetako ordenagailu-konpainiaren akzioak ehuneko 5 **hazi**$_{215}$ ziren.

Aipatzekoa$_{218}$ **da**$_{219}$ Apple Inc.-en akzio-kapitala **hirukoiztu**$_{226}$ egin zela *2004an*, **bikoiztu**$_{231}$ *2005ean* eta ehuneko 16 **hazi**$_{236}$ *2006an*. **Hazkunde**$_{239}$ jarraitua iPod eta ordenagailuen **salmenta**$_{244}$ handien ondorio ere **bada**$_{248}$. *Joan den astean*, konpainiak iPhone **aurkeztu**$_{256}$ zuen.

KroniXa takes the information extracted by bTime and EusHeidelTime as the basis for timeline building. Once the events in text have been identified and linked to a time point, we have arranged them in order to create the timelines. For this, we have organized the events according to their position towards the DCT.

First, a preliminary event ordering has been done taking into account the relation type between the events and the DCT[2]. This first classification enables

[2]It is to be born in mind that at this stage KroniXa only receives relations between event

arranging those events in five groups:

- The event happens before the time the document is created
- The event happens immediately before the time the document is created
- The event happens the time the document is created
- The event happens immediately after the time the document is created
- The event happens after the time the document is created

The temporal relation type classification we have defined in EusTimeML is shown Figure 3.

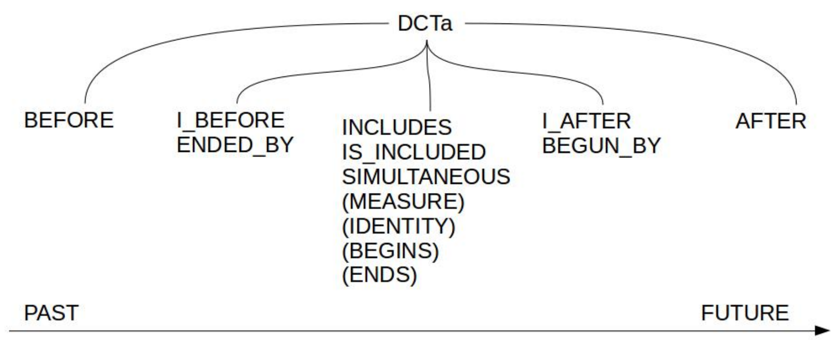


Figure 3: Temporal relation type clustering according to time

This first classification is a starting point. On one hand, only events expressed by tensed verbs are present in that first timeline. On the other hand, all events are anchored to a single time point.

4.1. Temporal relation tunning

Once all the possible event-time anchor pairs have been gathered, they have been ordered according to the normalized value of the time anchor. This process has consisted of the following steps:

expressed by tensed events and the DCT, due to the reduced amount of other temporal links in the gold standard corpus for system training purposes.

- As aforementioned, we aim at creating timelines based on temporal event anchoring. When we state that an event is included or is simultaneous to a time point, we assume the event has happened or will have in that certain moment in time. Therefore, we have had to transform the relations towards the DCT.

  The fact is that events placed before or after the DCT are not simultaneous to it and cannot be anchored to a timeline as they do not have a *simultaneous* anchor. Nonetheless, they can be anchored to auxiliary time anchors as follows: we have employed the undefined `XXXX-XX-XX` normalized value to create undefined time anchors for past and future events. As a result, we have obtained the simultaneity relations required for the timeline.

- After giving all the relations the `TIME ANCHOR --> simultaneity --> EVENT` format, we have ordered relations according to the normalized value of the time anchor. Although both past and future events (relative to the DCT) take the same `XXXX-XX-XX` normalized value, we have also taken into account the position of the event towards the DCT to place the new relations accordingly in the timeline.

- When all the events have been correctly placed in the timeline, a position identifier (position ID) has been assigned to each time anchor-event pair. All the pairs containing the same time anchor have been assigned the same identifier, for they are considered simultaneous. An example of the resulting timelines is presented in Figure 4.

5. Experimentation

KroniXa builds timelines automatically and we have evaluated its performance. For this, we have built a gold standard timeline set (Section 5.1) and we have defined the evaluation method and metrics (Section 5.2). We present the results obtained by KroniXa in Section 5.3. Nonetheless, as we have mentioned,

Position ID	Time expression ISO value	Doc ID-token ID-event
1	XXXX-XX-XX	7539-190-*hazi* (to grow/have grown)
1	XXXX-XX-XX	7539-215-*hazi* (to grow/grew)
1	XXXX-XX-XX	7539-89-*igo* (to increase/increased)
1	XXXX-XX-XX	7539-179-*saldu* (to sell/sold)
1	XXXX-XX-XX	7539-155-*iragarria* (to announce/announced)
1	XXXX-XX-XX	7539-134-*izan* (to be/were)
1	XXXX-XX-XX	7539-79-*hautsi* (to break/broke)
1	XXXX-XX-XX	7539-256-*aurkeztu* (to present/presented)
1	XXXX-XX-XX	7539-226-*hirukoiztu* (to triple/had tripled)
1	XXXX-XX-XX	7539-55-*bikoiztu* (to double/had doubled)
1	XXXX-XX-XX	7539-44-*izan* (to be/were)
1	XXXX-XX-XX	7539-142-*izan* (to be/were)
1	XXXX-XX-XX	7539-6-*bikoiztu* (to double/have doubled)
1	XXXX-XX-XX	7539-17-*hazi* (to increase/have increased)
2	2007-01-18	7539-108-*egon* (to be/to be)
2	2007-01-18	7539-111-*zaio* (to be-AUX)
2	2007-01-18	7539-123-*iragarri* (to annonuce/to announce)
2	2007-01-18	7539-184-*nahi* (to want/wants)
2	2007-01-18	7539-205-*jakin* (to know/to know)
2	2007-01-18	7539-219-*da* (to be/is)
2	2007-01-18	7539-248-*bada* (to be/is)
2	2007-01-18	7539-9-*irabaziak* (n. profits)
3	XXXX-XX-XX	7539-153-*izango* (to be/will be)
3	XXXX-XX-XX	7539-26-*amaitu* (to end/finished)

Figure 4: Example of automatically generated timeline from the relations extracted from bTime

temporal information obtained by bTime is rather scarce for timeline creation and, thus, we have reimplemented KroniXa adding information acquired from syntactic dependencies in a second experiment (Section 5.4).

5.1. Gold standard timelines

In order to be able to evaluate the performance of KroniXa, we have built a test dataset of timelines. More precisely, we have manually created the timelines of the 15 documents in the test corpus in EusTimeBank. For this task, two annotators were asked to create timelines from the temporal information annotated following EusTimeML in those texts. They were asked to place all the events marked in text in a timeline and to provide every event a time anchor according to the information in text.

Additionally, they were requested to take into account the following considerations:

- Only information in text was to be considered. Nothing could be inferred nor deduced neither world knowledge could be used.
- All the events annotated in text were to appear in their corresponding timelines.
- Only time expressions referring to time points (dates and times) could be anchors for only points could be pointed in chronology.
- All the simultaneous events[3] were to be anchored to the same time point and the same position in timeline.
- Co-referring events were to be considered individually, but were said to be simultaneous.

In Figure 5 a part of a timeline manually created from the sample text in Section 4 is displayed. As on the figure, the gold standard timelines contain the following information: i) the position identifiers are displayed in the first column; ii) normalized values of the time anchors are presented in the second column; and iii) the third column lists the events anchored to each time and their document and token identifiers.

5.2. Evaluation method and metrics

We have measured the performance of KroniXa following a methodology inspired in the SemEval 2015 Task 4 (Minard et al., 2015). This methodology uses the evaluation metrics in TempEval 3 (UzZaman et al., 2012) which take into account temporal awareness scores: "the performance of an annotation as identifying and categorizing temporal relations, which implies the correct recognition and classification of the temporal entities involved in the relations".

[3]For this experiment we have considered that partially overlapping events happen simultaneously.

Position ID	Time expression ISO value	Doc ID-token ID-event
3	2006	7539-w64-*irabazitako* (adj. earned)
3	2006	7539-w89-*igo* (to increase/increased)
3	2006	7539-w236-*hazi* (to grow/grew)
4	FY2006-Q1	7539-w179-*saldu* (to sell/sold)
4	FY2006-Q1	7539-w182-*horrek* (pron. that)
4	FY2006-Q1	7539-w186-*salmentak* (n. sales)
4	FY2006-Q1	7539-w190-*hazi* (to grow/have grown)
5	2006-Q4	7539-w17-*hazi* (to increase/have increased)
5	2006-Q4	7539-w100-*hazi* (to grow/have grown)
5	2006-Q4	7539-w115-*zor* (to owe/owe)
6	2006-12	7539-w48-*salmenta* (n. sale)
7	2006-12-30	7539-w26-*amaitu* (to end/finished)
8	2007-01	7539-w36-*salmentetako* (adj. sold)
8	2007-01	7539-w42-*eskari* (n. demand)
8	2007-01	7539-w44-*izan* (to be/were)
9	FY2007	7539-w55-*bikoiztu* (to double/had doubled)
9	FY2007	7539-w59-*salmentak* (n. sales)
9	FY2007	7539-w68-*konparatuz* (to compare/comparing)
9	FY2007	7539-w76-*irabazita* (adj. earned)
9	FY2007	7539-w79-*hautsi* (to break/broke)
9	FY2007	7539-w133-*hobeak* (adj. better)
9	FY2007	7539-w134-*izan* (to be/were)
9	FY2007	7539-w142-*izan* (to be/were)
9	FY2007	7539-w153-*izango* (to be/will be)
10	XXXX-XX-XX	7539-w205-*jakin* (to know/to know)
10	XXXX-XX-XX	7539-w215-*hazi* (to grow/grew)
11	2007-W2	7539-w256-*aurkeztu* (to present/presented)
12	2007-01-18	7539-w183-*esan* (to say/to say)

Figure 5: Partial timeline created manually

As a preliminary step, timelines are converted to a temporal relation graph in which simultaneity and precedence relations are made explicit (Figure 6). Then those graphs are compared to the gold standard graphs. The scores are based on the micro-average of the individual precision, recall and F_1 scores for each timeline. The Precision is calculated by checking the number of reduced system relations that can be verified from the reference annotation's temporal closure graph, out of the number of temporal relations in the reduced system relations. Similarly, the Recall is obtained by checking the number of reduced reference annotation relations that can be verified from the system output's temporal closure graph, out of the number of temporal relations in the reduced reference annotation.

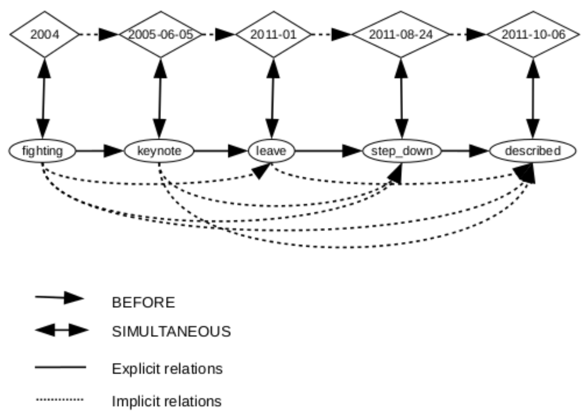


Figure 6: Example graph of a timeline extracted from

5.3. KroniXa's results

We have assessed KroniXa's performance comparing the automatically obtained timelines to manually created timelines. KroniXa has achieved the evaluation results shown in Table 4. In the table, we present the Temporal Awareness Scores of KroniXa when building timelines from temporal information automatically extracted with bTime (Auto) and from manually annotated temporal information (Gold). In what concerns the fully automated timeline creation, KroniXa reaches the 24.65 for F_1. When building timelines from manually annotated corpora, instead, F_1 score rises up to 34.38.

Table 4: Kronixa's results according to the Temporal Awareness Scores

	P	R	F_1
Auto	37.79	18.29	24.65
Gold	52.50	25.55	34.38

Taking into account that at this point KroniXa can only order events in three groups (i) prior to DCT, ii) simultaneous to DCT and iii) after DCT) and that only events expressed by tensed verbs are included in those timelines, the results are acceptable.

Regarding time anchor identification, a manual observation has shown that the fully automated system is able to identify 42 different anchors in corpus, while the KroniXa run on manually annotated corpora obtains 44.

Although the results are not high, they look promising. We have not found any comparable experiment, but the results obtained by the systems taking part in the SemEval 2015 Task 4 can be taken as a reference. The subtrack A of that task consisted in building multi-document timelines for certain seed entities from raw text. Although our effort is restricted to document level, the fact that the temporal information has to be extracted in order to build the timelines make both tasks similar. In what concerns that subtrack, the only participant system achieved 1.69 for F_1. We presume our system could perform similarly when timelines for determined entities will be created.

Coming back to table 4, in a first sight, one could think that improving the recall could notably improve the results. It needs to be kept in mind that bTime can only deal with certain temporal relations and that the temporal graph is not as rich as desired. However, we admit there is still room for improvement as results on the manually annotated data are still low.

5.4. Second experiment: Enhancing temporal information extraction through syntactic dependency parsing

We have used the existing syntactic dependency information to increase the amount of temporal information available to build timelines. Tensed verbs-DCT relations are not enough to represent the temporal relations inside the text and, hence, we have enhanced bTime by adding intra-sentential temporal relations based on syntactic information about dependencies.

Syntactic analysis of Basque is performed by the graph-based version of Mate parser, which adopts the second order maximum spanning tree dependency parsing algorithm (Otegi et al., 2016). The parser creates dependency relations (also known as binary relations) that connect pairs of words (a head and a dependent), and name the relationship between these parts.

We reckon that if there is a dependency relation between a word that has been marked as an event by bTime and a word that has been identified as a time expression by EusHeidelTime, there is a sentence-level temporal relation between both. We explicit those new relations and then, they can be processed by bTime in order to be classified. As a result, we obtain a denser temporal relation graph that will better represent temporal relations in text.

The temporal relations created from syntactic dependencies add new time anchors and events to the temporal graph: i) these relations capture time anchors present in text and, thus the number of anchoring point increases, and ii) any token marked as an event can be part of these relations, so events expressed by forms other than tensed events are added to the timelines.

In order to avoid inserting time expressions that could not be anchored to a chronology, we have only extracted dependency relations that involved time expressions annotated with `DATE` and `TIME` values, for they are the only that refer to time points. We also think that in this kind of relations there is normally an overlap (total or partial) between the event and the time point[4] they are related to. As a consequence, we have assigned an `IS_INCLUDED` relation type to all the relations obtained based on syntactic dependency information.

These new relations have been transformed into the `TIME ANCHOR --> simultaneity --> EVENT` format and added to the timelines created by KroniXa. Figure 7 displays the enhanced timeline corresponding to the sample text we have employed in this work.

As can be seen the events *7539-w256-aurkeztu*,*7539-w26-amaitu* and *7539-w89-igo* appear twice (in italics). This is due to the fact i) that they are tensed verbs and, thus, have been identified by bTime and ii) that they are related to a time expression and are part of a dependency relation and, hence, have been added to the timeline in this second experiment.

We reckon that an event cannot happen in two different moments. If an

[4]Since we deal with anchors in a timeline, we consider times points of variable granularity.

event appears to happen more than once, we assume we are dealing with two iterations of the same event, although they refer to individual actions. However, this is not the case in our corpus as each event has been annotated so as to express a unique iteration. In a well-formed timeline, the ones created from dependencies should be prioritized when an event is involved in two relations. For example, as can be seen in Figure 7, the event *igo* expressed by the token *w89* has been anchored in the past by bTime and in *2006* according to the dependency processing. Since the second is more descriptive, we should opt for it.

Figure 7 also shows some inconsistencies on anchor ordering. Ideally, to order anchors that partially share the normalized value, the one with the largest granularity should be placed first. E.g. the time anchor *2006* should be placed prior to *2006-12-30*, as it expresses a year while the second expresses a day. As can be seen in Figure 7, *2006* in presented in position 4 while *2006-12-30* appears in the 3rd position.

Figure 8 represents the events and time anchors in Figure 7 in a timeline.

Table 5: Results obtained by KroniXa when information from syntactic dependencies is added

	P	R	F_1
Auto	33.09	19.27	24.36
Gold	45.61	24.02	31.47

Table 5 shows the results achieved by the enhanced system. Surprisingly the results are lower compared to the ones KroniXa achieves employing only information extracted from bTime and EusHeidelTime. If compared to the results achieved by the first kroniXa configuration, precision drops notably— from 37.79 and 52.5 to 33.09 and 45.61 for fully automated and semi-automated respectively— whereas recall for the fully automated system increases (from 18.15 to 19.27).

A in-depth analysis of the resulting timelines has shown that, although the anchor number in the new timelines is higher, the impossibility to order them properly affects drastically the final results. In fact, the number of anchors has been increased by 2 in average for the fully automated system and by 2.2 for the system that takes manually annotated corpora as input. For example, the time anchors in Figure 7 widely outnumber the ones in Figure 4.

It is also to be taken into account that anchors and relations have been obtained automatically and, thus, even if the amount of anchors is larger, the event-anchor relations are not always correct. For example, *konparatuz* in position 2, should not be anchored to *2005* and should be anchored to the DCT (*2007-01-18*).

6. Discussion

Our experiments look promising, as we have extracted automatically many temporal information from text. More precisely, we are able to extract events, time expressions and some of the relations created among those, and we can also benefit from syntactic dependency information. In fact, dependency information has proved to be a very useful resource, as it has contributed to enrich the timeline adding a relevant amount of time anchors and events related to those. It should also be remarked that the use we made of dependency information helps getting profit of bTime's potential for it employs temporal information that could not be used by KroniXa. Nonetheless, we presume further analysis of the corpora and the timeline creation output will help exploiting those resources more efficiently.

At the moment, our timelines offer a general insight on what is going on in texts, but we admit deeper understanding is desirable. In order to produce more descriptive timelines, we have identified some areas for improvement. First, we assume enlarging our training corpus will improve the processing in bTime. This is a costly task, though, as manual annotation requires large amounts of time and annotator training. However, employing bTime for semi-automatic temporal information annotation may help ease the burden.

Secondly, we need to broaden the scope of bTime in what concerns relation identification. At the moment, the system is only able to deal with temporal relations between certain verbal events and the document creation time, but the relations inside the texts should also be treated, for they convey much temporal information. We have made a first approach creating temporal relations from dependencies, but this is clearly insufficient and new approaches that can cope with more types of relations should be taken into account. However, we have not made any attempts on event-event temporal ordering, which would add much useful information as would give evidence on the order of events that are now clustered in a same position.

In what concerns event information, while events denoted by verbs are identified and their information extracted conveniently, that is not the case for the rest of the events. As a matter of fact, the second major way of expressing events in Basque are nouns (Altuna et al., 2018a). However, these are not as easy to annotate as verbs, since the same form is less likely to always express an event. In what concerns semantic information extraction, Basque NomBank is still under construction and we cannot yet extract the semantic information of nominal events. The development of Basque NomBank will definitely contribute to enrich our timeline, as we will be able to link more events and entities taking part in those.

In this paper we have not addressed entity nor event co-reference. However, we consider solving co-reference a crucial step to good timelines. In fact, event co-reference resolution will avoid duplicities offering good quality summaries. More so, we have seen state-of-the-art systems deal with entity-based timelines. In our case, identifying all the co-referring entity expressions will lead to being able of building rich entity-based timelines.

7. Conclusion

In this paper we have described the first attempt of building a timeline for the events present in texts and based on the linguistic information given by EusHeidelTime and bTime, tools developed for the temporal information extraction in texts written in Basque. The basic temporal relations given by bTime have been enriched with new temporal relations we have extracted from dependencies in such a way that these new relations facilitate to set new time anchors.

Nonetheless, there is still much room for improvement. In future works, all the events in text should be eligible to be represented in a timeline. This requires some intermediate works: i) improving event detection, ii) improving time expression recognition and normalization, and iii) dealing with more types of temporal relations (including implicit relations). More precisely, the third work should consist of identifying as many explicit temporal relations as possible and trying to order events one relative to another, even if they cannot be anchored to any time point.

Another recurrent subject in temporal information processing is the anchoring of duration expressions, that is to say, identifying their beginning and end points. This would increment the amount of time anchors in a timeline. Minard et al. (2015) also show awareness for this issue, but it does not seem to have been addressed by the community yet.

Storylines, as defined in Minard et al. (2015), are an overview of what happened, to whom, when, and where. Being our aim to build a storyline building tool, we should extract the participants of the events taking part in the timeline in order to be able to create a chronological account of the action an entity has been involved.

Finally, we should try to build cross-document timelines in which the information in various texts is merged into a single chronology. For this effort event co-reference resolution is a major prerequisite, as the same events can be mentioned in different texts. Thus, a co-reference processing module should also be added to the processing pipeline.

Acknowledgements

This work has been supported by DETEAMI (Basque Government, ref.: 2014111003), by SINPATICO (MINECO, ref.: TIN2015-64395-R) and by PROSAMED (MINECO, ref.: TIN2016-77820-C3-1-R) projects.

References

Ahn, D., Adafre, S. F., & de Rijke, M. (2005). Extracting Temporal Information from Open Domain Text: A Comparative Exploration. *Journal of Digital Information Management*, *3*, 14–20.

Allan, J., Gupta, R., & Khandelwal, V. (2001). Temporal Summaries of New Topics. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* SIGIR'01 (pp. 10–18). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/383952.383954.

Alonso, O., Gertz, M., & Baeza-Yates, R. (2009). Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* CIKM '09 (pp. 97–106). New York, NY, USA: ACM. URL: http://doi.acm.org/10.1145/1645953.1645968. doi:10.1145/1645953.1645968.

Altuna, B., Aranzabe, M., & Díaz de Ilarraza, A. (2016). *Euskarazko denbora-egiturak etiketatzeko gidalerroak v2.0*. Technical Report Lengoaia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2016. https://addi.ehu.es/handle/10810/17305.

Altuna, B., Aranzabe, M. J., & Díaz de Ilarraza, A. (2017). EusHeidelTime: Time Expression Extraction and Normalisation for Basque. *Procesamiento del Lenguaje Natural*, *59*, 15–22. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5488.

Altuna, B., Aranzabe, M. J., & Díaz de Ilarraza, A. (2018a). Adapting TimeML to Basque: Event Annotation. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 565–577). Cham: Springer International Publishing.

Altuna, B., Aranzabe, M. J., & Díaz de Ilarraza, A. (2018b). An Event Factuality Annotation Proposal for Basque. In A. U. Frank, C. Ivanovic, F. Mambrini, M. Passarotti, & C. Sporleder (Eds.), *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)* (pp. 15–24). Vienna, Austria: Gerastree Proceedings volume 1. `https://www.oeaw.ac.at/ac/crh2/proceedings/`.

Bartalesi Lenzi, V., Moretti, G., & Sprugnoli, R. (2012). CAT: the CELCT Annotation Tool. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (pp. 333–338). Istanbul, Turkey: European Language Resources Association (ELRA).

Bauer, S., & Teufel, S. (2015). A Methodology for Evaluating Timeline Generation Algorithms based on Deep Semantic Units. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015* (pp. 834–839). Beijing, China volume 2: Short Papers. URL: `http://aclweb.org/anthology/P/P15/P15-2137.pdf`.

Bethard, S. (2013). ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 10–14). Association for Computational Linguistics. URL: `http://aclanthology.coli.uni-saarland.de/pdf/S/S13/S13-2002.pdf`.

Chieu, H. L., & Lee, Y. K. (2004). Query Based Event Extraction Along a Timeline. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* SIGIR'04 (pp. 425–432). New York, NY, USA: ACM. URL: `http://doi.acm.org/10.1145/1008992.1009065`. doi:`10.1145/1008992.1009065`.

Fokkens, A., Soroa, A., Beloki, Z., Ockeloen, N., Rigau, G., van Hage, W. R., & Vossen, P. (2014). NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation* (p. 9). Reykjavik, Iceland.

Jung, H., & Stent, A. (2013). ATT1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)* (pp. 20–24). volume 2.

Minard, A.-L., Speranza, M., Agirre, E., Aldabe, I., van Erp, M., Magnini, B., Rigau, G., & Urizar, R. (2015). SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 778–786). Denver, Colorado: Association for Computational Linguistics. URL: `http://www.aclweb.org/anthology/S15-2132` urlhttp://www.aclweb.org/anthology/S15-2132.

Otegi, A., Ezeiza, N., Goenaga, I., & Labaka, G. (2016). A Modular Chain of NLP Tools for Basque. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Proceedings of the 19th International Conference on Text, Speech and Dialogue — TSD 2016, Brno, Czech Republic* (pp. 93–100). Springer International Publishing volume 9924 of *Lecture Notes in Artificial Intelligence*.

Pustejovsky, J., Castaño, J. M., Ingria, R., Saurí, R., Gaizauskas, R. J., Setzer, A., Katz, G., & Radev, D. R. (2003). TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, *3*, 28–34.

Salaberri Izko, H. (2017). *Rol semantikoen etiketatzeak testuetako espazio-denbora informazioaren prozesamenduan daukan eraginaz*. Ph.D. thesis Universidad del País Vasco/Euskal Herriko Unibertsitatea Donostia, Euskal Herria.

Saurí, R., Batiukova, O., & Pustejovsky, J. (2009). *Annotating Events in Spanish. TimeML Annotation Guidelines*. Technical Report Technical Report Version TempEval-2010., Barcelona Media-Innovation Center.

Saurí, R., Littman, J., Gaizauskas, R., Setzer, A., & Pustejovsky, J. (2006). TimeML Annotation Guidelines, Version 1.2.1.

Setty, V., Bedathur, S., Berberich, K., & Weikum, G. (2010). InZeit: Efficiently Identifying Insightful Time Points. *Proceedings of the VLDB Endowment*, *3*, 1605–1608. URL: http://dx.doi.org/10.14778/1920841.1921050. doi:10.14778/1920841.1921050.

Strötgen, J., & Gertz, M. (2013). Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, *47*, 269–298. doi:10.1007/s10579-012-9179-y.

TimeML Working Group (2010). *TimeML Annotation Guidelines version 1.3. Manuscript*. Technical Report Brandeis University.

Tonelli, S., Sprugnoli, R., Speranza, M., & Minard, A.-L. (2014). *NewsReader Guidelines for Annotation at Document Level. Version FINAL*. Technical Report Fondazione Bruno Kessler. http://www.newsreader-project.eu/files/2014/12/NWR-2014-2-2.pdf.

Tran, G. B., Herder, E., & Markert, K. (2015). Joint Graphical Models for Date Selection in Timeline Summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015* (pp. 1598–1607). Beijing, China volume 1: Long Papers. URL: http://aclweb.org/anthology/P/P15/P15-1154.pdf.

UzZaman, N., Llorens, H., Allen, J. F., Derczynski, L., Verhagen, M., & Pustejovsky, J. (2012). TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. *CoRR*, *abs/1206.5333*. http://arxiv.org/abs/1206.5333.

Yan, R., Wan, X., Otterbacher, J., Kong, L., Li, X., & Zhang, Y. (2011). Evolutionary timeline summarization: A balanced optimization framework via iterative substitution. In *In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'11* (pp. 745–754). ACM.

Position ID	Time expression ISO value	Doc ID-token ID-event
1	XXXX-XX-XX	7539-190-*hazi* (to grow up/have grown)
1	XXXX-XX-XX	7539-215-*hazi* (to grow up/grew)
1	XXXX-XX-XX	*7539-89-igo* (to increase/increased)
1	XXXX-XX-XX	7539-179-*saldu* (to sell/sold)
1	XXXX-XX-XX	7539-155-*iragarria* (to announce/announced)
1	XXXX-XX-XX	7539-134-*izan* (to be/were)
1	XXXX-XX-XX	7539-79-*hautsi* (to break/broke)
1	XXXX-XX-XX	*7539-256-aurkeztu* (to present/presented)
1	XXXX-XX-XX	7539-226-*hirukoiztu* (to triple/had tripled)
1	XXXX-XX-XX	7539-55-*bikoiztu* (to double/had doubled)
1	XXXX-XX-XX	7539-44-*izan* (to be/were)
1	XXXX-XX-XX	7539-142-*izan* (to be/were)
1	XXXX-XX-XX	7539-6-*bikoiztu* (to double/have doubled)
1	XXXX-XX-XX	7539-17-*hazi* (to incease/have increased)
2	2005	7539-231-*bikoiztu* (to double/had doubled)
2	2005	7539-68-*konparatuz* (to compare/comparing)
3	2006-12-30	*7539-26-amaitu* (to finish/finished)
4	2006	*7539-89-igo* (to increase/increased)
5	2006-W51	*7539-256-aurkeztu* (to present/presented)
6	2007-01-18	7539-108-*egon* (to be/to be)
6	2007-01-18	7539-111-*zaio* (to be-AUX)
6	2007-01-18	7539-123-*iragarri* (to announce/to announce)
6	2007-01-18	7539-184-*nahi* (to want/wants)
6	2007-01-18	7539-205-*jakin* (to know/to know)
6	2007-01-18	7539-219-*da* (to be/is)
6	2007-01-18	7539-248-*bada* (to be/is)
6	2007-01-18	7539-9-*irabaziak* (n. profits)
7	XXXX-XX-XX	7539-196-*konparatuz* (to compare/comparing)
8	XXXX-XX-XX	7539-153-*izango* (to be/will be)
8	XXXX-XX-XX	*7539-26-amaitu* (to end/finished)

Figure 7: Example of automatically generated timeline

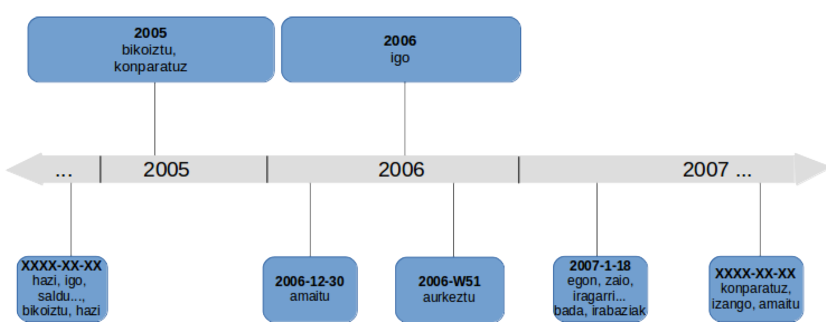


Figure 8: Visual timeline representation of the events and time expressions in Figure 7

CONCLUSIONS AND FUTURE WORK

6 Conclusions and Future Work

As mentioned in the introductory chapter, understanding temporal information contributes to understanding textual information. Hence, we have developed the resources for temporal information processing in Basque. Being able to deal with that information has made possible creating advanced temporal information processing tools such as chronlogy creation systems.

In this chapter we present the contributions of our work, the conclusions and the future work we envision.

6.1 Contributions

There are many resources that have to be developed for proper temporal information processing. In fact, in the case of Basque, this is the first in-depth work that focuses at temporal information processing. In this section we give account of our contributions to temporal information processing in Basque.

6.1.1 Temporal Information Analysis

As a first step of our work we have analysed which elements take part in temporal information.

- First, we have analysed events and time expressions. That is to say, what and when happens. In general, we have followed the main trends in the area, although a little discussion has been required sometimes.

For example, in what concerns event definition we have opted for a rather wide one and we have considered events generic situations, for we consider they can be placed in time and for we aimed at a general-use temporal information analysis.

- We have also analysed the relations created between events and time expressions, their directionality and class. More precisely, we have analysed three kinds of relations: temporal relations, subordination relations and aspectual relations. The selection of these three relation types is justified by the fact that we have chosen TimeML as the basis of our mark-up language. The relations proposed in that mark-up scheme seemed the most comprehensive for the task, although we reckon causal relations would also be of interest. Additionally, we have also analysed the signals that make the temporal relations in text explicit.

- Finally, we also consider factuality information of ultimate importance for temporal information processing. On one hand both share similar linguistic features. On the other hand, knowing whether an event has effectively happened is of utmost importance when understanding temporal information.

6.1.2 EusTimeML: Mark-up Language for Temporal Information in Basque

We have created a TimeML inspired mark-up language for temporal information in Basque. We will now list the main contributions when creating the mark-up language:

- We have defined the EusTimeML mark-up language for temporal information processing in Basque. For that we have defined the tags, the attributes and the attribute values that best describe the linguistic information concerning time in Basque. EusTimeML shares tags with other TimeML schemes in order to ease the task of corpora comparison or parallel corpora building.

- We have opted for preserving token-level annotation and the tags have been assigned according to the lemmas. Opposite to the mots of the languages that have a TimeML-style scheme, Basque is a morphologically rich language and that has led us to take that annotation decision.

- We have mostly transferred the attributes from other TimeML schemes for they have proved worthy for temporal information normalisation. Nevertheless, some attributes (e.g. vform and mood) have been discarded since they did not express relevant information for Basque and some others (aspect and tense) have seen their values changed in order to accommodate Basque linguistic features. In this step we have always taken into account the information in the Basque linguistic processing chain, as the temporal information processing tools take parsed texts as input.

- We have added factuality attributes to EusTimeML as we have decided to integrate factuality in temporal information processing.

- We have defined the EusTimeML annotation guidelines and we have evaluated them.

We have measured the quality of the annotation guidelines through inter-annotator agreement analysis. we have achieved an agreement over 80 % in event recognition and some for some of the attributes we have gone even further. We have achieved 87 % on grammatical category agreement, 98 % in polarity and 77 % in factuality. Tiem expressions have been unanimously annotated in 92 % of the cases and have been assigned the same type and normalised value in circa 70 % of the cases. In what concerns relations, we have reached 92 % of agreement in relation classification. Considering the annotation guidelines have been created properly, we have started corpus annotation.

6.1.3 The EusTimeBank corpus

EusTimeBank has been a crucial resource for temporal phenomena analysis and tool development. EusTimeBank contains i) news texts for they are closely related to the time they are created and ii) history texts as they provide narrations of events in time. The corpus is formed of 164 documents from which 60 have been employed to create the EusTimeBank gold standard corpus. The gold standard subcorpus has been used to train bTime (Salaberri Izko, 2017) and to evaluate EusHeidelTime and bTime. Furthermore, the corpus has been used as the basis for gold standard timeline creation for the evaluation of KroniXa.

All the documents have been annotated following EusTimeML and are available in NAF format. Additionally, EusTimeBank has also been used for negation information analysis and thus, a selection of the documents also contain negation information. The EusTimeBank gold standard corpus is freely available[1].

6.1.4 The EusHeidelTime Tool

We have developed EusHeidelTime which identifies, classifies and normalises Basque time expressions. More precisely we have defined 313 rules, 58 patterns and 29 normalisation files and we have adapted HeidelTime (Strötgen and Gertz, 2013) to work with Basque processing chains.

While some rules have been transferred directly from other languages, many other have been built from scratch as Basque has some particularities neighbouring languages do not share. Nonetheless, the rule-based approach was the most appropriate in our case due to the size of the available corpus and the little variation of time expressions in Basque.

It is relevant to mention that EusHeidelTime achieves 80 % on strict time expression recognition, which is a state-of-the-art performance. In the case of partial recognition of time expressions, performance rises up to 90 %, which proves that EusHeidelTime is already a useful tool for textual annotation. Apart from that, EusHeidelTime can deal with NAF documents and thus, can be easily integrated in more complex systems.

6.1.5 The KroniXa System

As mentioned in this work temporal information processing contributes to more advanced natural language processing systems. In our case, we have employed temporal information in KroniXa, the timeline creation system for Basque. KroniXa takes as input documents annotated following EusTimeML and creates timelines based on the relations between events and time expressions in text. Additionally, we have enriched the temporal graphs obtained from stricty speaking temporal information processing with dependency information, as we consider that if there is a syntactic dependency relation between a token annotated as an event and a token annotated as a time expression, there is a temporal relation between both.

[1] `http://ixa2.si.ehu.es/eusheideltime`

In order to assess KroniXa's performance, we have manually created a timeline dataset that has been confronted with the automatically obtained ones. KroniXa has scored 24.36 for Temporal Awareness Score F-measure in its fully automated version. Results are not high, but they are very promising since we correctly extracted notable amounts of information, but we admit there is still information which we have not been able to catch.

6.2 Conclusions

In the introductory chapter of this dissertation we have listed the main goals of our research and in the following chapters we have presented the fullfilment of those. From our research we have extracted these conclusions:

- As we have chosen nonrestrictive definitions of temporal constructions, we have obtained definitions that can be used in many contexts.
- EusTimeML permits the complete and accurate annotation of temporal information in Basque.
- Since EusTimeML follows the TimeML style, annotations in different languages are easily comparable and parallel corpora can be easily built.
- We have developed EusHeidelTime for time expression extraction and normalisation. The good results obtained by the tool show that the rule-based approach was the right one.
- As EusHeidelTime reads and writes data in NAF format, EusHeidelTime can be easily integrated in the Basque processing chain.
- Robust temporal information analysis and processing are the pillars for advanced temporal information processing tools. When creating KroniXa, we have seen that the more temporal information is available, the more representative timelines are obtained.

We have made the first steps in Basque temporal information processing. Nonetheless, there are still aspects that can be analysed more in depth and some other still need to be addressed.

6.3 Future Work

We foresee some works in the future to go further in temporal information processing in Basque.

- We will continue developing EusTimeML and we may add variations for different types of texts.
- We will enlarge the EusTimeBank corpus to open new lines of research. For example, a larger corpus would enable its use for machine learning methods. We may also add text of other domains to get a more representative portion of language.
- We will continue analysing events and will try to compare our definition with those of PropBank (Palmer *et al.*, 2005) and NomBank (Meyers *et al.*, 2004). We aim at merging our work with predicate analysis.
- We will work on event co-reference to better understand what happens in texts.
- We will continue studying complimentary information to better represent temporal information. For example, analysing causal relations will improve event ordering.
- We will use temporal information in advanced tools such as automatic question-answer generation or natural language generation.

Bibliography

Agerri R., Agirre E., Aldabe I., Altuna B., Beloki Z., Laparra E., de Lacalle M.L., Rigau G., Soroa A., and Urizar R. NewsReader project. *Procesamiento del Lenguaje Natural*, 53:155–158, 2014.

Ahn D., Adafre S.F., and de Rijke M. Extracting Temporal Information from Open Domain Text: A Comparative Exploration. *Journal of Digital Information Management*, 3(1):14–20, 2005.

Allan J., Gupta R., and Khandelwal V. Temporal Summaries of New Topics. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, 10–18, New York, NY, USA, 2001.

Allen J.F. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

Alonso O., Gertz M., and Baeza-Yates R. Clustering and exploring search results using timeline constructions. *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, 97–106, New York, NY, USA, 2009.

Altuna B. Análisis de la información temporal en euskera. In Martínez Barco P., Navarro Colorado B., Vázquez Pérez S., and Romá Ferri M.T., editors, *Actas del XXXI Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Universidad de Alicante, Alicante, Spain, 2015a.

Altuna B. Euskarazko denbora-egituren tratamendu automatikorako azterketa. In Alegria I., Latatu A., and Omaetxebarria M.J., editors, *I. Ikergazte, Nazioarteko ikerketa euskaraz*, 46–53, Bilbao, Spain, 2015b.

Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. *Linguamática*, 6(2): 13–24, 2014a. http://linguamatica.com/index.php/linguamatica/article/view/v6n2-1.

Altuna B., Aranzabe M., and Díaz de Ilarraza A. Euskarazko denbora-egiturak etiketatzeko gidalerroak. Barne-txostena, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2014, 2014b. http://ixa.si.ehu.es/Ixa/Argitalpenak/Barne_txostenak/1414871293/publikoak/Denbora-egiturak%20etikeatzeko%20gidalerroak.

Altuna B., Aranzabe M., and Díaz de Ilarraza A. Euskarazko denbora-egiturak etiketatzeko gidalerroak v2.0. Barne-txostena, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2016, 2016a. https://addi.ehu.es/handle/10810/17305.

Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. Euskarazko denbora-informazioaren tratamendu automatikoa TimeMLren eta HeidelTimeren bidez. *Ekaia*, 30:153–165, 2016b.

Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. EusHeidelTime: Time Expression Extraction and Normalisation for Basque. *Procesamiento del Lenguaje Natural*, 59(0):15–22, 2017a. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5488.

Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. Euskarazko ezeztapenaren tratamendu automatikorako azterketa. In Alegria I., Latatu A., Ormaetxebarria M.J., and Salaberri P., editors, *II. IkerGazte, Nazioarteko Ikerketa Euskaraz: Giza Zientziak eta Artea*, 127–134, Bilbo, 2017b. http://ixa.si.ehu.eus/sites/default/files/dokumentuak/8863/IKERGAZTE.2017.GIZAZIENTZIAKetaARTEA.pdf.

Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. Adapting TimeML to Basque: Event Annotation. In Gelbukh A., editor, *Computational Linguistics and Intelligent Text Processing*, 565–577, Cham, Switzerland, 2018a.

Altuna B., Aranzabe M.J., and Díaz de Ilarraza A. An Event Factuality Annotation Proposal for Basque. In Frank A.U., Ivanovic C., Mambrini F., Passarotti M., and Sporleder C., editors, *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2)*, 1 lib., 15–24, Vienna, Austria, 2018b. `https://www.oeaw.ac.at/ac/crh2/proceedings/`.

Altuna B., Minard A.L., and Speranza M. The Scope and Focus of Negation: A Complete Annotation Framework for Italian. *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, 34–42. 2017c. `http://aclweb.org/anthology/W17-1806`.

Altuna B., Soraluze A., Aranzabe M.J., Arregi O., and Díaz de Ilarraza A. KroniXa: Timeline Creation from Basque Texts. *Journal of Information Processing and Management*, Under revision.

Altuna P., Salaburu P., Goenaga P., Lasarte M.P., Akesolo L., Azkarate M., Charriton P., Eguskitza A., Haritschelhar J., King A., Larrarte J.M., Mujika J.A., Oyharçabal B., and Rotaetxe K. *Euskal Gramatika Lehen Urratsak (EGLU) II.* Euskaltzaindiko Gramatika Batzordea, Euskaltzaindia, Bilbao, 1987.

Altuna P., Salaburu P., Goenaga P., Lasarte M.P., Akesolo L., Azkarate M., Charriton P., Eguskitza A., Haritschelhar J., King A., Larrarte J.M., Mujika J.A., Oyharçabal B., and Rotaetxe K. *Euskal Gramatika Lehen urratsak (EGLU) I.* Euskaltzaindiko Gramatika batzordea, Euskaltzaindia, Bilbo, 1985.

Asahara M., Yasuda S., Konishi H., Imada M., and Maekawa K. BCCWJ-Timebank: Temporal and Event Information Annotation on Japanese Text. *27th Pacific Asia Conference on Language, Information, and Computation*, 206–214, Wenshan, Taipei, Taiwan, 2013.

Bach E. The algebra of events. *Linguistics and philosophy*, 9(1):5–16, 1986.

Bauer S., Clark S., and Graepel T. *Learning to Identify Historical Figures for Timeline Creation from Wikipedia Articles*, 234–243. Springer International Publishing, Cham, Switzerland, 2015.

Bayer J. The Role of Event Expression in Grammar. *Studies in Language*, 1 (10):1–52, 1986.

Bethard S. ClearTK-TimeML: A minimalist approach to TempEval 2013. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 10–14. 2013.

Bethard S., Derczynski L., Savova G., Pustejovsky J., and Verhagen M. SemEval-2015 Task 6: Clinical TempEval. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 806–814, Denver, Colorado, 2015.

Bethard S. and Martin J.H. Identification of Event Mentions and Their Semantic Class. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, 146–154, Stroudsburg, PA, USA, 2006. `http://dl.acm.org/citation.cfm?id=1610075.1610098`.

Bijankhan M., Sheykhzadegan J., Bahrani M., and Ghayoomi M. Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation*, 45(2):143–164, 2011. `http://dx.doi.org/10.1007/s10579-010-9132-x`.

Bindi R., Monachini M., and Orsolini P. Il Corpus di Riferimento della lingua italiana contemporanea, Documentazione. Barne-txostena, ILC-TLN-1, 1989.

Bird S., Day D., Garofolo J., Henderson J., Laprun C., and Liberman M. ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, 2000. ACL Anthology Identifier: L00-1138.

Bittar A. Annotation of Events and Temporal Expressions in French Texts. *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, 48–51, Stroudsburg, PA, USA, 2009. `http://dl.acm.org/citation.cfm?id=1698381.1698388`.

Bittar A. *Building a TimeBank for French: a Reference Corpus Annotated According to the ISO-TimeML Standard.* Doktoretza-tesia, Université Paris Diderot, Paris, 2010. `http://www.linguist.jussieu.fr/~abittar/docs/Bittar-PhD.pdf`.

Bittar A., Hagège C., Moriceau V., Tannier X., and Teissèdre C. Temporal Annotation: A Proposal for Guidelines and an Experiment with Inter-annotator Agreement. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., and Piperidis S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 3741–3745, Istanbul, Turkey, 2012. ACL Anthology Identifier: L12-1142.

Boser B.E., Guyon I.M., and Vapnik V.N. A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, 144–152, New York, NY, USA, 1992. http://doi.acm.org/10.1145/130385.130401.

Broda B., Marcińczuk M., Maziarz M., Radziszewski A., and Wardyński A. KPWr: Towards a Free Corpus of Polish. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., and Piperidis S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 3218–3222, Istanbul, Turkey, 2012. ACL Anthology Identifier: L12-1574.

Canh T.V., Markert K., and Nejdl W. RussianFlu-DE: A German Corpus for a Historical Epidemic with Temporal Annotation. *Research and Advanced Technology for Digital Libraries - 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings*, 61–73, 2017.

Carlson G.N. *Reference to Kinds in English.* Doktoretza-tesia, University of Massachusetts Amherst, Amherst, Massachusetts, 1977. https://scholarworks.umass.edu/dissertations/AAI7726414.

Caselli T. It-TimeML: TimeML Annotation Scheme for Italian. Barnetxostena, Istituto di Linguistica Computazionale, 2010. Version 1.3. 1. Technical Report.

Caselli T., Lenzi V.B., Sprugnoli R., Pianta E., and Prodanof I. Annotating Events, Temporal Expressions and Relations in Italian: The It-TimeML Experience for the Ita-TimeBank. *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, 143–151, Stroudsburg, PA, USA, 2011.

Caselli T. and Vossen P. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. *Proceedings of the Events and Stories in the News Workshop*, 77–86, Vancouver, Canada, 2017.

Cassidy T., McDowell B., Chambers N., and Bethard S. An Annotation Framework for Dense Event Ordering. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 501–506, Baltimore, Maryland, 2014.

Chang A.X. and Manning C. SUTime: A library for recognizing and normalizing time expressions. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., and Piperidis S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 3735–3740, Istanbul, Turkey, 2012. ACL Anthology Identifier: L12-1122.

Cheng Y., Asahara M., and Matsumoto Y. Constructing a Temporal Relation Tagged Corpus of Chinese Based on Dependency Structure Analysis. *Proceedings of the 14th International Symposium on Temporal Representation and Reasoning*, TIME '07, 59–69, Washington, DC, USA, 2007. `http://dx.doi.org/10.1109/TIME.2007.47`.

Chieu H.L. and Lee Y.K. Query Based Event Extraction Along a Timeline. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, 425–432, New York, NY, USA, 2004.

Chinchor N. and Sundheim B. Message Understanding Conference (MUC) 6 LDC2003T13. Barne-txostena, Linguistic Data Consortium, Philadelphia, USA, 2003. `https://catalog.ldc.upenn.edu/LDC2003T13`.

Cieri C. and Liberman M. TIDES Language Resources: A Resource Map for Translingual Information Access. *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*, 1334–1339. 2002.

Costa F. and Branco A. TimeBankPT: A TimeML Annotated Corpus of Portuguese. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., and Piperidis S., editors, *Proceedings of* *the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 3727–3734, Istanbul, Turkey, 2012. ACL Anthology Identifier: L12-1096.

Croft W. Force Dynamics and Directed Change in Event Lexicalization and Argument Realization. In de Almeida G.R. and Manouilidou C., editors, *Cognitive Science Perspectives on Verb Representation and Processing*, 103–129. Springer International Publishing, Cham, Switzerland, 2015. `http://dx.doi.org/10.1007/978-3-319-10112-5_5`.

Cybulska A. and Vossen P. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., and Piperidis S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 4545–4552, Reykjavik, Iceland, 2014. ACL Anthology Identifier: L14-1646.

Day D., Aberdeen J., Hirschman L., Kozierok R., Robinson P., and Vilain M. Mixed-initiative Development of Language Processing Systems. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, 348–355, Stroudsburg, PA, USA, 1997.

Day D.S., McHenry C., Kozierok R., and Riek L. Callisto: A Configurable Annotation Workbench. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, 2073–2076, Lisboa, Portugal, 2004.

Derczynski L. and Bontcheva K. Pheme: Veracity in Digital Social Networks. *Proceedings of the 10th Joint ACL – ISO Workshop on Interoperable Semantic Annotation (ISA)*, Lisboa, Portugal, 2014.

Derczynski L., Llorens H., and UzZaman N. TimeML-Strict: Clarifying Temporal Annotation. *Computing Research Repository (CoRR)*, abs/1304.7289, 2013. `http://arxiv.org/abs/1304.7289`.

Dice L.R. Measures of the Amount of Ecologic Association between Species. *Ecology*, 26:297–302, 1945.

Dowden B. Time Supplement. In Fieser J. and Dowden B., editors, *Internet Encyclopedia of Philosophy*. 2009.

Etcheverry M. and Wonsever D. Time Expressions Recognition with Word Vectors and Neural Networks. In Schewe S., Schneider T., and Wijsen J., editors, *24th International Symposium on Temporal Representation and Reasoning (TIME 2017)*, 90 lib. of *Leibniz International Proceedings in Informatics (LIPIcs)*, 12:1–12:20, Dagstuhl, Germany, 2017.

Ferro L., Gerber L., Mani I., Sundheim B., and Wilson G. TIDES 2003 Standard for the Annotation of Temporal Expressions. Barne-txostena, MITRE, McLean, USA, 2003. http://www.mitre.org/sites/default/files/pdf/ferro_tides.pdf.

Ferro L., Gerber L., Mani I., Sundheim B., and Wilson G. ACE Time Normalization (TERN) 2004 English Evaluation Data v1.0. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2010T18, 2010. Accessed: 2013-07-03.

Ferro L., Kozierok R., Gerber L., Sundheim B., Mani I., and Wilson G. Annotating temporal information: From theory to practice. *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, 226–230, San Francisco, CA, USA, 2002.

Fiscus J.G., Doddington G.R., Garofolo J.S., and Martin A.F. NIST's 1998 topic detection and tracking evaluation (TDT2). *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999*, Budapest, Hungary, 1999.

Fokkens A., Soroa A., Beloki Z., Ockeloen N., Rigau G., van Hage W.R., and Vossen P. NAF and GAF: Linking linguistic annotations. *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 9, Reykjavik, Iceland, 2014.

Forăscu C. and Tufiş D. Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., and Piperidis S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 3762–3766, Istanbul, Turkey, 2012. ACL Anthology Identifier: L12-1451.

Galton A. A Critical Examination of Allen's Theory of Action and Time. *Artificial Intelligence*, 42(2-3):159–188, 1990.

Girardi C., Speranza M., Sprugnoli R., and Tonelli S. CROMER: a Tool for Cross-Document Event and Entity Coreference. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., and Piperidis S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3204–3208, Reykjavik, Iceland, 2014. ACL Anthology Identifier: L14-1568.

Goranko V. and Galton A. Temporal logic. In Zalta E.N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2015 edition, 2015.

Grishman R. and Sundheim B. Message Understanding Conference-6: A Brief History. *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, 466–471, Center for Sprogteknologi, Copenhagen, Denmark, 1996.

Grover C., Tobin R., Alex B., and Byrne K. Edinburgh-LTG: TempEval-2 System Description. *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, 333–336, Stroudsburg, PA, USA, 2010. `http://dl.acm.org/citation.cfm?id=1859664.1859738`.

Guerrero Nieto M., Saurí R., and Poveda M.Á.B. ModeS TimeBank: A Modern Spanish TimeBank Corpus. *Procesamiento del lenguaje natural*, 47:259–267, 2011.

Gupta N., Joshi A., and Bhattacharyya P. A temporal expression recognition system for medical documents by taking help of news domain corpora. In Sharma D.M., Sangal R., and Sherly E., editors, *Proceedings of the ICON 2015: 12th International Conference on Natural Language Processing*, Hyderabad, India, 2015.

Hacioglu K., Chen Y., and Douglas B. Automatic Time Expression Labeling for English and Chinese Text. In Gelbukh A., editor, *Computational Linguistics and Intelligent Text Processing*, 548–559, Berlin, Heidelberg, 2005.

Hualde J.I. and de Urbina J.O. *A grammar of Basque*, 26 lib. Walter de Gruyter, Boston, USA, 2003.

Im S., You H., Jang H., Nam S., and Shin H. KTimeML: Specification of Temporal and Event Expressions in Korean Text. *Proceedings of the 7th* *workshop on Asian Language Resources in conjunction with ACL-IJCNLP 2009*, 115–122, Suntec City, Singapore, 2009. Association for Computational Linguistics.

ISO Commitee. Information interchange: representation of dates and times. International Standard ISO 8601:2004, International Organization for Standardization, Switzerland, 1997. `https://www.iso.org/standard/40874.html`.

ISO-TimeML working group. Language resource management – Semantic Annotation Framework (SemAF) – Part 1: Time and events. International Standard ISO/CD 24617-1(E), International Organization for Standardization, Switzerland, 2008. `http://lirics.loria.fr/doc_pub/SemAFCD24617-1Rev12.pdf`.

Jang S.B., Baldwin J., and Mani I. Automatic TIMEX2 Tagging of Korean News. *ACM Transactions on Asian Language Information Processing (TALIP) - Special Issue on Temporal Information Processing*, 3(1):51–65, 2004. `http://doi.acm.org/10.1145/1017068.1017072`.

Jędrzejko E. The problematics of describing periphrastic predication between word and image. *Studies in Polish Linguistics*, 6(1):27–44, 2011.

Jeong Y.S. and Choi H.J. Language independent feature extractor. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 4170–4171, Texas, USA, 2015.

Jeong Y.S., Kim Z.M., Do H.W., Lim C.G., and Choi H.J. Temporal Information Extraction from Korean Texts. 279–288, 2015.

Jeong Y.S., Lim C.G., and Choi H.J. ExoTime: Temporal Information Extraction from Korean Texts Using Knowledge Base. *Journal of the Korea Society of Computer and Information*, 22:35–48, 2017.

Kawai H., Jatowt A., Tanaka K., Kunieda K., and Yamada K. Chronoseeker: Search engine for future and past events. *Proceedings of the 4th International Conference on Uniqiutous Information Management and Communication*, ICUIMC '10, 25:1–25:10, New York, NY, USA, 2010. `http://doi.acm.org/10.1145/2108616.2108647`.

Kocoń J. and Marcińczuk M. Supervised approach to recognise Polish temporal expressions and rule-based interpretation of timexes. *Natural Language Engineering*, 23(3):385–418, 2017.

Kocoń J., Marcińczuk M., Oleksy M., Bernaś T., and Wolski M. Temporal Expressions in Polish Corpus KPWr. *Cognitive Studies/Études Cognitives*, 15:293–317, 2015.

Lafferty J., McCallum A., and Pereira F.C. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 282–289, San Francisco, CA, USA, 2001. `http://dl.acm.org/citation.cfm?id=645530.655813`.

Lenzi V.B., Moretti G., and Sprugnoli R. CAT: the CELCT Annotation Tool. In Calzolari N., Choukri K., Declerck T., Doğan M.U., Maegaard B., Mariani J., Odijk J., and Piperidis S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, 333–338, Istanbul, Turkey, 2012. ACL Anthology Identifier: L12-1072.

Li L., Zhang J., He Y., Zhang Y., and Huilin W. Chinese Temporal Relation Resolution Based on Chinese-English Parallel Corpus. *New Technology of Library and Information Service*, (31):42–48, 2014.

Lin C.Y. Rouge: A package for automatic evaluation of summaries. In Marie-Francine Moens S.S., editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81, Barcelona, Spain, 2004.

Llorens H., Saquete E., and Navarro B. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, 284–291, Stroudsburg, PA, USA, 2010. `http://dl.acm.org/citation.cfm?id=1859664.1859727`.

Mani I. and Wilson G. Robust Temporal Processing of News. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 69–76, Stroudsburg, PA, USA, 2000.

Mansouri B., Zahedi M., Campos R., Farhoodi M., and Rahgozar M. ParsTime: Rule-Based Extraction and Normalization of Persian Temporal

Expressions. *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018*, 715–721, Grenoble, France, 2018.

March O. and Baldwin T. Automatic Event Reference Identification. *Australasian Language Technology Association Workshop 2008*, 6 lib., 79–87, 2008.

Marcińczuk M., Kocoń J., and Janicki M. *Liner2 – A Customizable Framework for Proper Names Recognition for Polish*, 231–253. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

Marcińczuk M., Oleksy M., Bernaś T., Kocoń J., and Wolski M. Towards an Event Annotated corpus of Polish. *Cognitive Studies/Études Cognitives*, 15:253–267, 2015.

Marcińczuk M., Oleksy M., and Kocoń J. Inforex — a Collaborative System for Text Corpora Annotation and Analysis. *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 473–482, Varna, Bulgaria, 2017.

Martí M.A., Martín-Valdivia M.T., Taulé M., Jiménez-Zafra S.M., Nofre M., and Marsó L. La negación en español: análisis y tipología de patrones de negación. *Procesamiento del Lenguaje Natural*, 57:41–48, 2016.

Mazur P. and Dale R. The DANTE Temporal Expression Tagger. In Vetulani Z., editor, *Proceedings of the 3rd Language And Technology Conference (LTC)*, Poznań, Poland, 2007.

Mazur P. and Dale R. WikiWars: A New Corpus for Research on Temporal Expressions. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, 913–922, Stroudsburg, PA, USA, 2010. `http://dl.acm.org/citation.cfm?id=1870658.1870747`.

Meyers A., Reeves R., Macleod C., Szekely R., Zielinska V., Young B., and Grishman R. The NomBank Project: An Interim Report. In Meyers A., editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, 24–31, Boston, Massachusetts, USA, 2004.

Minard A.L., Marchetti A., and Speranza M. Event Factuality in Italian: Annotation of News Stories from the Ita-TimeBank. *Proceedings of CLiC-it 2014, First Italian Conference on Computational Linguistic*, 2014.

Minard A.L., Speranza M., Agirre E., Aldabe I., van Erp M., Magnini B., Rigau G., and Urizar R. SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 778–786, Denver, Colorado, 2015.

Minard A.L., Speranza M., Urizar R., Altuna B., van Erp M., Schoen A., and van Son C. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In Calzolari N., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., and Piperidis. S., editors, *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*. 2016.

Mirza P. and Minard A.L. HLT-FBK: a Complete Temporal Processing System for QA TempEval. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 801–805, Denver, Colorado, 2015.

Mirza P. and Tonelli S. An Analysis of Causality between Events and its Relation to Temporal Information. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2097–2106. 2014.

Morante R., Schrauwen S., and Daelemans W. Annotation of negation cues and their scope: Guidelines v1. *Computational linguistics and psycholinguistics technical report series, CTRS-003*, 2011.

Moriceau V. and Tannier X. French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., and Piperidis S., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, 2014.

Mostafazadeh N., Chambers N., He X., Parikh D., Batra D., Vanderwende L., Kohli P., and Allen J. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 839–849, San Diego, California, 2016.

Murat A., Li X., Jiang T., Yang Y., Zhou X., and Wang L. Research on Temporal Expression Recognition: A Case Study of Uyghur. *ICIC Express Letters*, 8(4):703–710, 2017.

Orasmaa S. Towards an Integration of Syntactic and Temporal Annotations in Estonian. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., and Piperidis S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1259–1266, Reykjavik, Iceland, 2014.

Orasmaa S. *Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience*. Doktoretza-tesia, University of Tartu, Estonia, 2016. `http://hdl.handle.net/10062/54752`.

Palmer M., Gildea D., and Kingsbury P. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

Pennington J., Socher R., and Manning C. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543, Doha, Qatar, 2014.

Pustejovsky J., Belanger L., Castaño J., Gaizauskas R., Hanks P., Ingria B., Katz G., Radev D., Rumshisky A., Sanfilippo A., Saurí R., Setzer A., Sundheim B., and Verhagen M. NRRC Summer Workshop on Temporal and Event Recognition for Question Answering Systems. Barne-txostena, National Regional Research Center (NRRC), 2002.

Pustejovsky J., Castaño J.M., Ingria R., Saurí R., Gaizauskas R.J., Setzer A., Katz G., and Radev D.R. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3: 28–34, 2003a.

Pustejovsky J., Hanks P., Saurí R., See A., Gaizauskas R., Setzer A., Radev D., Sundheim B., Day D., Ferro L., and Lazo. M. The TimeBank Corpus. In Archer D., Rayson P., Wilson A., and McEnery T., editors, *Proceedings of Corpus Linguistics 2003*, number 16, 647–656, Lancaster, UK, 2003b.

Pustejovsky J., Moszkowics J., and Verhagen M. ISO-Space: The Annotation of Spatial Information in Language. In Blunt H., editor, *Proceedings of the Sixth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-6)*, 1–9, Oxford, UK, 2011.

Pustejovsky J., Verhagen M., Saurí R., Littman J., Gaizauskas R., Katz G., Mani I., Knippen R., and Setzer A. TimeBank 1.2. Barne-txostena, Linguistic Data Consortium, 2006. https://catalog.ldc.upenn.edu/LDC2006T08.

Radinsky K. and Horvitz E. Mining the web to predict future events. *Proceedings of the sixth ACM international conference on Web search and data mining*, 255–264. ACM, 2013.

Reichenbach H. The tenses of verbs. *Elements of Symbolic Logic*, 287–298. The Macmillan Company, New York, 1947.

Reynaert M., Oostdijk N., Clercq O.D., van den Heuvel H., and de Jong F. Balancing SoNaR: IPR versus Processing Issues in a 500-Million-Word Written Dutch Reference Corpus. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M., and Tapias D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2693–2398, Valletta, Malta, 2010.

Robaldo L., Caselli T., Russo I., and Grella M. From Italian Text to TimeML Document via Dependency Parsing. *Computational Linguistics and Intelligent Text Processing*, 177–187. Springer, 2011.

Rose T., Stevenson M., and Whitehead M. The Reuters Corpus Volume 1-from Yesterday's News to Tomorrow's Language Resources. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. 2002.

Salaberri H., Altuna B.A., Aranzabe M.J., Arregi O., and Díaz de Ilarraza A. bTime: a Hybrid Architecture for Capturing Temporal Information in Basque. *Knowledge-Based Systems*, Under revision.

Salaberri Izko H. *Rol semantikoen etiketatzeak testuetako espazio-denbora informazioaren prozesamenduan daukan eraginaz*. Doktoretza-tesia, Universidad del País Vasco/Euskal Herriko Unibertsitatea, Donostia, Euskal Herria, 2017.

Sandhaus E. The New York Times Annotated Corpus Overview. Barne-txostena, The New York Times Company, Research and Development, New York, 2008. https://catalog.ldc.upenn.edu/docs/LDC2008T19/new_york_times_annotated_corpus.pdf.

Saurí R. *A Factuality Profiler for Eventualities in Text.* Doktoretza-tesia, Faculty of the Graduate School of Arts and Sciences, Waltham, MA, USA, 2008. AAI3304029.

Saurí R. and Badia T. Catalan TimeBank 1.0. Barne-txostena, Linguistic Data Consortium, Philadelphia, Pennsylvania, 2012. `https://catalog.ldc.upenn.edu/LDC2012T10`.

Saurí R., Batiukova O., and Pustejovsky J. Annotating Events in Spanish. TimeML Annotation Guidelines. Barne-txostena, Technical Report Version TempEval-2010., Barcelona Media-Innovation Center, 2009.

Saurí R., Knippen R., Verhagen M., and Pustejovsky J. Evita: A Robust Event Recognizer for QA Systems. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, 700–707, Stroudsburg, PA, USA, 2005.

Schilder F. and Habel C. From Temporal Expressions to Temporal Information: Semantic Tagging of News Messages. *Proceedings of the Workshop on Temporal and Spatial Information Processing - Volume 13*, TASIP '01, 9:1–9:8, Stroudsburg, PA, USA, 2001. `http://dx.doi.org/10.3115/1118238.1118247`.

Schuurman I., Hoste V., and Monachesi P. Interacting Semantic Layers of Annotation in SoNaR, a Reference Corpus of Contemporary Written Dutch. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M., and Tapias D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, 2471–2477, Valletta, Malta, 2010.

Setty V., Bedathur S., Berberich K., and Weikum G. InZeit: Efficiently Identifying Insightful Time Points. *Proceedings of the VLDB Endowment*, 3(1-2):1605–1608, 2010.

Setzer A. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study.* Doktoretza-tesia, University of Sheffield, Sheffield, UK, 2001. `ftp://ftp.dcs.shef.ac.uk/home/robertg/theses/setzer_thesis.pdf`.

Skukan L., Glavaš G., and Šnajder J. HeidelTime.HR: Extracting and Normalizing Temporal Expressions in Croatian. *Proceedings of the 9th Language Technologies Conference*, 99–103, 2014.

Smith C.S. *The parameter of aspect.* Kluwer, 1991.

Strötgen J. and Gertz M. WikiWarsDE: a German Corpus of Narratives Annotated with Temporal Expressions. In Hedeland H., Schmidt T., and Wörner K., editors, *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*, 129–134, Hamburg University, 2011.

Strötgen J. and Gertz M. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013.

Stubbs A. MAE and MAI: Lightweight Annotation and Adjudication Tools. *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, 129–133, Stroudsburg, PA, USA, 2011.

Styler W.F., Bethard S., Finan S., Palmer M., Pradhan S., de Groen P.C., Erickson B., Miller T., Lin C., Savova G., *et al.*. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143–154, 2014.

Sun W., Rumshisky A., and Uzuner O. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46(Supplement): S5–S12, 2013. 2012 i2b2 NLP Challenge on Temporal Relations in Clinical Data.

Taulé M., Martí M.A., and Recasens M. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., and Tapias D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco, 2008. ACL Anthology Identifier: L08-1222.

TimeML Working Group. TimeML Annotation Guidelines version 1.3. Manuscript. Barne-txostena, Brandeis University, 2010.

Tonelli S., Sprugnoli R., Speranza M., and Minard A.L. NewsReader Guidelines for Annotation at Document Level. version final. Barne-txostena, Fondazione Bruno Kessler, 2014. `http://www.newsreader-project.eu/files/2014/12/NWR-2014-2-2.pdf`.

Tran G.B., Herder E., and Markert K. Joint Graphical Models for Date Selection in Timeline Summarization. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015*, 1: Long Papers lib., 1598–1607, Beijing, China, 2015.

UzZaman N. and Allen J.F. Event and Temporal Expression Extraction from Raw Text: First Step towards a Temporally Aware System. *International Journal of Semantic Computing*, 4(04):487–508, 2010.

van de Camp M. and Christiansen H. Resolving relative time expressions in Dutch text with Constraint Handling Rules. *Revised Selected Papers of the 7th International Workshop on Constraint Solving and Language Processing - Volume 8114*, CSLP 2012, 166–177, New York, NY, USA, 2013.

Vendler Z. Verbs and times. *The philosophical review*, 143–160, 1957.

Verhagen M. The Brandeis Annotation Tool. In Calzolari N., Choukri K., Maegaard B., Mariani J., Odijk J., Piperidis S., Rosner M., and Tapias D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

Verhagen M., Gaizauskas R., Schilder F., Hepple M., Katz G., and Pustejovsky J. SemEval-2007 Task 15: TempEval Temporal Relation Identification. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 75–80, Prague, Czech Republic, 2007.

Verhagen M., Knippen R., Mani I., and Pustejovsky J. Annotation of temporal relations with Tango. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2249–2252, Genoa, Italy, 2006.

Verhagen M., Mani I., Sauri R., Knippen R., Jang S.B., Littman J., Rumshisky A., Phillips J., and Pustejovsky J. Automating Temporal

Annotation with TARSQI. *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions*, ACLdemo '05, 81–84, Stroudsburg, PA, USA, 2005. http://dx.doi.org/10.3115/1225753.1225774.

Verhagen M. and Pustejovsky J. Temporal Processing with the TARSQI Toolkit. *22nd International Conference on on Computational Linguistics: Demonstration Papers*, COLING '08, 189–192, Stroudsburg, PA, USA, 2008. http://dl.acm.org/citation.cfm?id=1599288.1599300.

Vossen P., Rigau G., Serafini L., Stouten P., Irving F., and Hage W.V. NewsReader: recording history from daily news streams. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., and Piperidis S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2000–2007, Reykjavik, Iceland, 2014. ACL Anthology Identifier: L14-1374.

Widlöcher A. and Mathet Y. La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, Frantzia, 2009.

Wonsever D., Malcuori M., and Rosá A. SIBILA: Esquema de anotación de eventos. Barne-txostena, Instituto de Computación, Universidad de la República, 2008. https://www.colibri.udelar.edu.uy/handle/123456789/3419.

Wonsever D., Rosá A., Malcuori M., and Etcheverry M. TEMANTEX: A Markup Language for Spanish Temporal Expressions and Indicators. *Research in Computing Science*, 97:9–19, 2015.

Wonsever D., Rosá A., Malcuori M., Moncecchi G., and Descoins A. Event annotation schemes and event recognition in spanish texts. In Gelbukh A., editor, *Computational Linguistics and Intelligent Text Processing*, 7182 lib. of *Lecture Notes in Computer Science*, 206–218. Springer Berlin Heidelberg, 2012.

Wu M., Li W., Lu Q., and Li B. CTEMP: A Chinese Temporal Parser for Extracting and Normalizing Temporal Information. *Natural Language Processing–IJCNLP 2005*, 694–706. Springer, 2005.

Xue N., Xia F., Chiou F.D., and Palmer M. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(02):207–238, 2005.

Yaghoobzadeh Y., Ghassem-Sani G., Mirroshandel S.A., and Eshaghzadeh M. ISO-TimeML Event Extraction in Persian Text. *Proceedings of COLING 2012*, 2931–2944, Mumbai, India, 2012. http://www.aclweb.org/anthology/C12-1179.

Yan R., Wan X., Otterbacher J., Kong L., Li X., and Zhang Y. Evolutionary timeline summarization: A balanced optimization framework via iterative substitution. *In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '11*, 745–754. 2011.

Zhong X. and Cambria E. Time Expression Recognition Using a Time-related Tagging Scheme. *Proceedings of the International World Wide Web Conference 2018*, Lyon, France, 2018. http://sentic.net/time-expression-recognition.pdf.