Cultural Aspects of Spatiotemporal Analysis in Multilingual Applications

Ineke Schuurman, Vincent Vandeghinste

Centrum voor Computerlinguïstiek, K.U.Leuven
ineke.schuurman@ccl.kuleuven.be
vincent.vandeghinste@ccl.kuleuven.be

Abstract

In this paper we want to point out some issues arising when a natural language processing task involves several languages (like multilingual, multidocument summarization and the machine translation aspects involved) which are often neglected. These issues are of a more cultural nature, and may even come into play when several documents in a single language are involved. We pay special attention to those aspects dealing with the *spatiotemporal* characteristics of a text.

Correct automatic selection of (parts of) texts such as handling the same eventuality, presupposes spatiotemporal disambiguation at a rather specific level. The same holds for the analysis of the query. For generation and translation purposes, spatiotemporal aspects may be relevant as well. At the moment English (both the British and American variants) and Dutch (the Flemish and Dutch variant) are covered, all taking into account the perspective of a contemporary, Flemish user. In our approach the cultural aspects associated with for example the language of publication and the language used by the user play a crucial role.

1. Introduction

When a journalist adapts a news item for a Flemish[1] newspaper from an American source, she has to do more than a proper translation. Dates every American citizen is familiar with, like `Halloween` may be unknown to people in Flanders. The same holds for place names, such as `Dover`: what is its location? The journalist has to *localize* the text.

Cultural localization is also necessary when adapting news from a *local* to a *national* newspaper, using the Gricean maxims, as the intended audience changes (Schuurman, 2007b). Unwittingly, a journalist will use the Gricean maxims even when writing a news item for a specific newspaper as it determines for example the desired level of details.

The notion `Christmas`, without further specification, in document A does not necessarily refer to the same date(s) as the notion `Christmas` in document B. This is also true for unspecified references to `Dover`.

When automatically analyzing texts, non-lingual factors should not be factored out. Such factors can be related to *location, period, religion, observance, tradition*.

How can we achieve this for *spatiotemporal phenomena*? In the AMASS++ project, on which we focus in this paper and which deals with multi-document, multilingual summarization, we use the same approach as in the SoNar core corpus (Schuurman et al., 2010), which consists of one million words of texts with manually corrected syntactic and semantic (coreference resolution, named entity recognition, semantic role labeling) annotations.

In AMASS++ however we work with plain text which is often just tokenized and part-of-speech tagged. For English texts, this is possibly extended with named entity recognition and semantic roles. Contrary to SONaR, however, the annotations are not manually corrected.

In section 2. we describe the role of spatiotemporal characteristics in an application like AMASS++. Section 3. describes the general architecture of such a system. Section 4. focuses on the multicultural aspect of the STex annotation scheme and section 5. draws some conclusions.

[1]Flanders is the Dutch speaking region of Belgium

2. The role of spatiotemporal characteristics in multi-document, multilingual summarization

Multimedia and multilingual archives increasingly become an important source of information for governments, companies and citizens. There is a large need for effective and efficient tools for information retrieval. An automated synthesis of the information across media and languages is here of primordial importance.

The main objectives of the AMASS++ project are:

- the alignment of equivalent content across documents, media[2] and languages
- the generation of structured cross-media and cross-lingual summaries

As a test case we use news archives of Dutch (Flemish)-speaking and English-speaking broadcasters.

In this paper we give a general description of the role of spatiotemporal characteristics in section 2.1. and present an example in section 2.2.

2.1. Description

No matter what architecture is chosen for multi-document summarization, spatiotemporal characteristics may play an important role in

- the alignment of documents,
- the analysis of the query,
- the search for content,
- the generation of the summary, especially the ordering of elements, and
- the translation proces,

[2]These multimedia aspects will be left aside in this paper.

Process	Task
Monolingual Alignment	- explicitation of information - linking of information (equivalent information)
Crosslingual Alignment	- linking of information (translational equivalents)
Query	- analysis - translation
Search	- inference procedures - reasoning
Summarization I	- filtering - ordering
Machine Translation	- translational equivalents (esp. tense & aspect)
Summarization II	- filtering - ordering

Table 1: Use of spatiotemporal analysis

Table 1 shows the tasks and processes in which spatiotemporal analysis is used.

Figure 1 shows the alignment process: all documents are preprocessed for content-alignment per language, like NL (Dutch), EN (English). Ideally, the documents undergo the following preprocessing steps:

- part-of-speech tagging and lemmatization,
- topic identification,
- coreference resolution,
- named entity recognition,
- semantic role detection, and
- **spatiotemporal analysis**.

The annotations resulting from these preprocessing steps are used as features in monolingual content alignment. Parts of the documents (in the same document or in different documents) may be explicitly marked as containing equivalent content (cf. the ellipses in the middle of figure 1), although, thus far this only represents content in the same language.
Using, among other things, the cross-lingual identifiers in the different Wordnets (Vossen, 1998) and the language-independent spatiotemporal and named entity values, content is aligned between documents in different languages as well (bottom of figure 1).

The input for summarization consists of one or more content-aligned documents, possibly in a mix of several languages. The output of the system is a summary (preferably query-focused), in a language chosen by the user.
We sketch the different possibile architectures with respect to where MT is introduced in the processing chain in section 3.
In the AMASS++ project we do not dispose of all the mentioned annotation layers, and the same kind of information is not available for both languages involved: Dutch and English.

Although part of the material will be parsed, i.e. that part that will be translated automatically (see section 3.), most of the documents in an archive or another large collection of texts are just available in the original language and do not contain any annotation layers. Cheap and fast annotations, such as part-of-speech tagging can be applied, but parsing is computationally too heavy.

2.2. An Example Scenario

In this section we show by an example which role spatiotemporal characteristics can play in document analysis for multi-document summarization.

Example Scenario:

- suppose several bombs exploded in *Dover (UK)* in *January 2008*
- a journalist wants to consult the archive to detect whether there have been riots and other disturbances *in Dover and its broader environs* over *the last two decades*
- she also wants the background of the organizations involved in these cases
- she wants the relevant data in a chronological order

Such a query can only be answered when the documents in the archive are annotated with geospatial and temporal features. In this case, only *Dover* in the *United Kingdom*, in the county of *Kent*, will be of interest. Even when we are not able to tell exactly which town and villages belong to the *broader environs* (a vague notion as such), we are able to tell which places definitely do not. Therefore, rows in *Chestnut Knoll* or *Little Creek* (both in the *US*) are not of interest. Neither is a bomb near *Dover Castle* on *Christmas Eve 1914*, whereas the fact that on *September 22nd, 1989* eleven military bandsmen in *Dover* were killed by the IRA might be.
When the query is analyzed, *Dover* is recognized as a geospatial expression referring to the town of *Dover*, in *Kent, UK* with the following tag:

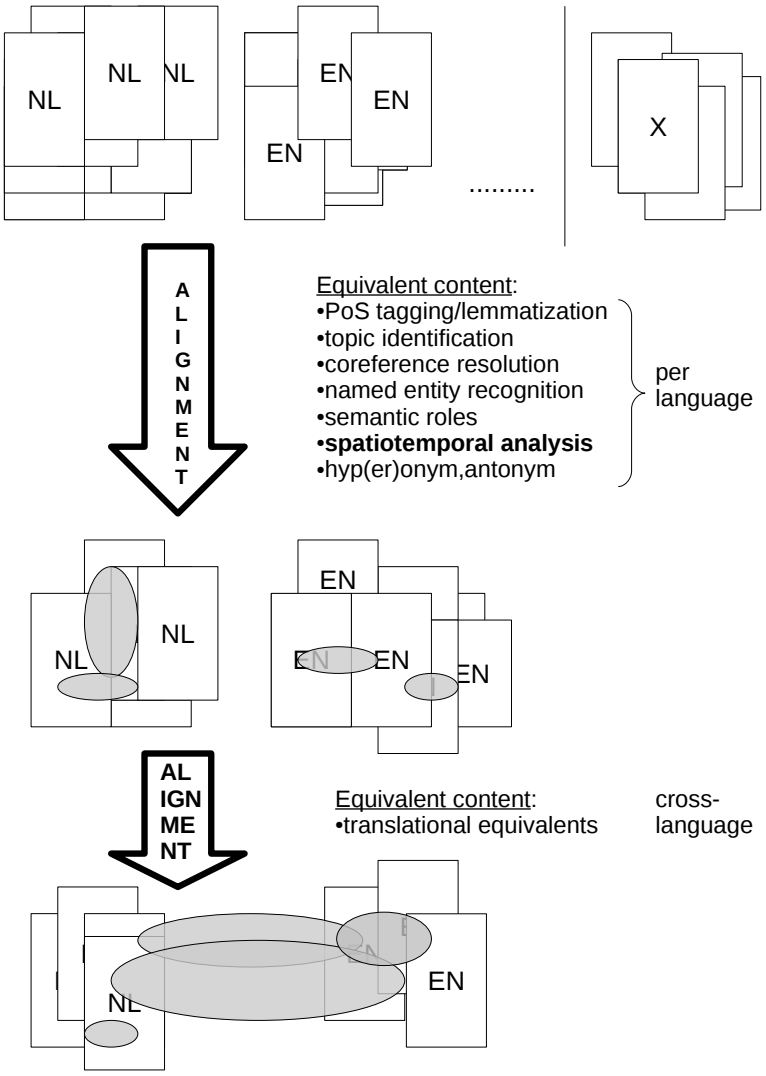

Figure 1: The alignment process

```
<geo type="place"
 id="55039"
 value="EU::GB::EN::Kent::Dover::Dover"
 coord="51.1295,1.3089"/>
```

The system needs to look for mentions of *Dover* in the archive, also using other spellings, alternative names, etc. Taking the tag of *Dover, Kent* (cf. above) into account, the correct *Dover* can be found in the documents. Finding hits for incidents in the unnamed surroundings of the *British Dover* can be done by taking into account those places that are in the same county (reflected in the `geo`-tags assigned) and/or have coordinates which indicate closeness.

3. Architecture

In multilingual multidocument summarization, there are several different architectural options. Figure 2 shows the different options as to where to insert machine translation in the full processing chain. Note that all $X_\lambda, Y_\lambda, Z_\lambda$ may have undergone some monolingual preprocessing steps, like the ones mentioned in section 2.1.

Approach A (MT of complete archive) is followed in the DUC task on multilingual summarization, see for example Evans et al. (2005). Such a translation of large amounts of documents is very expensive, while the summarization process is carried out on not always correct translations. Note that in our project, in which the user decides in which language she wants the summary, this may implicate a second translation in case she opts for a summary in another language than that used for the initial translation. The same holds for the approaches B and C, although the cost of translation is smaller.

When the translation is done at a much later stage (approach D), the initial, larger summarization task is done on documents in the original languages (which may lead to better summaries), but in this case a second summarization step is necessary after translation of the first versions in the desired language: two-stage summarization.

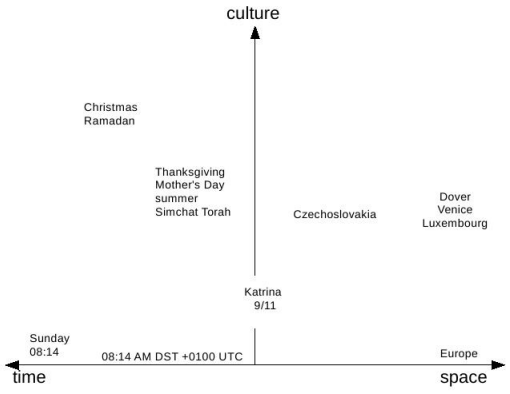

Figure 3: The spatiotemporal continuum

Translation at a later stage in the process is advocated in a.o. Lenci et al. (2002).
We select architecture D as it is the least costly and requires the least text to be translated. As MT can lead to ungrammatical output (which is less the case for summarization output) opting for MT as late in the processing chain as possible would keep the proliferation of errors due to the concatenation of several imperfect processing steps under control. It is this scenario that is described in table 1.

The machine translation engine we use for these purposes is a *Parse and Corpus-based Machine Translation engine* (called PaCo-MT) (Vandeghinste and Martens, 2010). It is an example-based translation engine with the architecture of a transfer rule-based MT system. The MT engine uses monolingual parsers to parse the source language sentences. Transfer rules are induced from a large parallel treebank (Tiedemann and Kotzé, 2009) and convert the source language tree into a target language tree, from which an output sentence is generated (Vandeghinste, 2009).
Of course, those parts of the text collection that are parsed in the context of the MT-job, will be saved as such. Over time spatial temporal analysis will be able to take parsed input into account as well, for example with respect to search for content and alignment of documents.

4. STEx and multiculturality

Geospatial and temporal analysis are often performed separately (Mani and others, 2008; Pustejovsky and Moszkowicz, 2008; TimeML Working Group, 2010), and in summarization usually only temporal analysis is used. As far as we are aware, cultural issues are not yet taken into account. In STEx, (Schuurman, 2007a; Schuurman, 2007b; Schuurman, 2008), integrated spatiotemporal analysis is at the heart of the matter to analyze expressions like `10:15 AM`, `summer`, `Christmas`, `Thanksgiving`, `Dover`, `Czechoslovakia` in a detailed way. Note that without mentioning the time zone (or the location in order to infer the time zone) an expression like `10:15 AM` is not informative enough for multi-document applications as it allows several interpretations. The same holds for expressions like `summer`: does it include the month of *May*? Geospatial information should also be as specific as possible: when in one document in *May 2006* a girl was murdered in `Heverlee`, and in another document in that same month a girl was murdered in `Leuven`, the system should be able to determine whether these two documents are talking about the same event: are these alternative names, is the one part of the other,... Therefore, the annotation needs to be as explicit as possible.
Combined spatiotemporal knowledge is needed to determine when and where expressions like those in figure 3 are located on a time axis or a map.
From a methodological point of view, linking the (geo)spatial and temporal approaches in STEx is quite obvious, as both approaches are similar (cf. Table 2).

temporal	**geospatial**
time of perspective	place of perspective
time of location	place of location
time of eventuality	place of eventuality
duration	distance
shift of perspective	shift of perspective
relations	relations

Table 2: Similar approaches

If there are several candidates when a specific spatiotemporal unit is analyzed,[3] STEx will select the one the intended audience, respectively the present-day user is expected to select, based on cultural properties: in Russia, *Christmas* will not be celebrated on the *25th of December*, but *13 days later*. For people in Flanders *Dover* will refer to the town in the *UK*, whereas many people in the *US* will not know of this British Dover, and instead associate the name with the capital of *Delaware*. Even the months associated with the very familiar notion *summer* are not the same all over the world (northern vs. southern hemisphere).
In STEx a large database containing spatiotemporal and (associated) cultural information is used to disambiguate such concepts.
The influences heaped together under *culture* in figure 3 can be found in many fields:

- tradition (Christian, Jewish, ...),
- geographical background,
- upbringing,
- social background,
- ....

As can be seen there are little temporal phenomena that can be interpreted without spatial knowledge, or, to put it differently, that do not contain (geo)spatial information. The same holds, albeit to a lesser extent, for (geo)spatial phenomena.[4] *Czechoslovakia* is a relevant example, as this

[3] Assuming that the context does not contain any indications for disambiguation.

[4] This may be due to the fact that we ignore the development of, say, a country over the years. For example: the current shape of the Netherlands is not identical to that of the country in the beginning of the 20th century, but we will abstract away from this.

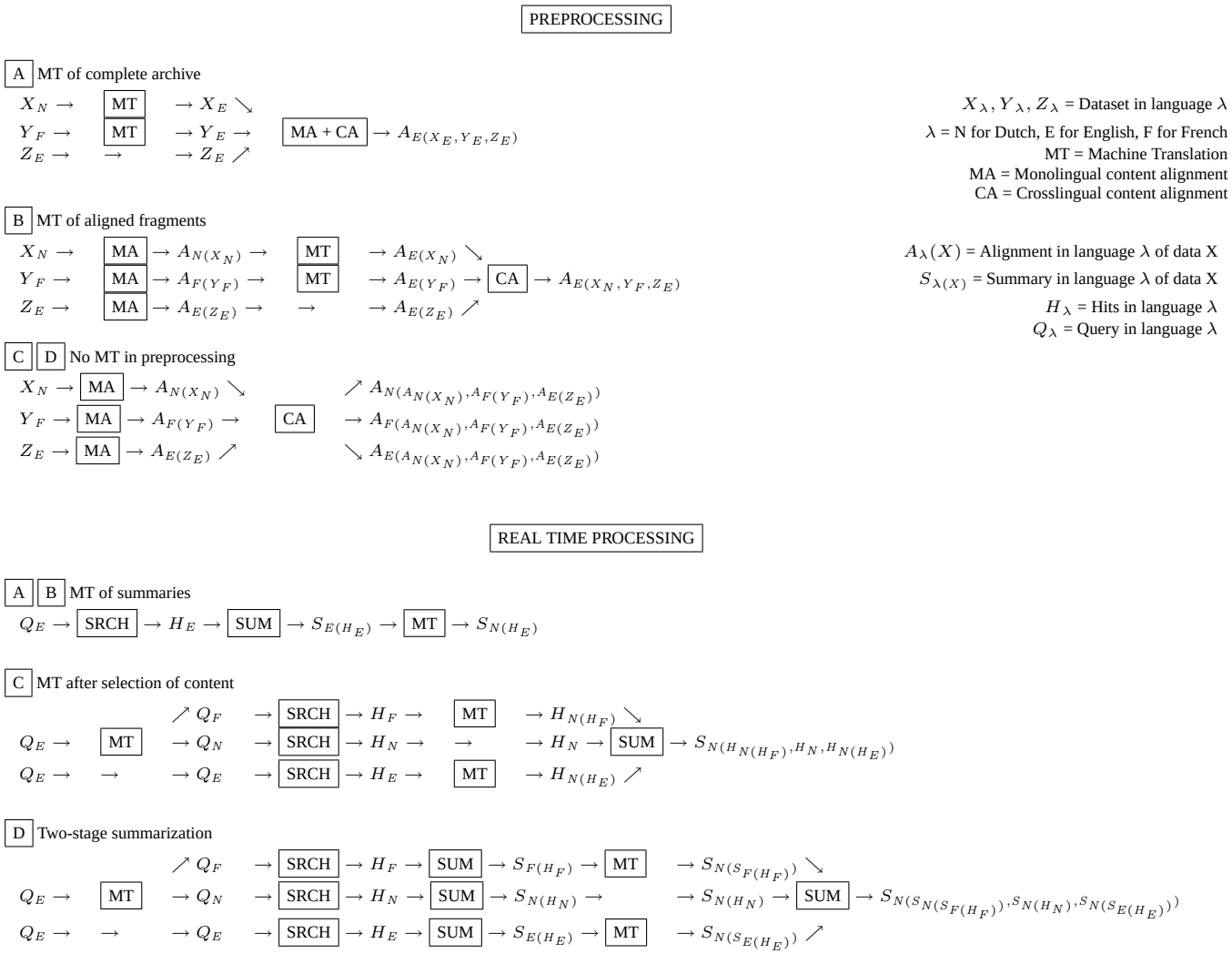

Figure 2: Some architectures for multilingual summarization

country does not exist anymore. As in STEx everything is mapped onto a contemporary map, *Czechoslovakia* is said to exist spatially of the current *Czech Republic* and *Slovakia*. A geospatial element like *Czechoslovakia* therefore gets a combined spatiotemporal tag:

```
<geo id="52467"
     type="country"
     val="EU::CS">
  <parts>
    <geo type="country"
         val="EU::CZ"/>
    <geo type="country"
         val="EU::SK"/>
  </parts>
  <temp type="cal"
        val="1918/1990"/>
</geo>
```

An expressions like *9/11* also has become an event instead of just a date and is associated with temporal and geospatial information. The temporal interpretation of expressions like *Thanksgiving, Mother's day* and *summer* depends on the country involved, sometimes even on the part of the country. *Moederdag* [EN: Mother's day], for example is celebrated in the *province of Antwerp* on *August 15*, while the default value in *Belgium* is the *second Sunday in May*: For Antwerp this results in the following tag:

```
<temp id="89762"
      type="cal"
      val="XXXX-08-15">
  <geo type="province"
       val="EU::BE::VL::Antwerpen"/>
</temp>
```

The default value for Belgium:

```
<temp id="89763"
      type="cal"
      val="XXXX-05-D07&8..14">
  <geo type="country"
       val="EU::BE"/>
</temp>
```

Our database contains several entries for *Moederdag*. Which one applies when no date is specified is a matter of culture.

When a specific event nowadays is celebrated on another date than in the past, which is for instance the case with the

Dutch *Koninginnedag* [EN: Queen's Day] this can be stated as follows.
Before 1948:

```
<temp id="90452"
      type="cal"
      val="XXXX-08-31">
  <temp type="cal"
        val="1898/1948"/>
  <geo type="country"
       val="EU::NL"/>
</temp>
```

Since 1949:

```
<temp id="90451"
      type="cal"
      val="XXXX-04-30">
  <geo type="country"
       val="EU::NL"/>
</temp>
```

In a 1930's Dutch newspaper, an unspecified use of *Koninginnedag* needs to be associated with entry 90452, while in a recent one it should be entry 90451.

These examples show that both geospatial and temporal expressions may be easily misinterpreted by humans and machines when they are not disambiguated thoroughly.

5. Conclusions

We described the importance of a detailed spatiotemporal analysis for AMASS++, emphasizing to take cultural factors into account. STEx uses an integrated, spatiotemporal annotation system, which makes ample use of such cultural information. The annotation reflects common knowledge the intended audience of the original documents and the user of the application are assumed to have. Our annotation is rather explicit and detailed. In this respect it deviates from TimeML (TimeML Working Group, 2010), resp. SpatialML (Mani and others, 2008). In the first, expressions like *summer* and *Christmas* are marked as temporal, without linking them to a time axis.[5] We consider this a serious drawback. The same holds for SpatialML, in which *Dover* could be recognized as a populated place in the UK,[6] but not that the UK is in Europe, or that Dover is in the county of Kent. For example for content alignment accross documents our system shows more potential. To us, it is certainly worth the efforts of building the STEx database, an ongoing effort.

6. Acknowledgements

This research was supported by the AMASS++ Project (http://www.cs.kuleuven.be/∼liir/projects/amass/) IWT (SBO IWT 060051).

[5] James Pustejovsky informed us that this might be changed in TimeML as well.

[6] Coordinates are provided as well.

7. References

D.K. Evans, K. McKeown, and J.L. Klavans. 2005. Similarity-based Multilingual Multi-Document Summarization. Technical report, Columbia University.

A. Lenci, R. Bartolini, N. Calzolari, A. Agua, S. Busemann, E. Cartier, K. Chevreau, and J. Coch. 2002. Multilingual Summarization by Integrating Linguistic Resources in the MLIS-MUSI Project. In *Proceedings of LREC*, Las Palmas, Spain.

I. Mani et al. 2008. SpatialML: Annotation Scheme, Corpora, and Tools. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

J. Pustejovsky and J.L. Moszkowicz. 2008. Integrating Motion Predicate Classes with Spatial and Temporal Annotations. In *Proceedings of Coling 2008*, pages 95–98, Manchester, U.K.

I. Schuurman, V. Hoste, and P. Monachesi. 2010. Interacting Semantic Layers of Annotation in SoNaR, a Reference Corpus of Contemporary Written Dutch. In *Proceedings of LREC*.

I. Schuurman. 2007a. Spatiotemporal Annotation on Top of an Existing Treebank. In K. De Smedt, J. Hajic, and S. Kuebler, editors, *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, pages 151–162, Bergen, Norway.

I. Schuurman. 2007b. Which New York, which Monday? The role of background knowledge and intended audience in automatic disambiguation of spatiotemporal expressions. In *Proceedings of CLIN 17*.

I. Schuurman. 2008. Spatiotemporal annotation using MiniSTEx: How to deal with alternative, foreign, vague and obsolete names? In *Proceedings of LREC 2008*, Marrakech, Morocco.

J. Tiedemann and G. Kotzé. 2009. Building a large machine-aligned parallel treebank. In *Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories*, Milan, Italy. Catholic University of the Sacred Heart.

TimeML Working Group, 2010. *TimeML Annotation Guidelines, version 1.3*, February 9.

V. Vandeghinste and S. Martens. 2010. Bottom-up transfer in example-based machine translation. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*.

V. Vandeghinste. 2009. Tree-based target language modeling. In Lluìs Màrquez and Harold Somers, editors, *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, pages pp.152–159, Barcelona, Spain. Universitat Politècnica de Catalunya.

P. Vossen, editor. 1998. *eurowordnet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.