WikiWars: A New Corpus for Research on Temporal Expressions

Paweł Mazur[1,2]

[1]Institute of Applied Informatics,
Wrocław University of Technology
Wyb. Wyspiańskiego 27,
50-370 Wrocław, Poland
pawel@mazur.wroclaw.pl

Robert Dale[2]

[2]Centre for Language Technology,
Macquarie University,
NSW 2109, Sydney, Australia
Pawel.Mazur@mq.edu.au
Robert.Dale@mq.edu.au

Abstract

The reliable extraction of knowledge from text requires an appropriate treatment of the time at which reported events take place. Unfortunately, there are very few annotated data sets that support the development of techniques for event time-stamping and tracking the progression of time through a narrative. In this paper, we present a new corpus of temporally-rich documents sourced from English Wikipedia, which we have annotated with TIMEX2 tags. The corpus contains around 120000 tokens, and 2600 TIMEX2 expressions, thus comparing favourably in size to other existing corpora used in these areas. We describe the preparation of the corpus, and compare the profile of the data with other existing temporally annotated corpora. We also report the results obtained when we use DANTE, our temporal expression tagger, to process this corpus, and point to where further work is required. The corpus is publicly available for research purposes.

1 Introduction

The reliable processing of temporal information is an important step in many NLP applications, such as information extraction, question answering, and document summarisation. Consequently, the tasks of identifying and assigning values to temporal expressions have recently received significant attention, resulting in the creation of mature corpus annotation guidelines (e.g. TIMEX2[1] and TimeML[2]), publicly available annotated corpora (ACE,[3] TimeBank[4]) and a number of automatic taggers (see, for example, (Mani and Wilson, 2000; Schilder, 2004; Hacioglu et al., 2005; Negri and Marseglia, 2005; Saquete, 2005; Han et al., 2006; Ahn et al., 2007)).

However, existing corpora have their limitations. In particular, the documents in these corpora tend to be limited in length and, in consequence, discourse structure. This impacts on the number, range and variety of temporal expressions they contain. Existing research carried out on the interpretation of temporal expressions, e.g. by (Baldwin, 2002; Ahn et al., 2005; Mazur and Dale, 2008), suggests that many temporal expressions in documents, especially news stories, can be interpreted fairly simply as being relative to a reference date that is typically the document creation date. This phenomenon does not carry over to longer, more narrative-style documents that describe extended sequences of events, as found, for example, in biographies or descriptions of protracted geo-political events. Consequently, existing corpora are not ideal as development data for systems intended to work on such historical narrations.

In this paper we introduce a new annotated corpus of temporal expressions that is intended to address this shortfall. The corpus, which we call WikiWars, consists of 22 documents from English Wikipedia that describe the historical course of wars. Despite the small number of documents, their length means that the corpus yields a large number of temporal expressions, and poses new challenges for tracking

[1]See http://fofoca.mitre.org.

[2]See http://timeml.org.

[3]See corpora LDC2005T07 and LDC2006T06 in the LDC catalogue (http://www.ldc.upenn.edu).

[4]See corpus LDC2006T08 in the LDC catalogue.

temporal focus through extended texts. The corpus has been made available for others to use;[5] to give an indication of the difficulty of processing the temporal phenomena in the texts, we also report on the performance of DANTE, our temporal expression tagger, on detecting and interpreting the temporal expressions in the corpus.

The rest of this paper is organised as follows. In Section 2 we describe related work, focusing on the TIMEX2 annotation scheme, and existing corpora that contain annotations of temporal expressions using this scheme. Section 3 describes the process of creation of the WikiWars corpus. In Section 4 we comment on some artefacts of Wikipedia articles that impact on the annotation process and the use of this corpus. Then, in Section 5 we analyse the differences between the WikiWars corpus and the widely-used ACE corpora. In Section 6 we report on the performance of our temporal expression tagger on this data set. Finally, in Section 7, we conclude.

2 Related Work

At the time of writing, there are two mature, wide-coverage schemes for the annotation of temporal information in texts: TIMEX2 (Ferro et al., 2005) and TimeML (Pustejovsky et al., 2003; Boguraev et al., 2005), which is soon to become an ISO standard (Pustejovsky et al., 2010).

These schemes were used to annotate corpora that are often used in research on temporal expression recognition and normalisation: the series of corpora used for training and evaluation in the Automatic Content Extraction (ACE) program[6] run in 2004, 2005 and 2007, and the TimeBank Corpus.

The ACE corpora were prepared for the development and evaluation of systems participating in the ACE program. However, the evaluation corpora have never been publicly released, and thus are currently, for all practical purposes, unavailable. The ACE 2004 corpus contains news data only (broadcast news, newspaper and newswire), while the ACE 2005 and 2007 corpora contain news (broadcast and newswire), conversations (broadcast and telephone), UseNet discussions and web blogs. The 2005 and 2007 ACE corpora are annotated with the latest version of TIMEX2 (2005), while the 2004 corpus is annotated with the older 2003 version of TIMEX2; however, the differences are not very significant.

Apart from the unavailability of the evaluation data, there are two issues with the ACE corpora. One is that most of the documents are relatively short, so that the average number of temporal expressions per document is low (typically between seven and nine per document, including the document time stamp as a metadata element). This results in very limited temporal discourse structure, and relatively few underspecified and relative temporal expressions. Unfortunately, these are the more difficult temporal expressions to handle, and so the ACE corpora may not serve as a good baseline for performance more generally.

A second problem is that the ACE corpora appear to contain a significant number of errors in the gold standard annotations, with respect to both the annotated extents and the semantic values assigned, which do not always follow the TIMEX2 guidelines.

TimeBank v1.2 is a revised and improved version of TimeBank 1.1 resulting in a number of errors fixed and inconsistencies removed (see (Boguraev et al., 2007)). Unfortunely, this corpus has the same limitations as the ACE corpora in regard to document length and complexity of discourse structure. Further, TimeBank is annotated with TimeML, a scheme more complex than TIMEX2 since it also encompasses the tagging of events and temporal relations. However, TIMEX2 is sufficiently sophisticated for the annotation of most types of temporal expressions, and our review of the literature reveals that the majority of existing temporal taggers output TIMEX2 annotations. Since automatic conversion between TIMEX2 and TimeML annotations is not straightforward, TimeBank is of limited use for those who work specifically with TIMEX2.

3 Creating WikiWars

Given the above concerns, we were particularly interested in developing a corpus that would allow more rigorous testing of techniques for tracking time across extended narratives, since these give rise to more complex temporal phenomena than are found in simpler documents. To avoid copyright issues that might arise in the development and distribution of such a

[5] See `www.TimexPortal.info/WikiWars`.

[6] See `www.itl.nist.gov/iad/mig/tests/ace`.

corpus, we decided to use Wikipedia as a source. After considering various types of historical narrative, we settled on descriptions of the course of wars and conflicts as being particularly rich in the kinds of phenomena we wanted to explore.

3.1 Selecting Data

We queried Google with two phrases, 'most famous wars in history' and 'the biggest wars', and in each case chose the top-ranked result. One of the pages found proposed a list of the 10 most famous wars in history, and the other listed the names of the 20 biggest wars that happened in the 20th century, measured in terms of the number of military deaths. We combined the two lists, eliminated duplicates, and searched Wikipedia for articles describing these wars. Wikipedia did not contain an article for one war, and we considered two articles as inappropriate for our purposes since they did not describe the course of the wars, but rather some general information about the conflicts. This resulted in a final set of 22 articles. More details of the selection process and the URLs of the chosen Wikipedia articles are provided in the documentation distributed with the corpus.

3.2 Text Extraction and Preprocessing

To prepare the corpus, we first manually copied text from those sections of the webpages that described the course of the wars. This involved manual removal of picture captions and cross-page links. We then ran a script over the results of this extraction process to convert some Unicode characters into ASCII (ligatures, spaces, apostrophes, hyphens and other punctuation marks), and to remove citation links and a variety of other Wikipedia annotations.

Finally, we converted each of the text files into an SGML file: each document was wrapped in one `DOC` tag, inside which there are `DOCID`, `DOCTYPE` and `DATETIME` tags. The document time stamp is the date and time at which we downloaded the page from Wikipedia to our local repository. The proper content of the article is wrapped in a `TEXT` tag. This document structure intentionally follows that of the ACE 2005 and 2007 documents, so as to make the processing and evaluation of the WikiWars data highly compatible with the tools used to process the ACE corpora.

3.3 Creating Gold Standard Annotations

Having prepared the input SGML documents, we then processed them with the DANTE temporal expression tagger (see Mazur and Dale (2007)). DANTE outputs the original SGML documents augmented with an inline TIMEX2 annotation for each temporal expression found. These output files can be imported to Callisto,[7] an annotation tool that supports TIMEX2 annotations. Using a temporal expression tagger as a first-pass annotation tool not only significantly reduces the amount of human annotation effort required (creating a tag from scratch requires a number of clicks in the annotation tool), but also helps to minimize the number of errors that arise from overlooking markable expressions through 'annotator blindness'. The annotations produced by DANTE were then manually corrected in Callisto via the following process. First, Annotator 1 (the first author) corrected all the annotations produced by DANTE, both in terms of extent and the values provided for TIMEX2 attributes. This process also included the annotation of any temporal expression missed by the automatic tagger, and the removal of spurious matches. Then, Annotator 2 (the second author) checked all the revised annotations and prepared a list of errors found and doubts or queries in regard to potentially problematic annotations. Annotator 1 then verified and fixed the errors, after discussion in the case of disagreements.

The final SGML files containing inline annotations were then transformed into ACE APF XML annotation files, this being the stand-off markup format developed for ACE evaluations. This transformation was carried out using the `tern2apf` tool developed by NIST for the ACE 2004 evaluations, with some modifications introduced by us to adjust the tool to support ACE 2005 documents and to add a document ID as part of the ID of a TIMEX2 annotation (so that all annotations would have corpus-wide unique IDs).

The resulting corpus is thus available in two formats: one contains the original documents enriched with inline annotations, and the other consists of stand-off annotations in the ACE APF format.

[7] See `http://callisto.mitre.org`.

3.4 Some Deficiencies of TIMEX2

The annotation process described above revealed some issues with the use of TIMEX2 in practice. First, the flexibility of the TIMEX2 scheme, which can be at first seen as an advantage, actually makes it ambiguous. One instance of this phenomenon relates to the fact that the TIMEX2 guidelines state that the provision of some attribute values for what are called **event-based expressions** (such as *three weeks after the siege of Boston began* or *the first year of the American invasion*) is optional. Since our corpus has a significant number of such expressions, the decision as to whether or not to provide semantic values in such cases has a potentially large impact on the perceived performance of a tagger. In such cases, we decided only to provide the value when it is very clear from the article itself what the value should be.

Another area where TIMEX2 is not ideal is in regard to the annotation of time zones. First, only whole-hour time differences are supported, which eliminates some time zones (e.g. Afghanistan lies in UTC+04:30). Second, time zone information is supposed to be marked only for expressions which have it explicitly stated. However, it can often be inferred from the context that subsequent unadorned time references should inherit the same time zone as an earlier time reference.

We also found that, in a not insignificant number of cases, it is impossible to provide a precise and correct value for a temporal expression. For example, the TIMEX2 guidelines stipulate that the anchors of durations cannot have a MOD attribute, so if the anchor is *mid-August*, the value of the anchor must refer to August, which is not entirely correct as the semantics of *mid-* is lost.

TIMEX2 only supports nonspecific expressions which have explicit information about granularity. Expressions such as *a very short time* or *a short period of time* therefore cannot be provided with any value, since the context does not indicate whether the period involved should be measured in days, weeks, or months. One might consider using the typical durations of events of the corresponding types in such cases, but this solution also has problems (see (Pan et al., 2006)).

As is acknowledged in the TIMEX2 guidelines, the treatment of set expressions (i.e. recurring times and durations and frequencies, e.g. *twice a month*) is underdeveloped. One rule states that set expressions should not be anchored (Ferro et al., 2005, p. 42); this has the consequence that the full semantics of the expression *annually since 1955* cannot be provided, and the expression is therefore treated as two separate expressions, *annually* and *1955*.

Finally, alternative calendars are not supported, so an expression like *February in the pre-revolutionary Russian calendar* cannot receive a value unless it appears in an appositive construction which provides an alternative description. Similarly, consider Example (1):

(1) On *9 November 1799* (18 Brumaire of *the Year VIII*) Napoleon Bonaparte staged the coup of 18 Brumaire which installed the Consulate.

Here, *18 Brumaire of the Year VIII* is a date in an alternative calendar used in France, but we annotated only *the Year VIII* based on the trigger *year*. Note that *18 Brumaire* also occurs later in the sentence, but is not annotated.

3.5 Corpus Statistics

The corpus contains 22 documents with a total of almost 120,000 tokens[8] and 2,671 temporal expressions annotated in TIMEX2 format. In Table 1 we compare the WikiWars corpus with the other existing corpora. While the ACE 2005 Training corpus remains the largest corpus, WikiWars is larger than the ACE 2005 and 2007 evaluation corpora and the TimeBank v1.2 corpus, both in terms of number of tokens and TIMEX2 annotations. WikiWars has an order of magnitude more temporal expressions in each document, and a slightly higher density of temporal expressions than the other corpora.

Table 2 presents statistics on the individual documents that make up the corpus. The documents vary considerably in size, the smallest consisting of only 1,455 tokens, and the largest being eight times larger at 11,640 tokens. The density of TIMEX2 annotations varies from 1 in 23.1 tokens to 1 in 72.1 tokens, but for the majority of documents the ratio lies between 30 and 60.

[8] All token counts presented in Tables 1 and 2 were obtained using GATE's default English tokeniser; hyphenated words, e.g. *British-held* and *co-operation*, were treated as single tokens. For more information on GATE see (Cunningham et al., 2002).

Corpus	Docs	KB	Tokens	Temp. Expr.	$\frac{\text{Tokens}}{\text{TIMEX}}$	$\frac{\text{TIMEX}}{\text{Doc}}$
ACE05 Train.	599	1,733	318,785	5,469	58.3	9.13
ACE05 Eval.	155	350	63,217	1,154	54.8	7.45
ACE07 Eval.	254	561	104,779	2,028	51.7	7.98
WikiWars	22	631	119,468	2,671	44.7	121.41
TimeBank1.2	183	816	78,444	1,414	55.5	7.73

Table 1: Statistics of the Wikipedia War corpus compared to those of other corpora.

4 The Nature of Wikipedia Articles

Wikipedia articles may be edited by a large number of people over a significant number of revisions. We checked how often the articles constituting WikiWars were modified in the period from January 2008 to February 2010. On average, each article was changed almost 52 times per month, with the monthly number of changes for a single article ranging from 1 to 372.[9] The minimum average for an individual document was 13.08 (17_AlgerianWar), and the maximum was 171.77 (07_IraqWar).

The nature of the revision process in Wikipedia leads to some artefacts that may be not typical of other document sources, such as news, where the text is usually carefully prepared by its author and checked by an editor. This is not to say that Wikipedia content is necessarily of low quality; this is an encyclopedia with many people and bots controlling its quality, and there exist manuals of style for authors to help them avoid errors and ambiguity and to ensure maximum consistency.[10] However, given the large number of editors with various degrees of fluency and experience in writing and editing, it would not be surprising if some parts of the texts are not perfect. In the process of preparing the gold standard annotations for the WikiWars corpus, we have made the following observations.

[9]Note that these numbers are for the articles as a whole, and not just the sections which we extracted (although these are usually the major part of the article). Additionally, these edits include both major changes (e.g. adding a new section), minor changes (e.g. correcting a grammar error or adding a comma), vandalism (deletion of the page content or the on-purpose provision of false information) and restoring the page after an act of vandalism has been detected.

[10]See, for example, the manual of style concerning formating dates and numbers, located at `http://en.wikipedia.org/wiki/Wikipedia:DATE`.

Document ID	Tokens	TIMEX2	$\frac{\text{Tokens}}{\text{TIMEX2}}$
01_WW2	5,593	169	33.1
02_WW1	10,370	264	39.3
03_AmCivWar	3,529	75	47.1
04_AmRevWar	5,695	146	39.0
05_VietnamWar	11,640	243	47.9
06_KoreanWar	5,992	147	40.8
07_IraqWar	8,404	247	34.0
08_FrenchRev	9,631	174	55.4
09_GrecoPersian	7,393	129	57.3
10_PunicWars	3,475	57	61.0
11_ChineseCivWar	3,905	103	37.9
12_IranIraq	4,508	98	46.0
13_RussianCivWar	3,924	103	38.1
14_FirstIndochinaWar	3,085	70	44.1
15_MexicanRev	3,910	77	50.8
16_SpanishCivilWar	1,455	63	23.1
17_AlgerianWar	7,716	130	59.4
18_SovietsInAfghanistan	5,306	110	48.2
19_RussoJap	2,760	62	44.5
20_PolishSoviet	5,137	106	48.5
21_NigerianCivilWar	2,091	29	72.1
22_2ndItaloAbyssinianWar	3,949	69	57.2
Total for the whole corpus	119,468	2,671	44.7
Average per document	5,430	121	–
Standard deviation	2,663	63	–

Table 2: Statistics of the Wikipedia War corpus.

4.1 Broken Narratives

In some articles we have found situations where a sentence does not appear to cohere with those on either side of it. This may be the result of a number of modifications made by different authors, or it may be due to a lack of writing skill on the part of the person who wrote the paragraph in question. Example (2) below provides an example of this phenomenon: the sentence about de Gaulle being elected president contains a temporal expression which progresses the temporal focus in the narrative to 1959, but the later context of the article strongly suggests that the subsequent reference to *October* is in fact October 1958.

(2) ALN commandos committed numerous acts of sabotage in France in *August*$_{[1958]}$, and the FLN mounted a desperate campaign of terror in Algeria to intimidate Muslims into boycotting the referendum. Despite threats of reprisal, however, 80 percent of the Muslim electorate turned out to vote in *September*$_{[1958]}$, and of these 96 percent approved the constitution. In *February 1959*, de Gaulle was elected president of the new Fifth Republic. He visited Constantine in

October$_{[1958]}$ to announce a program to end the war and create an Algeria closely linked to France.

It would appear that the reference to *February 1959* is a later addition to the text which has been made without the surrounding text being appropriately revised to accommodate this change. Clearly such instances of incoherence will cause problems for any process that attempts to track the temporal focus.

4.2 Ambiguous Writing

We have also found cases of a lack of precision in writing, which leads to ambiguous statements. Consider the following example:

(3) The Afghan government, having secured a treaty in *December 1978* that allowed them to call on Soviet forces, repeatedly requested the introduction of troops in Afghanistan in *the spring* and *summer of 1979*. They requested Soviet troops to provide security and to assist in the fight against the mujahideen rebels. On *April 14, 1979*, the Afghan government requested that the USSR send 15 to 20 helicopters with their crews to Afghanistan, and on *June 16*, the Soviet government responded and sent a detachment of tanks, BMPs, and crews to guard the government in Kabul and to secure the Bagram and Shindand airfields. In response to this request, an airborne battalion, commanded by Lieutenant Colonel A. Lomakin, arrived at the Bagram Air Base on *July 7*. [...]

After *a month*, the Afghan requests were no longer for individual crews and subunits, but for regiments and larger units. In *July*, the Afghan government requested that two motorized rifle divisions be sent to Afghanistan. *The following day*, they requested an airborne division in addition to the earlier requests.

Here, in the first paragraph there are four temporal expressions related to the Afghan government asking for troops and equipment. There is also one date related to the Soviets' reply to these requests and sending of tanks, and one date related to the arrival of an airborne battalion. The second paragraph starts with *after a month*; the first possible interpretation is that this is a month after the 7th July mentioned in the previous paragraph; i.e. the month would end on the 6th of August. But the following sentence reveals that this is not the case, as it mentions some requests for larger units that were made in July. Usually a narrative progresses forwards in time, not backwards, so the month must start either on 14th April or 16th June: if the second sentence elaborates the first one, then it is a month from 16th June; if it just mentions one of the requests for larger units, then it is probably a month from 14th April.

It is also unclear whether the second paragraph talks about the same request for airborne forces which was mentioned in the first paragraph: both these events are dated July. The phrase *In response to this request* is in fact placed very oddly, as its preceding sentence does not mention any request, but rather talks about the Soviets' response to requests. This may suggest that what at first looks just like a careless and ambiguous use of the expression *after a month* is in fact a larger problem of lack of coherency in these two paragraphs.

4.3 Use of Deictic Expressions

One of the articles, `07_IraqWar`, contained a number of deictic temporal expressions, indicative of the fact that the events described were happening contemporaneously to the time of writing (as is often the case in news stories); for example:

(4) a. Democrats plan to push legislation *this spring* that would force the Iraqi government to spend its own surplus to rebuild.

b. A protester said that despite the approval of the Interim Security pact, the Iraqi people would break it in a referendum *next year*.

Obviously, after some time these expressions will no longer make sense, since there is no 'at-the-time-of-writing' time stamp associated with these sentences: for the reader of a Wikipedia article, the reference date is the time of reading. In the case of the above example, these sentences were written in April and December 2008, respectively.[11] Arguably, these sentences should be corrected, making the temporal expressions fully-specified (e.g. *in spring of 2009* and *in 2009*), or context-dependent (e.g. *in spring of that year* and *the following year*) if there is a context in the article which supports their correct interpretation. Of course, not only the temporal expressions need to be revised, but also the tense and aspect of the verbs used in the sentences. In the gold standard annotations, however, we provided the values by interpreting these expressions with respect to the document time stamp (i.e. `2010-SP` and `2010`), as the text itself does not provide any evidence that other dates were intended.

[11] Somewhat laborious document archaeology allows this information to be extracted from Wikipedia's archive.

Pos	Count	Token class or lexical form	Pos	Count	Token class or lexical form	Pos	Count	Token class or lexical form
1	4650	NUMBER_DIGIT_2	18	222	today	35	49	AMPM
2	1942	:	19	202	NUMBER_DIGIT_1	36	48	ORDINAL_DIGIT
3	1499	-	20	191	last	37	48	?
4	1329	NUMBER_DIGIT_4	21	171	WEEKDAYNAME_ABBR	38	45	recently
5	828	ARTICLE	22	145	NUMBER_DIGIT_8	39	43	year-old
6	765	TEMPORALUNIT	23	113	ago	40	42	later
7	634	TEMPORALUNIT_PLURAL	24	108	former	41	41	tonight
8	555	PREPOSITION	25	96	time	42	39	christmas
9	528	now	26	79	right	43	36	tomorrow
10	411	t	27	69	new	44	36	current
11	403	WEEKDAYNAME	28	69	future	45	35	couple
12	335	NUMBER_WORD	29	67	gmt	46	34	recent
13	329	MONTHNAME	30	65	next	47	33	earlier
14	242	MONTHNAME_ABBR	31	63	past	48	32	and
15	240	DAYPART	32	61	yesterday	49	31	early
16	233	DEMONSTRATIVE	33	59	few	50	31	DIRECT_FREQ
17	224	,	34	50	every	51	31	's

Table 3: The most frequent tokens in TEs in the ACE 2005 Training corpus.

Pos	Count	Token class or lexical form	Pos	Count	Token class or lexical form	Pos	Count	Token class or lexical form
1	1181	MONTHNAME	18	59	:	35	13	first
2	1157	NUMBER_DIGIT_4	19	51	end	36	11	future
3	674	NUMBER_DIGIT_2	20	49	-	37	11	earlier
4	490	ARTICLE	21	47	late	38	11	.
5	288	PREPOSITION	22	37	DAYPART	39	11	's
6	221	NUMBER_DIGIT_1	23	36	later	40	9	previous
7	211	TEMPORALUNIT	24	36	former	41	9	christmas
8	206	TEMPORALUNIT_PLURAL	25	32	next	42	8	last
9	165	,	26	27	same	43	8	AMPM
10	133	NUMBER_WORD	27	25	period	44	7	battle
11	99	SEASON	28	22	t	45	7	DIRECT_FREQ
12	98	NUMBER_DIGIT_3	29	20	mid-	46	6	short
13	82	bc	30	18	war	47	6	several
14	76	now	31	18	few	48	6	season
15	70	time	32	14	following	49	6	recent
16	67	early	33	14	ORDINAL_DIGIT	50	6	past
17	63	DEMONSTRATIVE	34	13	s	51	6	"

Table 4: The most frequent tokens in TEs in the WikiWars corpus.

4.4 Use of Time Zone Information

Consider the following example, which comes from the article `01_WW2`:

(5) On *December 7 (December 8 in Asian time zones), 1941*, Japan attacked British and American holdings with near simultaneous offensives against Southeast Asia and the Central Pacific.

The italicized temporal expression is difficult to detect, and it is not clear how it should be annotated. But it is also imprecise with respect to which time zone is intended: Asia encompasses 10 time zones. Therefore it is impossible to fully interpret the expression. Note also that the expression combines a time zone with a date, rather than with a time. While uncommon, this is not incorrect; but the TIMEX2 guidelines do not explicitly allow for this circumstance.

4.5 Quotes Missing a Time Stamp

Occasionally it happens that an article contains a quoted utterance, but there is no indication of when the utterance was made. For example, in the document `05_VietnamWar` we find the following:

(6) Nixon said in an announcement, "I am *tonight* announcing plans for the withdrawal of an additional 150,000 American troops to be completed during *the*

spring of next year. This will bring a total reduction of 265,500 men in our armed forces in Vietnam below the level that existed when we took office *15 months ago*."

It is impossible to determine what dates are meant by the three temporal expressions present in the announcement. In some cases this information may be provided in citation footnotes, but this is not always the case; when this is absent, such expressions can only be annotated at the level of textual extent and a localised, context-dependent semantics.

5 Comparing WikiWars to the ACE Data

A comparison of WikiWars with the ACE corpora reveals some interesting differences.

5.1 Vocabulary Differences

First, we found differences on the level of the lexical triggers that signal the presence of temporal expressions. Because of space limitations, we provide here only the main findings.

Tables 3 and 4 present the 51 most frequent tokens, including punctuation, in the ACE 2005 Training and WikiWars corpus, respectively. Some tokens are combined into what we call **trigger classes**; for example, all weekday names belong to the class `WEEKDAYNAME`.[12]

We can see that there are many classes that fall into the top 51 positions for both corpora, e.g. the names of temporal units (such as *month* and *year*). But there are also clear differences. Month names are the most frequent class in WikiWars, while they are not so frequent in ACE. Similarly, year seasons ranked very highly in WikiWars, but do not figure in the rankings shown for ACE. On the other hand, weekday names are quite frequent in the ACE corpus, but do not occur in the table for WikiWars. This suggests that these corpora make different use of temporal expressions: in WikiWars we find many references to the more distant past, thus the high use of month names, but ACE documents tend to discuss temporally local issues, so they are more likely to refer to days in the weeks preceding and following the reference date.

Looking at individual tokens, we can see that deictic expressions such as *today*, *tonight*, *yesterday* and *tomorrow* are in the top 51 positions for ACE, but almost never occur in WikiWars: there are only three instances of *today*, two of *tomorrow* and one of *tonight* in the corpus, and all of these appear only in quoted speech. Similarly, *ago* occurred 113 times in ACE, but only twice in WikiWars: once in quoted speech, and once used incorrectly instead of *earlier* in a context-dependent expression. Other tokens which are frequent in ACE but rare in WikiWars are *recent*, *recently*, *current* and *currently*.

5.2 Temporal Discourse Structure

A more interesting property that WikiWars exhibits, and which is noticeably absent from the simpler ACE data, is what we might think of as a discourse mechanism for resetting the temporal focus. This is a feature of complex texts in general, rather than something that is specific to Wikipedia as a source. In these cases, the discourse does not follow a single global timeline from the beginning to the end of the document, but is rather divided into subdiscourses which describe separate chains of events that often have common temporal starting points. This is typical in the description of big, often international, conflicts, where one can distinguish several theaters of the war, i.e. the eastern and western theaters.

In most cases the switch to a different 'part of the story' can be determined not only by analysing the events and their geographic locations, but by recognizing that the first date appearing in the new subdiscourse is generally fully specified. This is, however, not always the case, as shown in the following example extracted from the article `01_WW2`:

(7) In northern Serbia, the Red Army, with limited support from Bulgarian forces, assisted the partisans in a joint liberation of the capital city of Belgrade on *October 20*$_{[1944]}$. *A few days later*, the Soviets launched a massive assault against German-occupied Hungary that lasted until the fall of Budapest in *February 1945*. [...]

By *the start of July*$_{[1944]}$, Commonwealth forces in Southeast Asia had repelled the Japanese sieges in Assam, pushing the Japanese back to the Chindwin River while the Chinese captured Myitkyina. In China, the Japanese were having greater successes, having fi-

[12] The entries in the table correspond to the lexical and punctuation clues that drive detection of temporal expressions: the high rank of colons and dashes comes from their use in document time stamps, which are considered markable by the TIMEX2 guidelines. The *T* token is a separator that often occurs in timestamps, e.g. *2005-01-25T11:08:00*; the question mark appears very often because some of the ACE timestamps are of the form *????-??-??T19:33:00*.

nally captured Changsha in *mid-June*$_{[1944]}$ and the city of Hengyang by *early August*$_{[1944]}$. Soon after, they [...] by *the end of November*$_{[1944]}$ and successfully linking up their forces in China and Indochina by *the middle of December*$_{[1944]}$.

Clearly, quite sophisticated processing is required to handle this phenomenon adequately.

6 Automated Processing of WikiWars

After we developed the WikiWars corpus, we used it to evaluate our temporal expression tagger, DANTE, which had been developed for participation in ACE. Performance at finding temporal expressions in text is traditionally reported, for example by (Mani and Wilson, 2000; Negri and Marseglia, 2005; Teissèdre et al., 2010), in terms of precision, recall and F-measure. These can, however, be calculated in two ways, **lenient** and **strict**, corresponding to two tasks: **detection** (where a single character overlap between the gold standard and system annotation counts as a correct answer) and **recognition** (where an exact overlap is required).

Table 5 shows our tagger's initial performance on the data. While the lenient F-measure for extent recognition was comparable to that obtained for the ACE 2005 Training corpus (0.82 vs 0.78), the recall was much lower: 0.75 vs 0.87. The difference in strict results was even larger, where both precision and recall were lower for WikiWars than for ACE, resulting in an F-measure of 0.38. When evaluating also the VAL attribute, the strict F-measure was quite low for both corpora, but significantly lower for WikiWars: 0.17 vs 0.33. This illustrates how illusive it may be to trust the performance of a tagger measured on a single, possibly biased, data set.

In the light of the results of our comparison in Section 5, it is clear that at some of the performance loss here is simply due to domain differences with respect to lexical triggers. So, we extended DANTE's coverage with approximately 20 temporal triggers and modifiers to include the more common vocabulary that appeared in the WikiWars data; we also modified the recognition grammar to reduce the number of spurious matches and extent errors. These changes resulted in the improvements shown in Table 6. The performance on extent recognition improves significantly for both sets of data, but the gap between extent recognition and evaluation of the VAL attribute is much larger on WikiWars. This is most likely because the strategy of using the document time stamp for the interpretation of context-dependent expressions does not work at all for WikiWars documents, whereas it works well for ACE documents, in line with our earlier comments in regard to the genres of the documents. This emphasises the need to develop sophisticated methods for temporal focus tracking if we are to extend current time-stamping technologies beyond the relatively simplistic temporal structures found in currently available corpora.

	Lenient			Strict		
Corpus and Task	Prec	Rec	F	Prec	Rec	F
WW - Extent only	0.90	0.75	0.82	0.42	0.35	0.38
WW - Extent + VAL	0.22	0.18	0.20	0.19	0.16	0.17
ACE - Extent only	0.71	0.87	0.78	0.53	0.65	0.58
ACE - Extent +VAL	0.34	0.42	0.37	0.30	0.36	0.33

Table 5: Initial performance of DANTE on WikiWars and the ACE 2005 Training corpus.

	Lenient			Strict		
Corpus and Task	Prec	Rec	F	Prec	Rec	F
WW - Extent only	0.98	0.99	0.99	0.95	0.95	0.95
WW - Extent + VAL	0.59	0.60	0.59	0.58	0.59	0.58
ACE - Extent only	0.88	0.93	0.90	0.75	0.79	0.77
ACE - Extent +VAL	0.63	0.67	0.65	0.57	0.60	0.58

Table 6: Current performance of DANTE on WikiWars and the ACE 2005 Training corpus.

7 Conclusions and Future Work

We have presented a new corpus based on the historical descriptions of 22 wars sourced from English Wikipedia, and we have described in detail the methodology adopted to construct the corpus; the corpus can be easily extended in the same way. We annotated temporal expressions in these documents with TIMEX2 tags, which provide both the textual extents and the semantics of the expressions in the context of whole article.

Following an analysis of the differences between our new corpus and existing data sets, we then presented the results of automatic processing of the corpus. This demonstrates that differences in the vocabulary used for temporal expressions can be fairly straightforwardly incorporated in a tagging tool, but that appropriate processing of temporal structure in complex documents requires more sophisticated techniques than those required to handle existing corpora. The WikiWars Corpus provides data that tests these capabilities.

References

David Ahn, Sisay Fissaha Adafre, and Maarten de Rijke. 2005. Recognizing and Interpreting Temporal Expressions in Open Domain Texts. In *We Will Show Them: Essays in Honour of Dov Gabbay, Vol 1*, pages 31–50, October.

David Ahn, Joris van Rantwijk, and Maarten de Rijke. 2007. A cascaded machine learning approach to interpreting temporal expressions. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, Rochester, NY, USA, April.

Jennifer Baldwin. 2002. Learning Temporal Annotation of French News. Master's thesis, Dept. of Linguistics, Georgetown University, April.

Branimir Boguraev, Jose Castaño, Rob Gaizauskas, Bob Ingria, Graham Katz, Bob Knippen, Jessica Littman, Inderjeet Mani, James Pustejovsky, Antonio Sanfilippo, Andrew See, Andrea Setzer, Roser Saurí, Amber Stubbs, Beth Sundheim, Svetlana Symonenko, and Marc Verhagen. 2005. TimeML 1.2.1 – A Formal Specification Language for Events and Temporal Expressions, October.

Branimir Boguraev, James Pustejovsky, Rie Ando, and Marc Verhagen. 2007. TimeBank evolution as a community resource for TimeML parsing. *Language Resources and Evaluation*, 41(1):91–115, 02.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the ACL*.

Lisa Ferro, L. Gerber, I. Mani, B. Sundheim, and G. Wilson. 2005. TIDES 2005 Standard for the Annotation of Temporal Expressions. Technical report, MITRE, September.

Kadri Hacioglu, Ying Chen, and Benjamin Douglas. 2005. Automatic time expression labeling for english and chinese text. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 6th International Conference, CICLing'05*, Lecture Notes in Computer Science, pages 548–559, Mexico City, Mexico, February. Springer.

Benjamin Han, Donna Gates, and Lori Levin. 2006. From language to time: A temporal expression anchorer. In *Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning (TIME'06)*, pages 196–203. IEEE Computer Society, June.

Inderjeet Mani and George Wilson. 2000. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL '00)*, pages 69–76, Morristown, NJ, USA, October. Association for Computational Linguistics.

Pawel Mazur and Robert Dale. 2007. The DANTE Temporal Expression Tagger. In Zygmunt Vetulani, editor, *Proceedings of the 3rd Language And Technology Conference (LTC)*, Poznan, Poland, October.

Pawel Mazur and Robert Dale. 2008. What's the Date? High Accuracy Interpretation of Weekday Names. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 553–560, Manchester, UK, August. Coling 2008 Organizing Committee.

Matteo Negri and Luca Marseglia. 2005. Recognition and normalization of time expressions: Itc-irst at tern 2004. Technical Report WP3.7, Information Society Technologies, February.

Feng Pan, R. Mulkar, and J. R. Hobbs. 2006. Learning event durations from event descriptions. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 393–400, Sydney, Australia, July. Association for Computational Linguistics.

James Pustejovsky, J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *IWCS-5, Fifth International Workshop on Computational Semantics*, Tilburg, The Netherlands, January.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Mike Rosner Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Estela Saquete. 2005. *Temporal Expression Recognition and Resolution applied to Event Ordering*. Ph.D. thesis, Departamento de Lenguages y Sistemas Informaticos, Universidad de Alicante, June.

Frank Schilder. 2004. Extracting meaning from temporal nouns and temporal prepositions. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(1):33–50, March.

Charles Teissèdre, Delphine Battistelli, and Jean-Luc Minel. 2010. Resources for calendar expressions semantic tagging and temporal navigation through texts. In *Proceedings of LREC2010*, May.