ACCURAT

Analysis and Evaluation of Comparable Corpora
for Under Resourced Areas of Machine Translation

www.accurat-project.eu

Project no. 248347

Deliverable D3.4
Report on methods for collection of comparable corpora

Version No. 1.0

31/10/2011

Document Information

Deliverable number:	D3.4
Deliverable title:	Report on methods for collection of comparable corpora
Due date of deliverable:	31/10/2011
Actual submission date of deliverable:	31/10/2011
Main Author(s):	Monica Paramita, Ahmet Aker, Rob Gaizauskas, Paul Clough, Emma Barker, Nikos Mastropavlos, Dan Tufis
Participants:	USFD, RACAI, ILSP, TILDE
Internal reviewer:	CTS
Workpackage:	WP3
Workpackage title:	Methods and techniques for building a comparable corpus from the Web
Workpackage leader:	USFD
Dissemination Level:	**PU**: Public
Version:	V1.0
Keywords:	Retrieval methods, evaluation, sentence alignment

History of Versions

Version	**Date**	**Status**	**Name of the Author (Partner)**	**Contributions**	**Description/ Approval Level**
V0.1	01/09/2011	Draft	USFD	Skeleton	
V0.1	20/10/2011	Draft	USFD, ILSP, RACAI	Contributions	Edited intro, additional sections
V0.2	26/10/2011	Draft	USFD, ILSP, RACAI, CTS	Contributions	Edited main sections based on internal review
V1.0	28/10/2011	Final	USFD, ILSP, RACAI, CTS, Tilde	Revisions	First release, Submitted to PO

EXECUTIVE SUMMARY

The Web contains comparable documents in different languages which may be useful for machine translation. Effective retrieval methods are needed to correctly identify and gather these documents. This document describes the methods developed to retrieve comparable documents from different Web sources to build comparable corpora for general and narrow domains.

Table of Contents

Abbreviations

Abbreviation	Term/definition
FIFO	First-In First-Out
HTML	HyperText Markup Language
JWPL	Java Wikipedia Library
SMT	Statistical Machine Translation
MCC	Multilingual comparable corpus
MT	Machine Translation
URL	Uniform Resource Locator
UTF-8	Unicode Transformation Format

1. Introduction

The Web contains comparable documents in different languages which may be useful for machine translation. Effective retrieval methods are needed to correctly identify and gather these documents. Previous work in this area has been reviewed and reported in D3.3, which also contains preliminary research into various retrieval methods. A number of iterations were performed internally to evaluate the comparable documents gathered and to improve the retrieval methods. This evaluation framework has since then been improved and a final evaluation performed project-wide to validate the results.

The ACCURAT project aims to build two types of corpora: 1) general usage corpora for under-resourced languages, and 2) corpora for narrow domains. This report starts by describing the work performed in developing methods to gather documents from the Web to build a general usage corpora for under-resourced languages. We specifically focus on two sources of documents found on the Web: news stories and Wikipedia articles. Methods developed to retrieve these documents are described in Section 2. In Section 3, we describe methods to retrieve documents from narrow domains. A pool of retrieval documents were evaluated by human assessors to measure the performance of the developed retrieval methods. The evaluation framework and the results are described in Section 4. The conclusions of our work are described in Section 5.

2. Retrieval Techniques for General Usage Corpora

The first type of corpora collected in this project is general usage corpora. These corpora contain comparable documents for all ACCURAT under-resourced languages (Greek, Estonian, Croatian, Latvian, Lithuanian, Romanian and Slovenian), English and German . We specifically aimed to retrieve comparable documents from news sites and Wikipedia. The work performed for each source is described in more detail below.

2.1. News

One type of text which occurs in virtually all languages in large volumes, and in which one finds similar content being expressed across languages, is news text. We live in a highly interconnected world and significant events taking place in any part of the world are likely to be reported in major newspapers everywhere at more or less the same time. While comments on events may differ, we expect basic factual reporting to convey the same message everywhere: Michael Jackson has died; there has been a tsunami in Japan, etc.

That news texts should be rich sources of shared content, and hence are a type of comparable corpus of high potential value for SMT, has not gone unnoticed. Various researchers have proposed techniques for gathering news stories about the same events and then for finding and extracting shared content in them for use in SMT systems. Munteanu and Marcu (2005), for example, propose a technique for finding parallel sentences within large corpora of comparable news texts and in Munteanu and Marcu (2006) they go further and discuss techniques for extracting parallel sub-sentential fragments from comparable news texts. There has been a long tradition of work on bilingual lexicon extraction from comparable corpora ranging from (Fung, 1998) to (Li and Gaussier, 2011), much of which uses corpora of news texts. Comparable corpora have also been used in support of cross-language information retrieval (Braschler and Schauble, 1998).

To the best of our knowledge none of this prior work has resulted in public domain tools for gathering comparable news corpora. Hence, as one of our first goals in ACCURAT, we set out to build such tools. Furthermore, previous work, with the exception of Braschler and Schauble (1998), has not paid much attention to developing tools that aim to maximize the amount of shared content in the retrieved text pairs, instead focusing on techniques for extracting parallel material from whatever has been retrieved. Our tools, by contrast, are specifically designed to retrieve highly similar news texts. To gather news documents, we used news articles retrieved by Google and employed different retrieval techniques, and rules of thumb, to obtain potential documents useful for SMT.

In order to collect comparable corpora we first collect News titles along with publishing time and article urls through Google News Search and RSS News feeds. Please note that we only download current News articles and do not search for articles in News archives or in the entire Web. Searching in a bigger space causes more noise in the pairing process than when the focus is only on the current News. The number of different events in the current News (e.g. within one week period) is smaller compared to a bigger space which in turn causes less noise in the pairing process.

2.1.1. Document Crawling

We first collect an *initial corpus* of titles from News article mono-lingually using Google News. For each language that has news entries in Google News we iteratively download titles from news in different topic categories, such as *economics, world, politics*, etc. We set the iteration time to 15 minutes. Apart from the title for each search result we also have information about the *date & time* of publication and the *URL* to the actual article.

We make use of the Google News clustering of News articles that are found to be similar to each other and for each title in our *initial corpus* we collect titles of articles in that cluster as well. We refer to this corpus as *news corpus 1*.

We then use the titles from the *initial corpus* and *news corpus 1* as queries and perform a monolingual Google News search. We extract the titles from the search results and these constitute the *news corpus 2*.We further expand the collection of article titles by a fourth corpus, *news corpus 3*. For this, we take the article titles from the *initial corpus, news corpus 1* and *news corpus 2* for the English collection only. We parse them for named entities such as *person, location* and *organization* names. For each named entity type we do the following: we translate the entities into the language in which the search will be performed (using Google Translate) and perform a Google News Search using the translated entity as a query.

We also manually identify a good number of RSS News feeds for each language from which we extract similar information as in the Google News Search. We refer to this corpus as *RSSFeed corpus*. This step is especially needed for languages such as Lithuanian for which Google News Search does not provide news search.

2.1.2. Document Alignment

In the alignment phase, the goal is to pair the articles from the articles sets (*initial corpora, news corpus 1, news corpus 2, news corpus 3 and RSSFeed corpus*) in order to build comparable corpus. We use English as the source language and any one from the remaining languages as the target language. We pair the source articles with the target ones leading to eight comparable corpus and for each language pair such as English-Greek, we obtain comparable corpus. The alignment rules of thumb we investigate are the following:

- DateSim: We align documents published within seven days of each other. We score each pair by their publishing date difference 1/ (d+1), where d is the date difference, with d = [0; 7]). Articles published on the same date get a score of 1.
- TimeSim: We align documents published in the same day. Each pair is scored by 1/ (h+1), where h is the time difference in hours, with h = [0; 23]. Articles published within the same hour get a score of 1.
- TitleLengthDif : We remove stop words from both article titles using OpenNLP and measure the difference in their lengths (based on number of words). The target language title is first translated into English using Google. We score each pair by 1/ (w+1), where w is the difference in words count (starting from 0). For this feature we also ensure that both titles (after removing the stop words) have at least 5 words. Article titles with the same length get a score of 1.
- TitleSim: We compute the cosine similarity over the titles. Each pair is scored between 0 and 1. For this feature we also ensure that both titles (after removing the stop words) have at least 5 words. If the titles do not have at least 5 words they are not considered for further process.
- All : We use all previous scoring functions combined. We consider the scores from each one of the previous scoring functions equally and sum their scores leading to a maximum score of 4 for this technique.

2.2. Wikipedia

As an online encyclopedia, Wikipedia is another promising source of comparable documents in the Web. Documents in Wikipedia are constantly developed and written in many languages. Documents which contain information on the same topic are also linked to each other, which enables us to filter out documents on different topics very easily. We investigated the degree to which Wikipedia articles about the same topic but written in different languages are comparable. Our findings show that even though these documents contain the same topic, not all documents contain comparable text segments. This is caused by differences how a topic is covered or viewed, or the way the information is presented. Including these kinds of documents in comparable corpus may introduce noise and lead to reduced MT performance. Retrieval methods are therefore needed to identify and gather documents which are comparable or contain comparable segments. In this section we described methods developed to retrieve comparable documents in Wikipedia.

2.2.1. Anchor-Based Method

In the first method, we focused our analysis on retrieving comparable Wikipedia articles by specifically looking for comparable (or parallel) segments within the documents. To enable this process to be applicable to all under-resourced languages, we aimed to use features requiring no linguistic resources. We used a technique modified from Adafre & Rijke (2007), which uses information about the anchor texts in Wikipedia articles on the same topic to find parallel sentences in Wikipedia documents.

We first collected all documents which are connected by interlanguage links from Wikipedia dump in March 2010. These documents are used as the data set as they contain the same topic and therefore have higher probability to contain comparable segments than documents in different topic. Due to the same reason, we ignored other documents which were not paired by

Wikipedia. We used Java Wikipedia Library (JWPL)[1] to access the full content of all paired documents. The process involved in this phase is shown in detail in Figure 1.

[1] http://www.ukp.tu-darmstadt.de/software/jwpl/

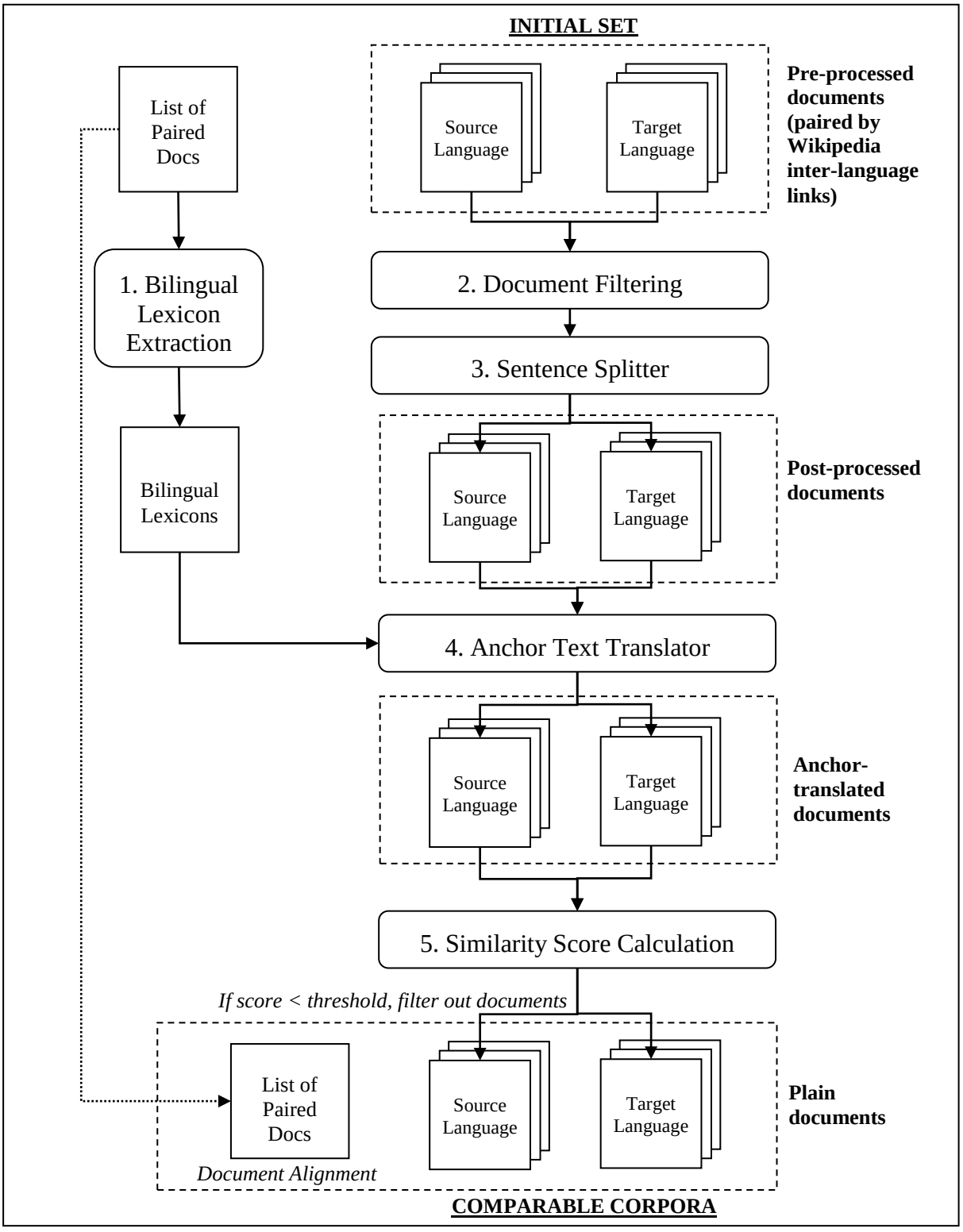

Figure 1 Proposed approach to collect comparable documents from Wikipedia

This retrieval method contains 5 main processes:

1. *Bilingual Lexicon Extraction*
   We build a bilingual lexicon by extracting document titles from the list of all paired documents in Wikipedia. This lexicon will not contain translations of all possible words in the dictionary; however, it will contain important terms, such as named entities. This process is an essential phase in this method as the result will eventually be used as a "dictionary" for the translation of anchor texts. Furthermore, no other linguistic resource will then be needed for the translation process.

2. *Document Filtering*
   Documents from Wikipedia contain formatting information which is not needed in our process and other information, such as image captions, tables, info boxes, etc. An example of this is shown in Figure 2.

```
{{atsauces+}}
{{Warbox
|conflict=Desmitdienu karš <br>([[Dienvidslāvijas kari]])
|campaign=
|colour_scheme=background:#bbcccc
|image=[[Attēls:Radenci, July 28, 1991.jpg|300px]]
|caption= Sadursmes vieta Radencos 28. jūnijā.
|date=[[27. jūnijs]] - [[6. jūlijs]], [[1991]]
|place=[[Slovēnija]]
|result=Slovēnijas uzvara
|combatant1=[[Image:Flag_of_Slovenia.svg|20px]] [[Slovēnija]], Zemssardze
|combatant2=[[Image:Flag of SFR Yugoslavia.svg|20px]] [[Dienvidslāvija]],
[[Dienvidslāvijas tautas armija]]
|commander1=[[Janezs Janša]]
|commander2=[[Veljko Kadijevičs]]
|strength1=16,000 zemessardze, 10,000 policijas spēki
|strength2=35,200 Dienvidslāvijas Tautas armija
|casualties1=18 nogalināti,<br> 182 ievainoti<br> (oficiālā informācija)
|casualties2=44 nogalināti,<br> 146 ievainoti<br> 5,000 saņemti gūstā<br> (pēc
slovēņu ziņām)
|}}'''Desmitdienu karš''', arī '''Slovēnijas neatkarības karš''' ([[slovēņu
valoda|slovēņu]] - ''Slovenska osamosvojitvena vojna'' (''Slovēnijas
neatkarības karš''), [[serbu valoda|serbu]] - ''Рат у Словенији'' (''Karš
Slovēnijā'')) - bruņots konflikts no 1991. gada 27. jūnija līdz 6. jūlijam
starp [[Slovēnija|Slovēniju]] un [[Dienvidslāvija|Dienvidslāviju]], kura
rezultātā nodibinājās neatkarīgā Slovēnijas valsts.
Desmitdienu karš bija iesākums daudz asiņainākiem kariem citās Dienvidslāvijas
republikās. Tas bija arī gandrīz vienīgais no Dienvidslāvijas kariem, pēc kura
neviena no pusēm netika apsūdzēta [[kara noziegumi|kara noziegumos]].
==Priekšvēsture==
...
```

Figure 2: An example pre-processed Latvian document

Since we are interested in eventually finding parallel, or comparable, sentences, we disregard all of this content and include only information (sentences) in paragraphs or lists. After the filtering step, the resulting documents which contain sentences and anchor information only (shown in [[anchors texts]]), as shown in Figure 3.

```
'''Desmitdienu karš''', arī '''Slovēnijas neatkarības karš''' ([[slovēņu valoda|slovēņu]] - ''Slovenska osamosvojitvena vojna'' (''Slovēnijas neatkarības karš''), [[serbu valoda|serbu]] - ''Рат у Словенији'' (''Karš Slovēnijā'')) - bruņots konflikts no 1991. gada 27. jūnija līdz 6. jūlijam starp [[Slovēnija|Slovēniju]] un [[Dienvidslāvija|Dienvidslāviju]], kura rezultātā nodibinājās neatkarīgā Slovēnijas valsts.
Desmitdienu karš bija iesākums daudz asiņainākiem kariem citās Dienvidslāvijas republikās. Tas bija arī gandrīz vienīgais no Dienvidslāvijas kariem, pēc kura neviena no pusēm netika apsūdzēta [[kara noziegumi|kara noziegumos]].
==Priekšvēsture==
...
```

Figure 3: Post-processed Latvian document

3. *Sentence Splitter*
   As described previously, this retrieval method aims to retrieve comparable documents which contain parallel or comparable sentences. Therefore, each document needs to be split into sentences to enable further analysis. This process will result in documents which contain one sentence per line.

4. *Anchor Text Translator*
   In this step, we make use of the extracted bilingual lexicon to translate all anchor texts from the source language into the target language. After replacing all the anchor texts in the previous example into English, we obtain the document shown in Figure 4 (in which all the replaced anchors are shown in bold). When an anchor text is not available in the bilingual lexicon, no translation will be performed and the original text will be used in the document.

```
'''Desmitdienu karš''', arī '''Slovēnijas neatkarības karš''' ([[slovene language]] - ''Slovenska osamosvojitvena vojna'' (''Slovēnijas neatkarības karš''), [[serbian language]] - ''Рат у Словенији'' (''Karš Slovēnijā'')) - bruņots konflikts no 1991. gada 27. jūnija līdz 6. jūlijam starp [[slovenia]] un [[yugoslavia]], kura rezultātā nodibinājās neatkarīgā Slovēnijas valsts.
Desmitdienu karš bija iesākums daudz asiņainākiem kariem citās Dienvidslāvijas republikās.
Tas bija arī gandrīz vienīgais no Dienvidslāvijas kariem, pēc kura neviena no pusēm netika apsūdzēta [[war crime]].
==Priekšvēsture==.
...
```

Figure 4: Documents translated based on using the anchor texts

5. *Calculation of Similarity Score*
   In this step, we calculate the similarity score of the document pairs by pairing sentences which contain the highest word overlap. No other translation is performed in the text.

Therefore, sentences are paired if they share the same anchor texts. Moreover, sentences which share overlapping words, such as named entities or numbers are also taken into account. We specified several filters for the sentences:

- Sentences which contain a large proportion of capital words are ignored.
- Sentences which contain a large proportion of numbers are ignored.
- Sentences which contain fewer than 3 words are ignored.

For each sentence from the smaller document of the pairs (regardless of the language), we aimed to find the best matching sentence from the bigger documents. Therefore, the maximum paired sentences for that pair is the number of sentences of the smaller documents. The comparability score for a document pair is computed as the average score of all paired sentences:

$$comparability\ score = \frac{\sum_{i=0}^{n} S_i}{n}$$

where S_i represents the Jaccard similarity score of word overlap between sentence *i* and the best matching sentence (or 0 if unpaired), and *n* represents the number of sentences in the shorter document. Based on preliminary experiments we selected a threshold value of 0.1 and include all document pairs which score higher than the threshold in the final set of comparable documents. For the final document, we use the plain version of the non-anchor-translated documents, as shown in Figure 5.

```
'''Desmitdienu karš''', arī '''Slovēnijas neatkarības karš''' slovēņu - ''Slovenska osamosvojitvena vojna'' (''Slovēnijas neatkarības karš''), serbu - ''Рат у Словенији'' (''Karš Slovēnijā'')) - bruņots konflikts no 1991. gada 27. jūnija līdz 6. jūlijam starp Slovēniju un Dienvidslāviju, kura rezultātā nodibinājās neatkarīgā Slovēnijas valsts.

Desmitdienu karš bija iesākums daudz asiņainākiem kariem citās Dienvidslāvijas republikās. Tas bija arī gandrīz vienīgais no Dienvidslāvijas kariem, pēc kura neviena no pusēm netika apsūdzēta kara noziegumos.

==Priekšvēsture==

...
```

Figure 5: Plain-text documents

2.2.2. Topic Identification Method

For building a strongly comparable corpus one step is to identify pairs of documents that are topically related. Work in this direction is reported in Munteanu (2006) who describes a method of identifying parallel fragments in MCC. Another experiment in pairing topically related documents is due to Tao and Zhai (2005). They tackle the problem of MCC acquisition by devising a language independent method based on the frequency correlation of words occurring in documents belonging to a given time scale. The intuition is that two words in languages A and B whose relative frequency vectors Pearson-correlate over *n* pairs of documents in languages A and B that are paired by a time point *i*, are translation equivalents. This relative frequency correlation is then used as a translation equivalence association score of words in languages A

and B for describing a measure of document relatedness. Vu et al. (2009) improve the accuracy of method described above by a margin of 4% on an English-Chinese corpus. Wikipedia as a comparable corpus has been studied and used by Yu and Tsujii (2009). They sketch a simple mining algorithm for MCC, exploiting the existence of inter-lingual links between articles.

Our goal is to extract good quality MCC in languages Romanian, English and German for use in the ACCURAT project[2]. We have employed two different methods of gathering MCC from Wikipedia:

1. the first one considers an input list of good quality Romanian articles (articles that senior Wikipedia moderators and the Romanian Wikipedia community think that they are complete, well written, with good references, etc.) from the Romanian Wikipedia (http://ro.wikipedia.org/) and for each such article, it searches for the equivalent in the English Wikipedia;
2. the second one uses the Princeton WordNet and extracts all the capitalized nouns (single-word or multi-word expressions) from all the synsets. Then, it looks for Wikipedia page names formed with these nouns, extracts them and their correspondent Wikipedia pages in Romanian and German (it these exist).

The first method of MCC compilation uses 3 different heuristics of identifying the English equivalent of a given Romanian article (they are tried in the listed order):

a) it searches for an English page with the exact name as the Romanian page. For instance, we have found the following exact-match English pages (starting from the Romanian equivalents): "Alicia Keys", "Hollaback Girl", etc.;

b) it searches for the English link from the Romanian page that would lead to the same article in those languages. The Romanian version of the page may or may not be a complete translation from English (we noticed that the translation is usually shuffled – the narrative order of the English page is rarely kept and it usually reflects the translator's beliefs with regard to the content of the English page);

c) it automatically transforms the Romanian page name into an English Wikipedia search query by using a translation dictionary that has scores for each translation pair. Thus, for each content word in the Romanian page name, generates the first *k* translations (*k*=2 in our experiments) and with this query, retrieves the first 10 documents from the English Wikipedia. We manually chose the right English candidate but an automatic pairing method based on document clustering is described below.

Using these heuristics, we managed to compile a very good Romanian-English comparable corpus that consists of 128 paired Romanian and English documents of approx. 502K words in English and 602K words in Romanian.

The second method of MCC compilation uses Princeton Wordnet for extracting a list of named entities. These named entities are then transformed into Wikipedia links by replacing the white spaces with underscore and adding the string "http://en.wikipedia.org/wiki/" in front of them. Then, an application performs the following steps:

a) it goes to every link and downloads the Wikipedia page if it exits;
b) every downloaded Wiki page is searched for links to correspondent Romanian and

[2] http://www.accurat-project.eu/

German Wiki pages; if such links exist, those pages are also downloaded;

c) all the html tags of every EN-RO or EN-DE pair of Wiki documents are stripped so that only the plain text remains (there is also the possibility of preserving some mark-ups for important terms highlighted in Wikipedia articles); The categories of the documents are kept in a simple database.

Using the categories of the documents one can select documents referring to specific subjects. However, due of the fact that we searched only for named entities, confusions might occur. For example, Wiki articles about Paris, Rome or London might be considered to be about sports as they are categorized, among others, as "Host cities of the Summer Olympic Games". In reality, these articles contain very few information about such a topic. Table 1 shows the amount of comparable data we extracted from Wikipedia using the described method.

Table 1 The amount of comparable data extracted from Wikipedia using the second method

Named Entities pages about:	EN-RO	DE-RO
Sports	1043.9 K	534.1 K
Software	63.3 K	35.8 K
Medical	617.7 K	400.9 K
Other	43,965 K	25,042.8 K
Total (in words)	**45,689.9 K**	**26,013.6 K**
Total size	**418.3 MB**	**239.2 MB**

Clustering is an unsupervised machine learning technique that groups together objects based on a similarity measure between them. This technique is appropriate for pairing documents in a comparable corpus as to their topic similarity. Classical document similarity measures rely on the supposition that the documents have common elements (words). But documents in different languages have actually very few common elements (numbers, formulae, punctuation marks, etc.) and in order to make documents in different languages similar, one approach is to replace the document terms with their equivalent translation pairs. In this approach, each document term is replaced with the translation equivalents pairs from a translation equivalents list. The document vectors for both source and target language documents are collections of translation equivalents pairs.

There are several difficulties in this approach that have to be surpassed:

1. TRANSLATION EQUIVALENTS SELECTION. Not all the translation equivalents pairs have the same discriminative degree in differentiating between comparable documents.
2. CLUSTERING ALGORITHM MODIFICATIONS. The algorithm should consider pairing only different language documents.

TRANSLATION EQUIVALENTS TABLE. The accuracy of the comparable documents selection depends directly on the quality of the translation equivalents table. The translation equivalents table contains only content-word translations of lemmas with *N*-gram maximum lengths. For RO-EN, we have a clean dictionary with more than 1.5 million entries which we use for various scopes, while the other language pairs were gathered using Giza++ by ACCURAT members.

Considering the fact that not all the translation equivalents have the same discriminative degree for selecting comparable documents, the translation equivalents table was filtered using a maximum translation equivalents entropy threshold (0.5 in our case). Using this filtering method, light verbs, nouns with many synonyms, and other spurious translation equivalents are removed.

DOCUMENT COLLECTION. The documents were tagged and lemmatized. Considering only the content words, for each n-gram from the document collection a set of translation equivalents were selected from the translation equivalents table. For example, the translation equivalents for "acetic acid" in both English and Romanian are: "acetic - acetic", "acetic acid - acid acetic", "acid - acid".

CLUSTERING FOR COMPARABLE DOCUMENTS IDENTIFICATION. This technique relies on the supposition that translation equivalents can be used as common elements that would make documents in different languages similar. We choose an agglomerative clustering algorithm. We tested several simple distance measures like Euclidean distance, squared Euclidean distance, Manhattan distance and percent disagreement. We found that percent disagreement differentiates better comparable and non-comparable documents. Considering the document vectors $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ of which elements are 1 or 0 depending on whether the corresponding vocabulary term belongs to the document or not, the percent disagreement is computed as:

$$d(x,y) = \frac{\sum_{i=1}^{n} x_i \neq y_i}{n}$$

with n representing the size of translation equivalent table. The distance measure has the restriction that the compared documents have to be in different languages. This simple distance measure gave us a precision of 72% (with a maximum translation equivalents entropy threshold of 0.5 and a maximum of 3 translation equivalents per document term) on the collection of 128 English and Romanian Wikipedia documents described

3. Retrieval Techniques for Corpora from Narrow Domains

Automatic acquisition of comparable corpora from the Web is a challenging task on its own; therefore the Narrow Domain limitation introduced significantly increased the challenge of the task, especially when dealing with less resourced languages. In this section we will discuss the general methodology we used for automatically acquiring comparable corpora in specific narrow domains from the Web and describe in detail each step of the proposed workflow.

3.1. Methodology

Since this task involves collecting texts from a wide range of topical areas, we could not rely on a specific set of Web portals for providing the required volume of data. It was obvious from the start that some kind of open Web crawler would have to be deployed, a tool that would be able to travel through the Web and somehow collect useful html documents, without being limited by a Web domain restriction.

We have considered two main approaches. The first approach could be based on some kind of crawling engine that would focus on retrieving documents for each language separately, either using a "domain specific" crawler and therefore pre-classifying each document with a domain-type or using a general monolingual crawler and afterwards use a domain-classifier on the collected texts. This approach would be faster and easier to implement (given the fact that many good quality monolingual crawlers already exist) but on the other hand, the task of aligning the texts (in document level) would be a lot more complicated.

The second approach could use some kind of crawler that is able to crawl and retrieve bilingual (or multi-lingual) documents with a good possibility that these documents are either parallel or topic-related. As a starting point, a parallel crawler can be used and later on expanded to enable the acquisition of not only parallel but also comparable texts. This approach would be harder to implement, since this kind of crawling is not supported by one of the readily available crawlers. However, it would provide a significant boost to the alignment process, especially if we use a focused crawler (a crawler that implements a domain-specific filtering).

An initial pursue of the second approach showed that although there is a small number of tools available for automatic detection of parallel html documents, they mostly exploit URL and HTML structure similarities (Resnik, 1998, Esplà-Gomis, 2010). The main assumption is that two parallel html documents in a Web domain tend to have very similar URLs (sometimes the only differences are the two letters denoting the document's language) and/or the same HTML structure (most bilingual Web sites use the same template for displaying the same article in different languages). However, this is never the case with comparable documents. In reality, two comparable documents (especially when in different Web domains) will probably have completely different URLs and HTML structures and therefore it is impossible to find them using such techniques.

In addition, comparable documents will rarely have link connections to each other, which is another characteristic of parallel Web pages often used for finding parallel pairs. This led us to the conclusion that the easiest way to find and pair comparable pages would have to be based purely on the content and its' classification as topic-related or not.

With that in mind, we decided to turn to the first approach. We designed and implemented a focused crawling system, a Web crawler with a built-in lightweight topic classifier able to decide for each given Web page whether it is relevant to the desired topic or not. Using such a system, we were able to produce narrow domain comparable corpora in all ACCURAT languages in a similar way as the one presented by Talvensaari et al. (2008) who used a focused crawling system

to produce comparable corpora in the genomics domain in English, Spanish and German languages.

3.2. Focused Crawling

The implemented crawling system incorporates several sub-tasks ranging from bootstrapping the crawler with an initial base of URLs to start the crawling to Web page parsing, classifying and processing in order to extract the required information. The essential steps of this system are illustrated in Figure 6.

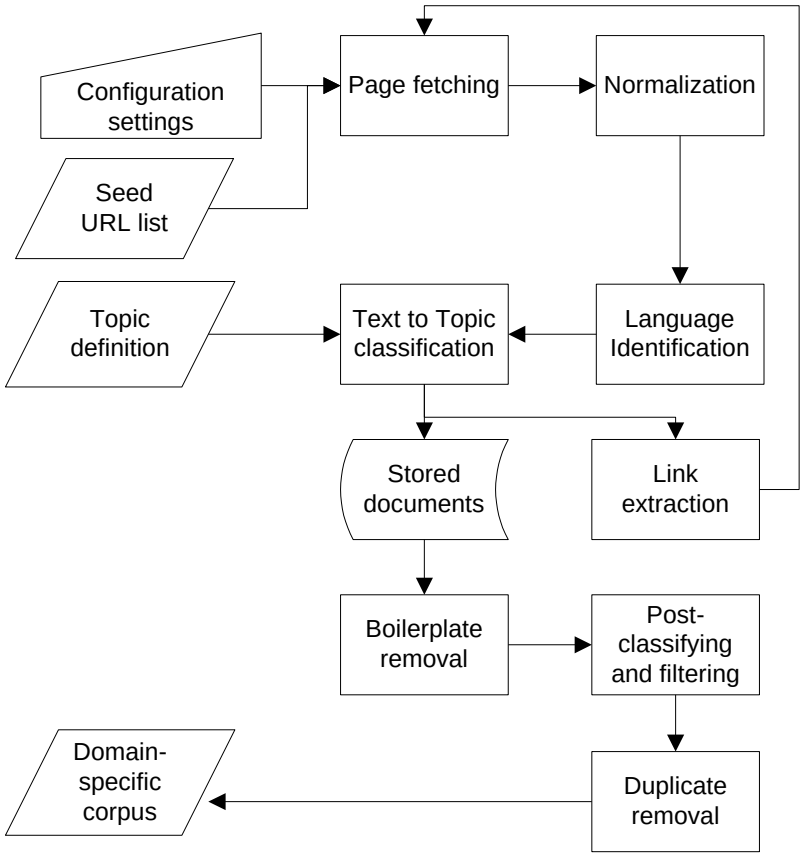

Figure 6: Essential steps of a focused crawler system

Bootstrapping the crawler was accomplished by providing two critical resources:

i. **Topic definition**. This is done by means of a list of weighted terms that are considered representative of a specific topic. Such a list can be constructed by manually selecting a representative set of terms and assigning weights. Online resources (e.g. Eurovoc) provide sets of words in different languages assigned in specific thematic categories and therefore can greatly assist in this process. Alternatively, these lists can be automatically extracted by small topic-specific corpora using tf-idf and term extraction algorithms. The topic definition consists of triplets representing the weight, term and the domain or subdomain it belongs to. A sample of the topic definition from the renewable energy domain is presented below:

100: natural resources=RenewableEN
100: natural processes=RenewableEN
100: biogas=RenewableEN

100: renewable power generators=RenewableEN

ii. **Seed URL list construction**. This is a simple list of URLs that are considered highly relevant to the topic. Again, such a list can be semi-automatically assembled with the help of one of the known search engines (e.g. yahoo, Google). BootCaT's (Baroni et al. 2004) tuple generation algorithm can also be used as follows: using the topic definition, a number of combinations of the topic terms are generated and then used as queries with Google. Top 5 or 10 URL results from each search are selected as candidates for the final seed URL list.

The next step in the chain of events is the actual crawling. Several crawling algorithms were examined:

1. The simplest and most common algorithm is Breadth-First (Pinkerton, 1994). Each page visited has its links extracted and inserted in the crawler's schedule, known as frontier. The frontier is filled in a First-In First-Out (FIFO) manner, meaning that the crawler visits links in the order they are found.
2. Fish-search (De Bra and Post, 1994) attempts to slightly improve the Breath-First algorithm. The main difference is that links, which were extracted from Web pages that were classified as irrelevant to the topic, are disregarded.
3. Best-First (Cho et al., 1998) was the logical evolution of Best-First. Instead of visiting pages in the order they are discovered, some kind of classifier is used which attempts to estimate a Web page's relevance to the topic. This relevance score is then used to sort the frontier and therefore ensuring that Best pages will be visited First.
4. Shark-search (Hersovici et al., 1998) is the first algorithm which attempts to measure the relevance of anchor texts (the text that is visible as an html link) and uses it to score each link. This was considered an improved version of Fish-Search.
5. PageRank (Brin and Page, 1998) introduced the concept of Web page popularity. Instead of scoring each page based on its content, this algorithm attempts to implement a ranking system by scoring a Web page depending on the number of other Web pages that have links to it. Therefore, a page is deemed as popular when there is a high number of links that lead back to it. Commonly used for indexing.
6. InfoSpiders (Menczer and Belew, 2000) is another algorithm that scores Web pages according to their relevance to the topic, but this time the user may assess the relevance of the documents visited by InfoSpiders up to a certain point. A distinct population of agents attempts to "sense" their local neighborhood by analyzing the text of the document where they are currently situated. The behavior of these agents can be subsequently altered by user's feedback therefore resulting to an adaptive environment.
7. Path algorithm (Passerini et al., 2001) again ranks pages based on their topic relevance, but also considers each page's distance from another relevant page (i.e. starting from a relevant page, how many links must be followed before reaching the current one).

Since we were seeking for a content-based solution, an algorithm that will prioritize most-relevant Web pages, a Best-First type of algorithm seemed the obvious choice. However, anchor texts can often indicate if a link will lead to a relevant Web page as well; therefore, a hybrid solution was used by employing the basic idea of a Best-First algorithm and the anchor text scoring introduced by Shark-Search.

3.3. Normalization and language identification

The text normalization phase involves detection of the formats and text encodings of the downloaded Web pages as well as conversion of these pages into a unified format (plain text) and text encoding (UTF-8).

In the language identification phase, each downloaded Web page is analysed and its language is identified. Documents that are not in the language of interest are discarded. *Lingua::Identify*[3], an open-source and flexible language identifier based on n-grams, is used for this task. *Lingua::Identify* did not originally support the Greek language; we provided a small corpus of Greek texts (taken from JRC Acquis) to the developer of the tool, who released a new version of the identifier, which we used throughout the subsequent work.

3.4. Text Classification

Our goal was to implement a "cheap" text classifier, so that it could be used during the crawling cycle without crippling the crawler's performance. In order to achieve a satisfying compromise between crawling speed (larger number of Web pages visited) and classification quality (less irrelevant pages actually fetched) we used a simple string-matching algorithm for the comparison of each crawled and normalized Web page to the topic definition. By adopting the method described in Ardö and Golub (2007), the score of relevance *c* for each Web page is calculated as follows:

$$c = \sum_{j=1}^{4} \sum_{i=1}^{N} \frac{w_j^l \cdot w_i^t \cdot n_{ij}}{l_j}$$

where N is the number of terms in the topic definition, w_j^l denotes the weight assigned to each location j of the HTML page (i.e. 10 for *title*, 4 for *metadata*, 2 for *keywords* and 1 for *main text*), w_i^t is the weight of term i, n_{ij} denotes the number of occurrences of term i in the location j and l_j is the number of words in the location j.

The calculated score models the likelihood that the page under consideration contains text relevant to the target domain. Therefore, if the score of relevance is under a predefined threshold[4], the page is classified as irrelevant and discarded. Otherwise, the page is stored and its links are extracted and added to the list of scheduled to be visited links.

In order to rank each link regarding the likelihood that the link points to a relevant Web page, we adopt an extension of the Shark-search (Hersovici et al., 1998) algorithm. Specifically, the potential score of each link is influenced by the estimated relevance of its anchor text (i.e. the visible, clickable text in a link), the text surrounding the link and the source Web page.

Analytically, the score $s_{l_{i,j}}$ of the *i*-th link of the *j*-th Web page ($l_{i,j}$) is calculated by the following formula:

[3] http://search.cpan.org/~ambs/Lingua-Identify-0.29/

[4] A typical value for this threshold is 100. This means that all web pages containing at least one term of weight 100 will pass at next processing steps (will not be discarded).

$$s_{l_{i,j}} = \sum_{m=1}^{N} w_m \cdot n_m + c_j$$

where w_m denotes the weight of the *m*-th term of the topic definition, n_m is the number of occurrences of the *m*-th term in the anchor text and the text surrounding the link, and c_j denotes the score of the source Web page (i.e. this value is constant for every link found in the *j*-th Web page).

3.5. Boilerplate Removal

Web pages often need to be cleaned from elements that are irrelevant to the content like navigation links, advertisements, disclaimers, etc. (often called boilerplate). Since we aim to collect comparable corpora useful for training MT systems, such parts of the HTML source are redundant. The algorithm we used for boilerplate removal (Kohlschütter et al., 2010), uses a set of shallow text features (link density, number of words in text blocks, etc.) for classifying individual text elements in a Web page as boilerplate.

3.6. Duplicate Detection

In (near) duplicate detection each new candidate document is checked against all other documents appearing in the corpus (e.g. by document similarity measures) before being added to the collection. An efficient algorithm for deduplication was used for this task (Theobald et al., 2008). The algorithm represents each document as a set of spot signatures, i.e. chains of words that follow frequent words as these are attested in a corpus. Each document is classified with respect to the cardinality of their set of spot signatures and so significantly reduces the time complexity of the task.

3.7. Post Filtering

The crawling engine used is able to score each page it visits by using the topic definition and a text-to-topic classifier. However, the crawler's scoring strategy will examine the whole HTML document (including title and metadata), since it focuses on finding not only Web pages that contain relative text blocks, but also pages with high probability to lead to other relative pages, even if they do not really contain any useful text on their own.

Therefore, in the final stage of the crawling framework, clean texts are examined and ranked in respect to the occurrences of words from the topic definition. This score along with thresholds regarding file size and type is used to filter out unwanted documents from the final collection.

4. Evaluation

In sections 2 and 3 above we described our tools for gathering comparable corpora. Here we begin to address the difficult question of how well these tools perform. One form of evaluation is to see to what extent parallel material extracted from these corpora may be used to improve MT systems, using well-known translation quality measures such as Bleu. ACCURAT will indeed carry out such evaluation as part of WP4. However, this is an indirect measure of the performance of the approaches to gathering comparable corpora since any improvements in MT performance resulting from exploitation of material extracted from comparable corpora depend not only on the quality of the gathered comparable corpora but on the tools for extracting parallel segments from them (WP2). Thus here we look at ways of directly assessing the comparability of the corpora we have gathered on the assumptions, which will be tested later, that (1) the measures of comparability we employ here will correlate with measures of the amount of parallel content extractable from the comparable corpora and (2) the parallel content extracted from the comparable corpora will improve MT quality.

Our intuition is that texts that "say the same thing", i.e. have considerable shared content, will be richer sources of parallel material than texts that do not. In the rest of this section we discuss work we have carried out, including development of new methods, to see whether our news and Wikipedia corpus gathering techniques are successful at gathering texts that do indeed share content.

4.1. News Evaluation

The idea that two news stories on the same event are likely to have some content in common is intuitive and one which has informed previous work on assembling comparable corpora (e.g. Braschler and Schauble (1998)). The basic idea is that two texts on the same event are talking about the same thing and therefore are likely to say similar things in their reports. However, if we take a closer look at how events are reported in news, both by examining real examples of news text and taking into account previous studies of news text, we find that the picture is not so straightforward.

First, events are multi-layered, complex objects. They come in different sizes, and have different complexity. This is often reflected in the content that is devoted to reporting such events. For example, consider texts 8, 9 and 10 in the Appendix, Table 1. These three texts are all about President Obama's one day visit to the Republic of Ireland and in that respect may be judged to be "on the same event". But if we look closely at the texts we can see that the texts are reporting different aspects of a complex event, and each has a different angle or slant to the story. Note also how the content, although overlapping in parts, is quite different. In text 8, the focus is on Obama's reference to ties between the US and Ireland, after emerging from a meeting with Taoiseach Enda Kenny; in 9 the focus and angle of the story are on what Obama had to say about the peace process and finally, in 10, the focus is on his reference to returning home to his roots and, in this text, the angle is on personal aspects of the visit.

Contrast this example with texts 1, 2 and 3, three texts about a Spanish earthquake. The first two texts are both initial reports issued when the quake had just struck and share a lot of content (in fact they are rewrites of the same newswire source). The third is a later report where the focus has shifted from announcing an earthquake to assessing the damage the quake has caused.

These examples suggest that a vague notion of "being about the same event" may not be sufficient to underpin an approach to gathering news texts with substantial shared content. To refine this notion we analyzed a wide range of news texts and reviewed the literature on the nature and structure of news reports. From this analysis we have developed a coarse-grained account of the functional structure of news texts on the basis of which we propose a scheme for classifying news text pairs into various classes reflecting the nature of their relatedness as reports of new events. Our hypothesis is that the amount of shared content between two news texts will correlate strongly with these relatedness classes. Work is on-going to verify this hypothesis, but is not reported here. What is reported here is:

- Our analysis of the nature and structure of news reports
- A scheme for classifying pairs of news reports based on their relatedness
- Experiments to confirm that human judges concur on assigning news text pairs to classes in our scheme
- Experiments to confirm that our news crawling tool gathers text pairs that are closely related according to our scheme.

Some of the experiments are still on going and results will be included in later deliverables.

4.1.1. Purpose of the Evaluation

The purpose of the evaluation is to test our methods and the extent to which our tool for gathering pairs of comparable news texts is able to gather texts which are closely related and hence likely to contain shared content.

4.1.2. Evaluation Methodology

Exploiting The Functional Structure of News Texts

While previous work has examined the structure of news texts from various perspectives (e.g. Bell, 1998 and Liddy et al., 1995), revealing patterns in news discourse, none of this work has carried out this analysis with a view to assessing content overlap between differing news accounts of the same or related events.

To further investigate functional patterns in news we hand-picked a small number of example texts from the on-line news domain, using particular news sites such as The BBC; The Guardian; The Telegraph and aggregated sources such as Google News. We restricted our searches to conventional news reports, i.e. texts reporting some new event or development in the world, excluding text types such as reviews, blogs and opinion or column pieces. Our development collection included: (1) texts on related events published at different points in time (e.g. reports of an Icelandic volcanic eruption and its potentially hazardous ash cloud, warnings of the knock-on disruption to airlines, warnings of the ash cloud risks to public health, etc.); texts on related events published at the same, or very similar points in time (e.g. reports on President Obama's 1 day visit to Ireland, early reports on the March 2011 Japanese earthquake and tsunami, etc.); and examples of similar, but different events at different points in time (e.g. reports on two different hurricanes).

We examined these texts and identified relations between events, both within a text and between different texts. We define an event as a specific thing that happens at a particular time and place. Key concepts are as follows:

Focal events: the event or events, which provide a focus for the text. Very often the most recent events in an unfolding news story, they also provide a particular angle or perspective for the report. Typically reported in the headline and first few lines of a news report, we may find in the body text: a fuller account (i.e. elaboration) of the focal events; background to the focal events and details of possible or actual subsequent events.

Background events: play a supporting role in the text, providing context for the focal events. May include: related events leading up to the focal events; examples of similar past events; and definitions, explanations or descriptions of things, people and or places which play a role in the focal events.

News_Events: a group of related events, broader than and including the focal event, which may be reported over time in *different* news text instalments. For example, initial reports on an earthquake News_event include details of a quake having occurred, while later reports cover rescue attempts, accounts of disaster aid and relief, etc. In such a case we view the texts as reporting on the **same** News_event. Note, in later reports, background events typically include details of previous events in the News_event.

Quotes: reported speech, typically indicated by quotation marks. May be part of the fuller account of the focal events or be part of the background.

An Event Relatedness Classification Scheme for News Text Pairs

Informed by our analysis of news, we developed a scheme for classifying pairs of news texts, where the classes are indicative of the relation holding between the events reported in the texts. A summary of the scheme is shown in Table 2, together with the types of shared content we can expect to see. We produced guidelines for analysing news texts, which describe the classes in the scheme, with examples. The guidelines are reproduced in Appendix 1.

Table 2 An event relatedness classification scheme for News text pairs

Class		Possible shared content patterns
SAME NEWS_EVENT	SAME FOCAL EVENTS	Common focal event, common elaboration, common background and common quotes.
	DIFFERENT FOCAL EVENTS	Focal event in one text appears as background in the other, common background and common quotes.
DIFFERENT NEWS_EVENTS, SAME TYPE	FOCAL_EVENTS SAME TYPE	Similar event structure for focal events, background in common (e.g. accounts of other similar events), details of one texts' news event appear as background in the other.
	FOCAL_EVENTS DIFFERENT TYPE	Background in common; details from one texts' news event appear as background in the other.
DIFFERENT NEWS EVENTS, DIFFERENT TYPE	RELATED via BACKGROUND	Details from one texts' news event appear as background in the other. Background in common.
	Other	No content in common

4.1.3. Evaluation Experiments

We specified a number of experiments to assess news texts using our event relatedness scheme. The experiments were based on evaluation datasets assembled from: (1) mono-lingual comparable corpora (pilot /full test set) and (2) multi-lingual comparable corpora (pilot /full test set). Details of the experiments and preparation of the evaluation corpora are as follows.

Web Based Assessment, Participants and Interface

In each task at least 2 human annotators assessed text pairs by answering a series of questions based on our scheme for annotating news texts. The questions are shown in Figure 7. We asked participants to read and follow our guidelines for analysing news text. Participants were ACCURAT partners.

Participants carried out the assessment tasks via a Web-based interface, which we developed in-house. The interface (see Figure 8) displays two texts side by side and then presents a set of questions with options for responses.

Evaluation Datasets: Extracting text from HTML

We created evaluation datasets which contained the news text only (i.e. each text is free from all images, adverts and other Web page content). To do this we stored local copies of the original html pages and then ran a custom built html-to-text converter, which removed all irrelevant content and html formatting. We then re-introduced light html to create machine-readable texts for display via a Web browser.

Question 1. Are both texts News Stories (i.e. and NOT opinion pieces, blogs, reviews, etc.)?
Yes, Please go to question 2. ⊙
No, Please go directly to the next pair. ⊙

Question 2. (Please answer all sub-parts)

Each text reports aspects of an underlying News Event. For this text pair, are the respective News Events:
A: The SAME News Event ⊙
(*IF A, THEN*)
1.1: Are the focal events in this text pair:
The same / partly the same ⊙
Different ⊙
1.2: Do these texts have any QUOTES in common?
Yes ⊙
No ⊙

B: DIFFERENT News Events ⊙
(*IF B, THEN*)
1.1: Are these news events the same type?
Yes ⊙
No ⊙
1.2: Are the focal events in these texts pair of the same type?
Yes ⊙
No ⊙
1.3: Is there any background in common?
Yes ⊙
No ⊙

Figure 7: Questions for Pilot News Assessment Interface

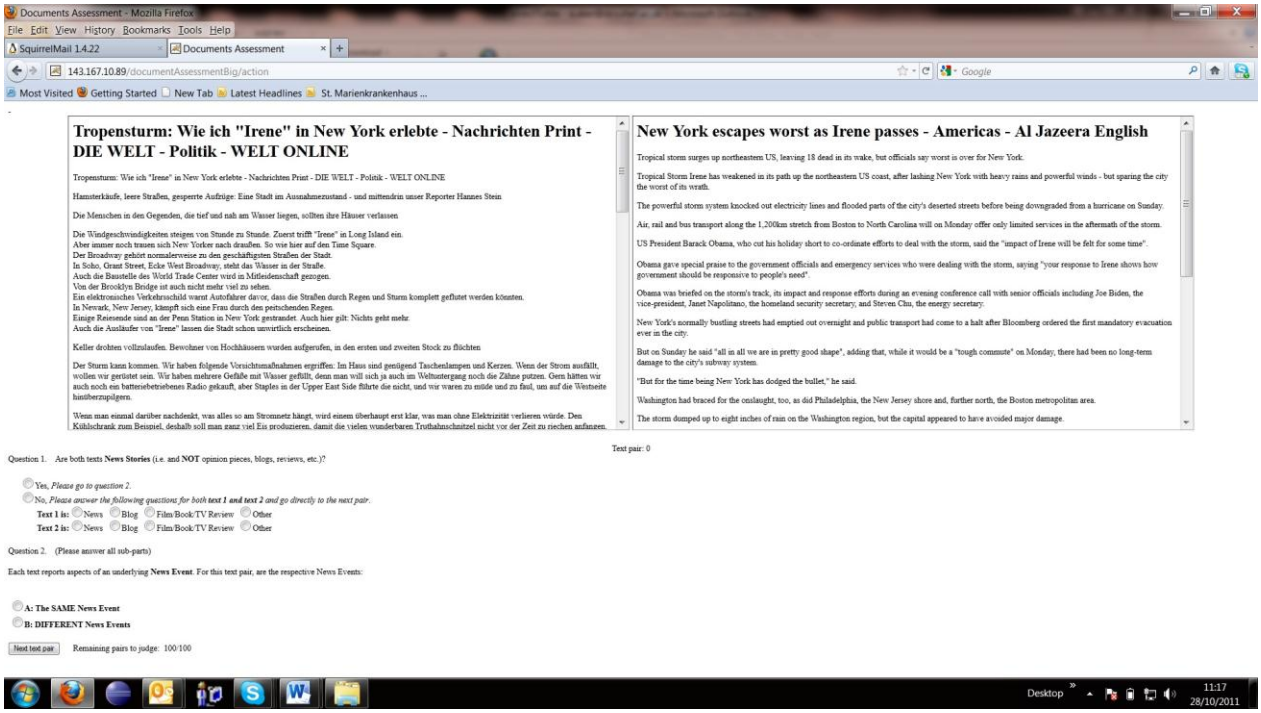

Figure 8: Interface for Assessing Comparability of News Texts

Mono-lingual Experiments

We first carried out exploratory experiments to test different versions of the interface and to obtain feedback on our scheme using mono-lingual English Corpora. We manually selected 50 news text pairs across a 24 hour period from the results of our method for automatically harvesting texts on news events. We selected stories from various points in the ranked list of results to obtain examples of texts in the different categories: particularly same news event, same focal events, and same news event, different focal events. We also collected 9 text pairs by hand to provide examples in the other categories. In the pilot exercise, initial results from 2 annotators showed high levels of coder agreement on key classes of same news event and same focal event; on the basis of user feedback from this exercise, we decided to omit questions about background content because they were judged to be too hard and too time consuming for the task to be feasible.

We plan to develop a full dataset of 100 mono-lingual comparable news texts in order to test more fully the categories in our scheme and for use in a second stage of experiments to investigate the amount of shared content which may be correlated with the different categories. This work will be reported elsewhere.

Experiments with Multi-lingual Corpora

In the pilot exercise based on multi-lingual corpora we asked annotators to classify a small dataset of 10 text pairs per language pair for seven European languages, (Latvian, Lithuanian, Estonian, Greek, German, Romanian and Croatian), paired in each case with English. The 10 text pairs chosen were those that came at the top of the ranking produced by our comparable

news corpus gathering system, described above in section 2.1. The aims of the pilot were as follows: to test our experimental setup: the interface; guidelines for analysing news and to

Table 3 % agreement for annotator responses, for different language text pairs.

Language pair	Is News Story?		News Events		Focal Events		Quotes		News Event Type		Focal Events Type		Background	
SL-EN	80.0	8/10	85.7	6/7	83.3	5/6	83.3	5/6	na		na		na	
RO-EN	90.0	9/10	100.0	8/8	100.0	8/8	100.0	8/8	na		na		na	
EL-EN	100.0	10/10	70.0	7/10	100.0	7/7	85.7	6/7	na		na		na	
LT-EN	100.0	9/9	88.9	8/9	87.5	7/8	75.0	6/8	na		na		na	
ET-EN	90.0	9/10	100.0	8/8	100.0	2/2	100.0	2/2	83.3	5/6	83.3	5/6	66.7	4/6
DE-EN	60.0	6/10	66.7	4/6	100.0	4/4	50.0	2/4	na		na		na	
HR-EN (a1-a2)	100.0	10/10	80.0	8/10	75.0	6/8	100.0	8/8	na		na		na	
HR-EN (a1-a3)	100.0	10/10	90.0	9/10	77.8	7/9	100.0	9/9	na		na		na	
HR-EN (a2-a3)	100.0	10/10	90.0	9/10	66.7	6/9	100.0	9/9	na		na		na	
LV-EN (a1-a2)	90.0	9/10	100.0	9/9	66.7	2/3	100.0	3/3	33.3	2/6	83.3	5/6	33.3	2/6
LV-EN (a1-a3)	90.0	9/10	62.5	5/8	100.0	3/3	100.0	3/3	50.0	1/2	100.0	2/2	50.0	1/2
LV-EN (a2-a3)	80.0	8/10	62.5	5/8	66.7	2/3	100.0	3/3	0.0	0/2	100.0	2/2	100.0	2/2
Average	**90.0**		**83.0**		**85.3**		**91.2**		**41.7**		**91.7**		**62.5**	

provide participants with a warm up task. The results are presented in Table 3. The right hand column gives the number of responses where both annotators were in agreement, out of the total number of responses where both annotators were asked to provide a response. The left hand column shows percentage figures. We designed the interface such that there were dependencies between questions: different response types meant that follow on questions would or would not be asked. The low number of responses for certain questions resulted from these dependencies, and the nature of the relations between the text pairs in this test sample. (Most examples were on the same news event, the category in which we were most interested in testing in the pilot study). There were three annotators for Latvian-English and Croatian-English pairs.

These initial figures show that agreement levels on the different questions between annotators were very high, typically between 70-90%. The results, although preliminary, suggested that the guidelines and assessment task are both valid and straightforward.

In a second experiment, currently underway, we asked the same annotators to complete the full task, involving 100 harvested text pairs per language pair. In order to investigate the full range of paired text results, as ranked by our system, we selected the evaluation dataset from deciles, 10 pairs randomly selected from each decile. The aims of this experiment were twofold: First, to provide feedback on the results of our tool for automatically harvesting comparable corpora from news sources, i.e. to show what relations exist between harvested text pairs. Second, we aim to measure inter-subjectivity agreement between annotators for all language pairs and for the 100 texts. We hypothesise that the results will show high agreement between subjects and hence validate the classes in our scheme. The results from this second experiment will be reported later in the project.

4.2. Wikipedia Evaluation

Wikipedia provides a multilingual corpus which is aligned in document level. Although these documents could be considered comparable, it is likely that the degree of similarity varies widely: some may be translations of each other, while others may have been developed independently and share little information. This evaluation method aims to assess the comparability between Wikipedia documents and to identify the characteristics of these comparable documents.

4.2.1. Evaluation Methodology

In order to develop a suitable evaluation method, a pilot task was run to test an initial similarity assessment scheme based on a 5-point Likert scale: (1) Very Different to (5) Very Similar. Assessments were made internally on 5 pairs of Slovenian-English documents. For this initial experiment, the Slovenian documents were translated into English using Google Translator. Ten assessors judged each document pair and initial findings showed the level of agreement to be 0.305 measured using Fleiss' Kappa. Assessors are also asked to specify reasons of chosen similarity scores of the five document pairs. Based on the findings from the pilot test we adapted the evaluation scheme as shown in the next figure.

4.2.2. Interface

We developed an online assessment tool which showed document pairs (in their original language) side by side and asked assessors to answer five questions, which were derived from our findings in the pilot task. These questions are shown Figure 9.

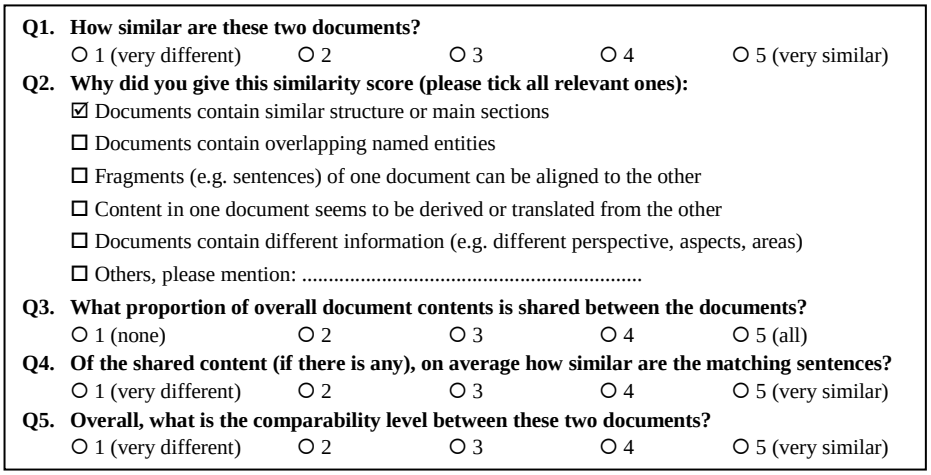

Figure 9: Interface of Wikipedia Assessment Tool

The first question (Q1) aims to identify the similarity between the two shown documents. To answer this question, assessors were asked to judge the similarity based on the aspects contained in the documents, rather than the similarity of the topic itself. Assessors were then asked to give reasons (Q2) to justify the chosen similarity scores by choosing one of the six options, which were previously derived from the internal pilot's results. In the third question (Q3), assessors were asked to provide a score on the proportion of shared contents between the documents, and assess the similarity of sentences within these shared contents in fourth question (Q4). Last (Q5),

we aim to identify whether documents' comparability level are related to similarity in general and ask assessors to specify a score for the comparability level of the document pairs.

4.2.3. Pooling of Documents

To select documents for the evaluation, we first sorted documents by their comparability scores resulted from the retrieval method and divided the set into 10 equal bins. Due to the unavailability of document pairs in some bins, we randomly selected 2 document pairs from each bin, resulting in a range of 11-18 document pairs to be selected for each language pair (with an average of 15 documents pairs).

4.2.4. Results

Out of the 105 chosen document pairs, 32 pairs were assessed by 1 assessor; the remaining 73 by 2 assessors thereby enabling the calculation of inter-assessor agreement (Table 4). Scores are calculated over the original 5-point scale and also for a 2-point scale created by aggregating the results for scores 1-2 (low similarity) and 3-5 (high similarity). Although the level of agreement between assessors on the four questions is low, the results are based on only 2 assessors. Question 3 is found to be the most disputable question with the lowest agreement score.

Table 4. Assessors' agreement on different questions

Questions	Percent Agreement		Cohen's Kappa	
	5 point scale	2 point scale	5 point scale	2 point scale
Q1. Similarity Reasons	38.40%	78.08%	0.165	0.163
Q3. Proportion	38.40%	69.86%	0.151	0.028
Q4. Similar Sentences	37.00%	86.00%	0.105	0.225
Q5. Comparability Level	43.80%	75.34%	0.117	0.287

As shown in **Table 4**, assessors' percent agreements almost doubled when we use score aggregation. However, this is not the case with Cohen's Kappa for Q1 and Q3, which suggests that people have certain disagreements even after score aggregration was used. Further analysis shows that assessors have different scoring ranges (some used scores from 1-5, some used scores from 2-4, etc.). Even though there are some disagreement, assessors seemed to agree that a document pair *m* is more similar than a document pair *n*. We aim to properly measure this issue in the near future.

Across all document pairs, we assessed the correlation between different questions and found that similarity level highly corresponds to the proportion of shared contents in the documents (Pearson=0.83). Interestingly, the similarity level and comparability level between documents correlate lower (Pearson=0.75), suggesting that people have different judgments in answering these two questions. The correlation for Q4 and Q5 is the lowest (Pearson=0.68), which shows that documents with less shared contents may still have highly similar sentences. Furthermore, this evaluation scheme has enabled us to analyse the characteristics of document pairs from each similarity score in more detail (Figure 10). The results suggest that the first three characteristics (similar structure, overlapping named entities and overlapping text fragments) may not necessarily predict the degree of similarity between multi-language document pairs; whilst highly similar (or parallel) document pairs will often contain what appear to be translations of the content.

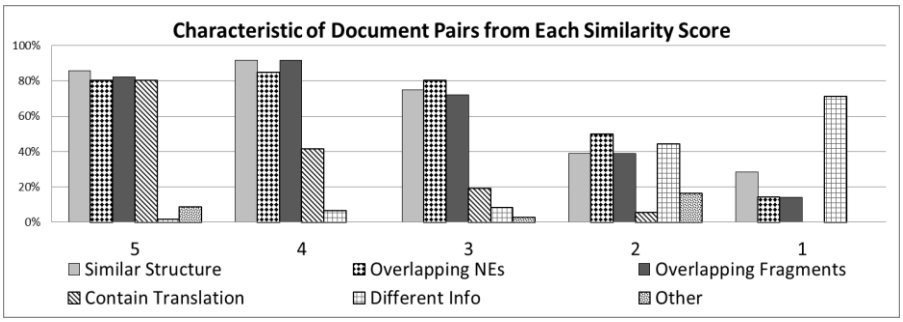

Figure 10: Characteristic of document pairs for different similarity scores

We conducted further experiments by comparing the anchor based similarity measures to a different similarity measures which use MT. In this case, source language documents are translated into the target language, enabling documents to score higher word overlap This analysis aims to find out whether simple language independent feature (*anchor and word overlap)* can be used to replace other complicated feature (*translation*). We found the Pearson correlation between these two measures to be 0.547, which shows that to some extent, simple language independent feature does correlate with MT measure. On the other hand, these will not mean anything if users do not agree with any of the similarity measures. We therefore calculate the correlation between the different similarity measures to human assessment. This finding is shown in Table 5.

Table 5. Pearson Correlation between Human Assessment and Comparability Measures

Measures	Correlation to human assessment
Translation	0.337
Anchor + Word Overlap	0.104

Based on the correlation scores in Table 6, we found that both measures do not accurately represent human's preference, which are shown by the low Pearson correlation scores. We attempt to analyse this further by looking into the correlation between these measures on different languages. We are aware that the evaluation set is very small (15 document pairs/language pair assessed by 1-2 people) to be able to provide a reliable conclusion on the performance on different languages. Therefore, we are currently running a bigger evaluation set which includes 100 document pairs/language pair to provide a more robust experiment. Meanwhile, the preliminary result for each language is shown in Table 6.

Table 6. Correlation between Human Assessment and Comparability Measures on Different Languages

Language Pair	Translation	Anchor + Word Overlap
DE-EN	0.768	0.650
EL-EN	0.459	-0.028
ET-EN	0.098	0.209
LT-EN	0.241	0.023
LV-EN	0.360	0.231
RO-EN	0.327	0.320
SL-EN	0.475	0.149

The results in Table 6 show that the measures depend on the language and the performance between these measures vary greatly. Similarity measures which use translation feature perform better in most language pairs than anchor and word overlap. Interestingly, the latter has very high correlation (even though not as high as translation) with human assessment in one language pair: German-English. Several reasons which may cause this are the amount of anchor texts in German documents or the fact that the language itself are more similar than other language pairs, such as Lithuanian, or Greek. However, none of these features currently are able to predict human's assessment with high accuracy. We are currently running an assessment task with bigger dataset (100 document pairs per language pair), which will enable us to perform further analysis on the retrieval method. We will report our findings in the near future.

5. Conclusions

In this deliverable, we described retrieval methods developed to gather comparable documents from different sources on the Web. We gathered news documents by making use of Google News and using different features: date, time, title length and title similarity, to align the documents. We developed different method to retrieve comparable Wikipedia documents, which is by making use of anchor information in Wikipedia to identify and retrieve documents which contain comparable (or parallel) sentences. We also analysed a different method to gather documents from Wikipedia by retrieving documents of the same topic. Lastly, we developed retrieval methods to collect documents from narrow domain corpora by using a focused crawling method.

Several evaluation tasks have been performed to assess the retrieval methods' performance. We reported the evaluation scheme we used and discussed the results from project pilot task. We are currently running the final evaluation task to gather more data to enable further experiments on the data and retrieval methods in general. Results from the final evaluation task will be reported in the extended deliverable.

6. References

Ardo, A., and Golub, K. 2007. Documentation for the Combine (focused) crawling system, http://combine.it.lth.se/documentation/DocMain/

Baroni, M., and Bernardini, S. 2004. BootCaT: Bootstrapping corpora and terms from the Web. In Proceedings of LREC 2004. 1313-1316.

Bell, A. 1998. The discourse structure of news stories. In Allan Bell and Peter Garrett, editors, Approaches to Media Discourse. Blackwell Publishers.

Braschler, M. and Schauble, P. 1998. Multilingual Information Retrieval Based on Document Alignment Techniques. In Research and advanced technology for digital libraries: second European conference, ECDL'98, Heraklion, Crete, Cyprus, September 21-23, 1998: proceedings, page 183. Springer Verlag, 1998.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems. 30, 1–7, 107–117.

Cho, J., Garcia-Molina, H., and Page, L. 1998. Efficient crawling through URL ordering, Computer Networks and ISDN Systems. 30, 1–7, 161–172.

De Bra P. M. E. and Post R. D. J. Information retrieval in the World-Wide Web: Making client-based searching feasible. Computer Networks and ISDN Systems, 27(2):183–192, 1994.

Espla-Gomis, M., and Forcada, M.L. 2010. Combining Content-Based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor. The Prague Bulletin of Mathematical Linguistics. 93, 77-86.

Fung, P. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In Parallel Text Processing, pages 1--17. Springer.

Hersovici, M., Jacovi M., Maarek, Y. S., Pelleg, D., Shtalhaim, M., and Ur S. 1998. The sharksearch algorithm—An application: TailoredWeb site mapping, Computer Networks and ISDN Systems, 30, 1–7, 317-326.

D. Huang, L. Zhao, L. Li, and H. Yu. Mining large-scale comparable corpora from Chinese-English news collections. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 472{480. Association for Computational Linguistics, 2010.

Kohlschütter, C., Fankhauser, P., and Nejdl, W. 2010. Boilerplate Detection using Shallow Text Features. The Third ACM International Conference on Web Search and Data Mining.

Li, B., Gaussier, E. and Aizawa, Akiko N. 2011. Clustering comparable corpora for bilingual lexicon extraction. In ACL (Short Papers), pages 473--478.

Liddy, E.D., Paik, W., and McKenna. M. 1995. Development and Implementation of a Discourse Model for Newspaper Texts, AAAI Technical Report SS-95-06.

Menczer, F. and Belew, R. 2000. Adaptive retrieval agents: Internalizing local context and scaling up to the Web. Machine Learning 39, 2–3, 203–242.

Moore, R. C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users, pages 135–144, London, UK, 2002. Springer-Verlag

Munteanu, D. Ş. (2006). Exploiting Comparable Corpora. PhD Thesis, University of Southern California, December 2006. ©2007 ProQuest Information and Learning Company

Munteanu, D.S and Marcu, D. Extracting parallel sub-sentential fragments from non-parallel corpora. In ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 81{88, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

Passerini, A., Frasconi, P., and Soda, G. 2001. Evaluation methods for focused crawling, Lecture Notes in Computer Science. 2175, 33–45.

Pinkerton, B. 1994. Finding what people want: Experiences with the Web Crawler. In Proceedings of the 2nd International World Wide Web Conference.

Resnik P., Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text, in D. Farwell, L. Gerber, and E. Hovy (eds.), Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas (AMTA-98), Langhorne, PA, Lecture Notes in Artificial Intelligence 1529, Springer, October, 1998.

P. Resnik. Mining the Web for bilingual text. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pages 527{534. Association for Computational Linguistics, 1999.Skadiņa, I., Vasiļjevs, A., Skadiņš, R., Gaizauskas, R., Tufiş, D. Gornostay, T.: Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. In Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010, Malta, pp. 6-14.

Talvensaari T., Pirkola A., Järvelin K., Juhola M. and Laurikkala J. October 2008. Focused Web crawling in the acquisition of comparable corpora. Information Retrieval, 11, 5, 427-445.

Tao, T., and Zhai, CX. (2005). Mining Comparable Bilingual Text Corpora for Cross-Language Information Integration. In Proceedings of the 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005.

Theobald, M., Siddharth, J., and Paepcke, A. 2008. SpotSigs: Robust and Efficient Near Duplicate Detection in Large Web Collections. In: 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2008).

Tufiş, D., Ion, R., Ceauşu, Al., and Ştefănescu, D. (2008). RACAI's Linguistic Web Services. In Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco, May 2008. ELRA - European Language Ressources Association. ISBN 2-9517408-4-0.

Yu, K., and Tsujii, J. (2009). Extracting Bilingual Dictionary from Comparable Corpora with Dependency Heterogeneity. In Proceedings of NAACL HLT 2009: Short Papers, pages 121–124, Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

Vu, T., Aw, A. T., and Zhang, M. (2009). Feature-based Method for Document Alignment in Comparable News Corpora. In Proceedings of the 12th Conference of the European Chapter of the ACL, pages 843–851, Athens, Greece, 30 March – 3 April 2009. ©2009 Association for Computational Linguistics

7. Appendix I

This appendix contains the full text of the guidelines given to human judges, who were asked to assess the comparability of news texts.

Assessing the Comparability of Texts: guidelines for analysing news text

In this exercise you will be asked to examine a pair of news texts and to identify and compare particular content in these texts. We are interested in conventional news texts and key features of news, including: the *Focal event(s)*, *Quotes*, *News_events and Background*. To assist you in this task, we provide guidelines and examples for identifying and comparing news content. We ask participants to read these carefully and to have a good understanding of the content types. Please don't hesitate to ask if you have questions and please report any feedback you may have from the task. Thankyou!

> **Important note on images:** please compare content based on **the text alone**. I.e. please ignore any images and try not to let images influence your judgements. Images can be misleading, remember, just one paparazzi image can be used in many different contexts!!

1 The News Story Text Type.

In Question 1 we ask you to decide if both texts in a pair are *news stories* i.e. texts reporting some new event or development in the world (see Table 1 for examples). Our scheme for analysing content is based on the conventional news story format and does not accommodate patterns found in other news media texts such as: blogs; reviews; columnist or opinion pieces; etc. If you decide that both texts are *not* news stories we will ask you to go to the next text pair.

2 Content in News: Focal Events

Typically a news text has a **focal event(s)**, i.e. the event, or events, which provide a focus for the text. We define an event as a specific thing that happens at a particular time and place. Very often the focal events in a text are the most recent events in an unfolding news story. We also find that journalists may choose to focus on certain events in order to provide a particular angle or perspective for their report.

We can identify focal events using textual cues. Events are typically brought into focus via the headline, sub-heading and/or the first few lines of a news text. Very often the focal events determine the "slant" of a text. In the remaining body text we may find a more detailed account of the focal events, background to the focal events and details of possible or actual subsequent events.

SOME EXAMPLES:

In Table 1, texts 1 and 2, we have highlighted in red the text that indicates focal events. Texts 1 and 2 have the same focal events, i.e.: two earthquakes in Spain resulting in fatalities, injuries and damage to property. The texts go on to describe the immediate official response and where the wounded were being treated.

These examples can be contrasted with text 3, which reports on further consequences of the quakes. In this text the focus (highlighted in red) is now on the aftermath of the quake, residents taking stock of the quake and its impact, and less on the quake itself. Details of the initial quakes are presented as background events (See Section 2 below). In sum, if we compare texts 1 and 3 or texts 2 and 3, we say the focal events are different.

Texts 4 and 5 also provide an example of a text pair with different focal events. In this case, both texts refer to consequences of a second ash cloud from Iceland, but text 4 focuses on the warning to lung patients whereas text 5 focuses on the disruption for airlines.

Finally, texts 8, 9 and 10 all report on President Obama's visit to Ireland, but the focal events in these texts are different. In 8 the text focuses on Obama's reference to ties between the US and Ireland. In 9 the focus is on Obama's reference to the peace process. Whereas in 10, the focus is on his reference to returning home to his roots and, in this text, the emphasis is on personal aspects of the visit.

2 Content in News: Background Events

Background events play a supporting role in the text, providing context for the focal events. Typically background events include: events leading up to the focal events; examples of similar events; other past events used to provide historical context for the focal events and definitions or descriptions of things, people and or places involved in the focal events.

For example, text 4, which reports on problems arising from the 2011 Icelandic ash cloud provides a good example of background events (highlighted in red). In this case the current ash cloud event is compared with the similar event of the 2010 ash cloud.

We only ask participants to look for background in a text pair when the **News_events** (see 4 below) are different.

3 Content in News: Quotes

Journalists take and use quotes relating to news events in different ways. We are interested if the text pairs have any quotes in common. Quotes are typically indicated by speech marks: " ". We note that due to issues in rendering the html, speech marks are sometimes displayed by substitute characters.

We are looking specifically for shared content in the quoted material, for example, shared propositions or assertions (and not simply duplicate names, or single words). For an example of quotes in common see the passages taken from President Obama's speech, in texts 8 and 9:

Text 8: "*On progress in Northern Ireland, Mr Obama said it spoke "**to the possibilities of peace and people in longstanding struggles being able to reimagine their relationships***".

Text 9 "*..the president said: "I want to express to the Irish people how ... inspired we have been by the progress made in Northern Ireland because it speaks **to the possibilities of peace and people in long-standing struggles to be able to re-imagine their relationships***."

4 News_events

Very often, the focal events reported in a particular news text can be viewed as part of a broader group of related events. For example, the two earth tremors that affected Spain were part of a series of events: the subsequent rescue of victims, the treatment of those wounded in the quakes, people being forced to sleep outdoors, queueing for food, and the response of the town residents to the damage (see Table 1, texts 1, 2, 3). We use the term **News_event** to refer to instances of such broad groupings of related events[5].

The various events associated with a News_event may be reported over a series of news text instalments. So, for example, details of a person being kidnapped may be in early news texts, while details of the subsequent ransom demand from the abductors may emerge a week later in a different text. In such a case we view the texts as reporting on the **SAME** News_event.

SOME EXAMPLES:

In table 1, examples of groups of texts reporting on the same News_event include: {1, 2 and 3}; { 4 and 5} and { 8,9 and 10}. By contrast, we find that different texts may refer to different News_events of the ***same type***. For example, see text 6, which reports on an earthquake, (the Christ Church Earthquake, 2011). This is the same type of News_event as that referred to in texts 1, 2 and 3 (type = earthquake).

5 Distinguishing News_events and News_event Types

Such groupings of related events are largely intuitive, and the task of assessing whether two different texts refer to the same News_event is usually straightforward. For example, most people would agree that an "earthquake" event might involve related events such as: the tremor itself; aftershocks; a measurement of the scale of the tremor; buildings being shaken; property damaged; people killed and injured; rescue attempts; assessments of damage; evacuation of residents; donations of aid; etc. Also, there are often clues about the News_event in the headline and the body text, and very often, a group of related events may be named or explicitly referred to in the text. Previous references to News_events include, for example: "Hurricane Katrina"; "Phone Hacking Scandal"; "The Royal Wedding"; "The 2012 Olympics".

[5] In everyday language people refer to such broader groupings of related events using terms such as "story": eg "the story refuses to go away"; "the on-going, phone hacking story". But we also find people use the term "story" in another sense, e.g., to refer to the telling of or *framing* of a particular event: "what's the story [angle] here?". In the latter, "story" refers to the human construction of a version of events, where events (often unrelated) are selected and woven into a particular interpretation of events in the world. In view of this ambiguity, and to indicate our focus on directly related events, we use the term "News_event".

There are however questions concerning the boundaries of news_events, which we believe are important to address for the purpose of this comparability exercise. First, we can imagine numerous "types" of News_event, from earthquakes, hurricanes and kidnaps to scientific discovery, political scandals and sporting news, and there are various ways of arranging such types. For example, earthquakes, tornadoes and hurricanes and floods are all types of natural disaster, but we can also view the latter as examples of extreme weather events. We note that different instances of a News_event may have different component events. For example: not all earthquakes may be associated with a Tsunami; certain hurricanes may result in a sea surge and massive flooding, others may not. Moreover, boundaries between related News_events are not always clear. For example, consider the March 2011 Japanese earthquake and tsunami and the subsequent rescue, evacuation and nuclear crisis. At which point do we consider that the crisis at the Sendai nuclear plant, the radiation threat and subsequent clean up should count as a distinct News_event from the events involved in the earlier Japanese Tsunami and earthquake?

In light of these issues and to encourage consistency in judgements we provide a reference set for the task of identifying and comparing News_events in news text. An appropriate "off-the-shelf" scheme is the list of topics and events defined by the Topic Detection Track (TDT) programme, for TDT4. Our definition of News_event is very similar to the TDT "Topic", which they define as: "an event or activity, along with all directly related events and activities"[6]. The TDT4 topic guidelines, reproduced below, comprise a set of 12 topic types and "rules for interpretation", a list of what types of events should be considered as related, for each particular topic. Note that a particular instance of a News_event need not involve all the listed events here.

Also, for each of the 12 topic types TDT distinguishes a subset of "seminal events" from other related "topic" events. This distinction is not relevant to our current task, and the events listed under both "seminal events" and "topic events" should be considered together (the seminal events are simply a more fine grained view of events listed under the topic set).

Appendix

When comparing News_events, please refer to the following scheme of event groupings, the TDT "**Annotation Guide"**, 2003. (Reproduced from: http://projects.ldc.upenn.edu/TDT4/Annotation/label_instructions.html 08/08/11)

"In TDT-4, there are twelve topic types with corresponding rules of interpretation, as follows:

[6] One difference between a News_event and TDT topic , is that the TDT task has placed emphasis on identifying the seminal or triggering events within a topic, and this informed their definition of a topic. "A TDT style event is defined as a specific thing that happens at a specific time and place along with all necessary preconditions and unavoidable consequences. It is defined by first defining an initial seminal event , which is the starting point for an event, For instance, when an U.S. Marine jet sliced a funicular cable in Italy in February 1998, the cable car's crash to earth and the subsequent injuries were all unavoidable consequences and thus part of the same event". We do not use the distinction of "seminal event" in our current task of assessing comparability of news texts.

1. Elections, e.g. 30030: Taipei Mayoral Elections

Seminal events include: a specific political campaign, election day coverage, inauguration, voter turnouts, election results, protests, reaction.

Topic includes: the entire process, from announcements of a candidate's intention to run through the campaign, nominations, election process and through the inauguration and formation of a newly-elected official's cabinet or government.

2. Scandals/Hearings, e.g. 30038: Olympic Bribery Scandal

Seminal events include: media coverage of a particular scandal or hearing, evidence gathering, investigations, legal proceedings, hearings, public opinion coverage.

Topic includes: everything from the initial coverage of the scandal through the investigation and resolution.

3. Legal/Criminal Cases, e.g. 30003: Pinochet Trial

Seminal events include: the crime itself, arrests, investigations, legal proceedings, verdicts and sentencing.

Topic includes: the entire process from the coverage of the initial crime through the entire investigation, trial and outcome. Changes in laws/policies as a result of a crime are not generally on-topic unless a clear and direct connection between the specific crime and the legislation is made.

4. Natural Disasters, e.g., 30002: Hurricane Mitch

Seminal events include: weather events (El Nino, tornadoes, hurricanes, floods, droughts), other natural events like volcanic eruptions, wildfires, famines and the like, rescue efforts, coverage of economic or human impact of the disaster.

Topic includes: the causal (weather/natural) activity including predictions thereof, the disaster itself, victims and other losses, evacuations and rescue/relief efforts.

5. Accidents, e.g., 30014: Nigerian Gas Line Fire

Seminal events include: transportation disasters, building fires, explosions and the like.

Topic includes: causal activities and all their unavoidable consequences like death tolls, injuries, economic losses, investigations and any legal proceedings, victims' efforts for compensation.

6. Acts of Violence or War, e.g., 30034: Indonesia/East Timor Conflict

Seminal events include: a specific act of violence or terrorism or series of directly related incidents (such as a strike and retaliation).

Topic includes: Direct causes and consequences of a particular act of violence such as preparations (including technological/weapons development), coverage of the particular action, casualties/loss of life, negotiations to resolve the conflict, direct consequences including retaliatory strikes. This topic type is difficult to define across the board, and can easily become extremely broad and far-reaching. As such, each topic of this type is treated individually and is defined in such a way as to sensibly limit its scope and make annotation manageable.

7. Science and Discovery News, e.g., 31019: AIDS Vaccine Testing Begins

Seminal events include: announcement of a discovery or breakthrough, technological advances, awards or recognition of a scientific achievement.

Topic includes: Any aspect of the discovery, impact on everyday life, the researchers or scientists involved, descriptions of research and technology directly involved in the discovery.

8. Financial News, e.g., 30033: Euro Introduced

Seminal events include: specific economic or financial announcements (like a specific merger or bankruptcy announcement); reactions to the event; direct impact on the economy or business world. General economic trends or patterns without a clear seminal event are not appropriate as TDT topics.

Topic includes: the specific event, its direct causes, impacts on finance, government interventions or investigations, public or business world reactions, media coverage and analysis of the event.

9. New Laws, e.g., 30009: Anti-Doping Proposals

Seminal events include: announcement of new legislation or proposals, acceptance or denial of the legislation, reactions.

Topic includes: the entire process, from announcement of the proposal, lobbying or campaigning, voting surrounding the legislation, reactions from within the political world and from the public, challenges to the proposal, analysis and opinion pieces concerning the legislation.

10. Sports News, e.g., 31016: ATP Tennis Tournament

Seminal events include: a particular sporting event or tournament, sports awards, coverage of a particular athlete's injury, retirement or the like.

Topic includes: training or preparations for a competition, the game itself, results. For tournament and championship events like the World Series or Superbowl, only direct precedents are considered on topic. Therefore, semi-finals and finals games leading up to the championship are on topic, but regular season play is not.

11. Political and Diplomatic Meetings, e.g., 30018: Tony Blair Visits China

Seminal events include: preparations for the meeting, the meeting itself, outcomes, reactions.

Topic includes: the whole process from the preparations and travel, the meeting itself, media coverage and public reaction, any outcome including legislation or policies adopted as a direct outcome of the meeting. Sources often report on one of a series of meetings between two officials or delegations; in these cases, only the current meeting part of the topic, although planning for a future meeting that is a direct outcome of the current meeting and is discussed as part of the current meeting will be considered on topic.

12. Celebrity/Human Interest News, e.g., 31036: Joe DiMaggio Illness

Seminal events include: most often involves the death of a famous person or other significant life events like marriage, or a noteworthy tidbit about some regular person, like someone setting a world record or giving birth to septuplets.

Topic includes the specific event, causes (such as illness in the case of a celebrity's death) or consequences (such as a funeral or memorial service), public reaction or media coverage, editorials and opinion pieces, retrospectives or life histories that are a direct consequence of the seminal event.

13. Miscellaneous News, e.g., 31024: South Africa to Buy $5 Billion in Weapons

Seminal events include all specific events or activities that do not fall into one of the above categories.

Topic includes the event itself, direct causes and unavoidable consequences thereof".

TABLE 1: Example News Texts

Text id	Source	Date	Author	Text
1	guardian.co.uk,	Wednesday 11 May 2011 21.49 BST	AP	[headline] Deadly earthquakes strike Spain [sub-heading] • Two earthquakes strike Lorca in south-east Spain Two earthquakes have struck south-east Spain in quick succession, killing at least 10 people, injuring dozens and causing major damage to buildings. The epicentre of the quakes – about two hours apart – with magnitudes of 4.4 and 5.2 was close to the town of Lorca, an official with the Murcia regional government said. A hospital in the town was evacuated so dozens of injured people were treated at the scene where a field hospital was being set up. The Spanish prime minister's office put the death toll at 10 ….
2	www.independent.co.uk	Wednesday 11 May 2011	AP	[headline] 10 dead as earthquakes rock southern Spain Two earthquakes struck southeast Spain in quick succession today, killing at least 10 people, injuring dozens and causing major damage to buildings, officials said. The epicenter of the quakes — with magnitudes of 4.4 and 5.2 — was close to the town of Lorca, and the second came about two hours after the first, an official with the Murcia regional government said on condition of anonymity in line with department policy. The Murcia regional government said a hospital in Lorca was being evacuated, dozens of injured people were being treated at the scene and a field hospital was being set up.

Text id	Source	Date	Author	Text
				The Spanish prime minister's office put the death toll at 10…
3	bbc online	12 May 2011 Last updated at 12:07		[headline]Spain earthquake: Lorca residents assess damage Residents in the Spanish town of Lorca are assessing the damage from quakes that killed eight people and forced thousands to spend the night outdoors. The mayor of the historic town, with a population of 90,000, said: "Almost no-one slept in their homes". Some 20,000 buildings are believed to have been damaged in what was Spain's worst earthquake for 50 years. The magnitude 5.2 tremor hit early on Wednesday evening, around two hours after a quake measuring 4.4… [[NOTE THE BACKGROUND TO THE FOCAL EVENT]] Regional officials say at least 130 people have been injured Lorca's Mayor Francisco Jodar said most of the town's population had spent the night sheltering in their cars, streets, public squares or other towns. Many people were queuing at first light for food and hot drinks from emergency workers.. Some were returning to their homes to assess the damage, although many were ordered to keep away until a safety assessment of their buildings had been carried out.. “We are very scared, because ours [house] didn't collapse, but they are very damaged," one resident, Jose Crespo, said. "All the village has fallen, everything... All the buildings

Text id	Source	Date	Author	Text
				have been affected."
4	Bbc online	23 May 2011 Last updated at 14:20	By Michelle Roberts Health reporter, BBC News	[headline] Lung patients warned about new ash cloud from Iceland The ash cloud Poor air quality can trigger breathing problems Continue reading the main story Medical experts are advising people with lung conditions, such as asthma, to be prepared for the ash cloud that is expected to reach the UK on Tuesday. The British Lung Foundation is advising those who might be susceptible to carry their medication as a precaution. If the cloud from the Grimsvotn volcano in Iceland hits the UK, air quality could be significantly reduced, causing breathing problems for some people. But experts predict it will not be as disruptive as last year's eruption. The Eyjafjallajokull volcano's unusual ash size distribution, combined with unusual weather patterns, made life difficult across Europe during the late spring and early summer of 2010. 'Relatively tame' About 20 countries closed their airspace and it affected hundreds of thousands of travellers. By comparison, the impact of the Grimsvotn volcano looks relatively tame, according to University of Iceland geophysicist Pall Einarsson. The ash particles from this eruption are said to be larger than last year and, as a result, fall to the ground more quickly.

Text id	Source	Date	Author	Text
				But lung experts still advise precaution. Dr Keith Prowse, of the British Lung Foundation, said: "In light of the latest news that ash from the volcanic eruption in Iceland could reach the UK by Tuesday, we would advise people living with a lung condition in affected areas to carry their medication as a precaution." Erica Evans, of Asthma UK, said: "We know that volcanic ash can trigger asthma symptoms like coughing, wheezing and shortness of breath. However, as the ash is very high in the atmosphere it does not pose an immediate problem. Asthma UK advises people with asthma to monitor the news to see whether the ash cloud moves closer to the UK. "People with asthma should make sure they maintain their regular asthma medicine and keep their emergency inhaler on them at all times." Both charities say they can offer advice via a telephone helpline to anyone who may be concerned.
5	Bbc online	23 May 2011 Last updated at 15:43		UK flight disruption cannot be ruled out - CAA Disruption to UK flights cannot be ruled out, the Civil Aviation Authority says, as volcanic ash is set to reach parts of the UK by Monday evening. The ash cloud from Iceland's Grimsvotn volcano is expected by analysts to reach Scotland and Northern Ireland first. It did not necessarily mean airspace would be closed, the Met Office said. The event comes a year after ash from the Eyjafjallajokull volcano spread across Europe, causing huge disruption. 'Better prepared'

Text id	Source	Date	Author	Text
				Andrew Haines, chief executive of the CAA, said: "Our number one priority is to ensure the safety of people both onboard aircraft and on the ground. I think we are far better prepared and we'll have far better information and intelligence”.. "We can't rule out disruption, but the new arrangements that have been put in place since last year's ash cloud mean the aviation sector is better prepared and will help to reduce any disruption in the event that volcanic ash affects UK airspace." … Loganair is cancelling almost all its flights in Scotland on Thursday.. A Loganair spokesman said Met Office forecasts indicated that a high density of ash would be present in large parts of Scottish airspace throughout Tuesday, clearing into Wednesday morning. The UK's air traffic control service, Nats, said volcanic ash was forecast to affect parts of Scotland between 1800 BST and midnight on Monday… The Met Office, which runs Europe's Volcanic Ash Advisory Centre, earlier said there was a possibility of ash moving across the UK towards the end of the week. But a spokesman said the weather was much more changeable than at the time of last year's eruption and there was a lot more uncertainty. The Civil Aviation Authority (CAA) and Nats said they were monitoring the situation closely. The CAA said ash levels would be graded as low, medium or high, and airlines would be notified if levels reached medium or high. Airlines would then consider whether to fly…

Text id	Source	Date	Author	Text
				The Icelandic Met Office said the ash cloud could touch north-west Scotland on Monday evening, reaching about 20,000ft (6,100 metres) below the normal cruising altitude of commercial aircraft… A spokesman said ash at higher altitudes than this was moving north-west and towards Greenland. The Grimsvotn volcano began erupting on Saturday with ash rising to 20km (12 miles) but, although still active, is now not as powerful with a plume of 13km (8 miles). The cloud is expected to cover a vast crescent across the North Atlantic from northern Russia to the British Isles. During last year's eruption UK airspace was shut down completely by the authorities as a precaution, but this time airlines will make their own decisions about whether it is safe to fly. The National Airspace Crisis Management Executive is meeting every six hours to assess the situation. Looking better Icelandic air traffic control has created a no-fly zone… Icelandic Foreign Minister Ossur Skarphedinsson told the BBC:… "It remains to be seen what kind of damage there will be. This particular volcano has a history showing that it takes two to 10, even 14 days, to blow its fury whereas last year's volcano was much more lasting." Large particles The Grimsvotn volcano lies beneath the ice of the uninhabited Vatnajokull glacier in south-

Text id	Source	Date	Author	Text
				east Iceland. The latest eruption is its most powerful eruption in 100 years. However, University of Iceland geophysicist Pall Einarsson said the eruption was on a different scale to the one last year. "It is not likely to be anything on the scale that was produced last year when the Eyjafjallajokull volcano erupted," he said. "That was an unusual volcano, an unusual ash size distribution and unusual weather pattern, which all conspired together to make life difficult in Europe." The ash particles from this eruption are said to be larger than last year and, as a result, fall to the ground more quickly…
6	BBC	22 February 2011 Last updated at 11:56		[Headline] New Zealand earthquake: 65 dead in Christchurch New Zealand's prime minister says at least 65 people have died after a 6.3-magnitude earthquake hit Christchurch. John Key said the toll was expected to rise further, adding: "We may be witnessing New Zealand's darkest day." The tremor caused widespread damage as it occurred at a shallow depth of 5km (3.1 miles) during lunchtime when Christchurch was at its busiest. The mayor of New Zealand's second-biggest city says 120 people have been rescued from the ruins. The country's deadliest natural disaster in 80 years struck at 1251 (2351 GMT on Monday), 10km (6.2 miles) south-east of the city…

Text id	Source	Date	Author	Text
7	BBC	Page last updated at 13:13 GMT, Monday, 7 March 2011		Christ Church in Crewe could close Christ Church, Crewe Because the central part of the church has been pulled down, many assume it is closed. A south Cheshire church is facing closure because its appearance is keeping potential worshippers away. The congregation at Christ Church in Crewe attracts fewer than a dozen regular attenders. It is currently costing up to £25,000 a year to run. Canon Bill Baker, vicar at Christ Church, blamed its unusual shape, which is the result of the demolition of the main body of the building. The church is more than 160 years old, built for the town's railway workers.
8	http://www.irishtimes.com/newspaper/breaking/2011/0523/breaking1.html?via=mr	Last Updated: Monday, May 23, 2011, 16:13	Elaine Edwards	Obama hails strong ties between US and Ireland The bond between the United States and Ireland is not just one of trade and commerce, but carries a "blood lineage", US president Barack Obama said today. Mr Obama was speaking after he met Taoiseach Enda Kenny at Farmleigh this morning. The two leaders discussed a range of issues, including the EU-IMF bailout, the banking situation and unemployment. They also discussed the peace process in Northern Ireland and the consequences of

Text id	Source	Date	Author	Text
				Queen Elizabeth II's visit here last week. Mr Obama and his wife Michelle arrived in Dublin this morning for a one-day visit to Ireland at the start of a week-long European tour. The pair arrived in Moneygall, Co Offaly by helicopter at 3.10pm and spent about 20 minutes meeting people in the village before visiting his ancestral home. They spoke to locals over pints of Guinness in Ollie Hayes' Bar. Mr Obama said "sláinte" before taking a healthy gulp of his drink and insisted that the president "always pays his bar tab". Mr Obama is expected to arrive back in Dublin for an open-air concert and rally between 5.30pm and 6pm at College Green. This public event is free - no tickets are required, and people can access the concert area from 2pm via security barriers at the intersection of Parliament Street and Dame Street. Westlife, Imelda May and Jedward, along with a number of actors and sports stars, are set to warm up the crowd before the president's speech. The Air Force One jet touched down shortly before 9.30am at Dublin airport and the Obamas were transferred to the Phoenix Park by helicopter, where they met President Mary McAleese at Áras an Uachtaráin. Mr Obama then travelled to Farmleigh for a meeting with the Taoiseach. Mr Kenny said the two leaders had discussed issues such as the use of Shannon airport by the US and the role of Irish peacekeepers in Afghanistan. Speaking after the 40-minute meeting, Mr Obama said he was "extraordinarily grateful" for the welcome he and his wife had received from the Taoiseach and the Irish people. He said the friendship and bond between the United States and Ireland "could not be stronger".

Text id	Source	Date	Author	Text
				"Obviously it is not just a matter of strategic interests. It's not just a matter of foreign policy, for the United States and Ireland carries a blood lineage," he said. "For millions of Irish-Americans this continues to symbolise the homeland and the extraordinary traditions of an extraordinary people." Noting that he and Mr Kenny had already had an opportunity to meet in Washington, he said he was glad to see progress on a number of issues, which the Taoiseach was “more than up to the task of achieving”. Mr Obama said the US wanted to help strengthen the bonds of trade and commerce between the two countries, and to do everything it could to help Ireland on the path to recovery. “Ireland is a small country but punches above its weight on a range of issues,” he said. He noted Ireland's strong voice in areas such as human rights, and also its work within the EU. On progress in Northern Ireland, Mr Obama said it spoke “to the possibilities of peace and people in longstanding struggles being able to reimagine their relationships”. He noted the “mutual warmth and healing” that accompanied the visit of the Queen here last week, which sent a signal not just here in Ireland but around the world, he said. “It sends what Bobby Kennedy once called a ripple of hope.” Mr Obama paid tribute to all those who had “worked tirelessly” to bring about peace in Northern Ireland. The president said he was proud of the part that America had played in getting both sides to talk and to provide a space for that conversation to take place. This morning’s meeting was also attended by Mr Gilmore. Initially billed as a courtesy call, it was upgraded to an official bilateral meeting.
9	http://www.newsletter.co.uk/news/local/obama_ha	Monday 23 May 2011 08:59		Obama hails Ulster peace on Republic visit US President Barack Obama has hailed the Northern Ireland peace process as an

Text id	Source	Date	Author	Text
	ils_ulster_peace_on_republic_visit_1_2704857			example to the rest of the world, during a visit to the Irish Republic. Mr Obama was speaking following a meeting with Irish taoiseach Edna Kenny at the Irish state guesthouse at Farmleigh House in Phoenix Park, Dublin. Commencing a week-long tour of Europe, the president touched down in the Republic on Monday morning prior to a state visit to the UK on Tuesday. Later, Mr Obama, joined by his wife Michelle, will visit his ancestral home of Moneygall, Co Offaly, and address an open air rally in Dublin city centre. Following talks with Mr Kenny, the president said: “I want to express to the Irish people how ... inspired we have been by the progress made in Northern Ireland because it speaks to the possibilities of peace and people in long-standing struggles to be able to re-imagine their relationships.” Reflecting on the Queen’s historic four-day state visit to the Republic last week, he added: “To see Her Majesty, the Queen of England, come here and to see the mutual warmth and healing that took place as a consequence of that visit, to know that the former taoiseach Dr (Garret) FitzGerald was able to witness the Queen coming here, that sends a signal not just in England, not just here in Ireland but around the world.” Mr Obama was earlier officially welcomed to the Republic by Irish president Mary McAleese. On Monday afternoon, the US president will be taken by helicopter to Moneygall. His great-great-great-great-grandfather was a shoemaker in the village and his son, Falmouth Kearney, left for New York in 1850. Residents queued for up to six hours on Thursday to secure a ‘golden ticket’ to see his homecoming. It will involve a trip down Main Street to the Kearney ancestral home, where he will be greeted by John Donovan, the owner of the house, and his family. The president and first lady will then visit Ollie Hayes’ pub to meet extended family members.

Text id	Source	Date	Author	Text
				Returning to Dublin, Mr Obama is due to address an open air College Green audience at the end of entertainment involving many well-known Irish artists. The president will fly out of Dublin on Tuesday to start his UK state visit.
10	http://www.hellomagazine.com/celebrities-news-in-pics/23-05-2011/56549/	23-05-2011		'Thrilled' Barack Obama returns to his Irish roots "We are thrilled to be here," President Barack Obama said as he greeted Ireland's president Mary McAleese on Monday. Despite the grey skies looming over the Emerald Isle, the 49-year-old added: "The sun's coming out – I can feel it.". First Lady Michelle Obama accompanied her husband for the trip to Ireland which saw him return to his ancestral home. Barack visited the village of Moneygall, where his maternal great-great-great-grandfather was born, and was welcomed by 3,000 people. He also met a distant cousin, whom he greeted with a big hug. Earlier in the day, he held private talks in the drawing room of the Aras with the Irish Prime Minister, Taoiseach Enda Kenny, before posing with a hurling stick he received as a present (pictured). This seven-day trip will also see the couple visit Britain, Poland and France.
11	BBC NEWS	11 May 2011 Last updated at 14:23	Megan Lane	Who, what, why: Can earthquakes be predicted? In Italy, Asia and New Zealand, long-range earthquake predictions from self-taught forecasters have recently had people on edge. But is it possible to pinpoint when a quake will strike?

Text id	Source	Date	Author	Text
				It's a quake prediction based on the movements of the moon, the sun and the planets, and made by a self-taught scientist who died in 1979. But on 11 May 2011, many people planned to stay away from Rome, fearing a quake forecast by the late Raffaele Bendandi - even though his writings contained no geographical location, nor a day or month. In New Zealand too, the quake predictions of a former magician who specialises in fishing weather forecasts have caused unease. After a 6.3 quake scored a direct hit on Christchurch in February, Ken Ring forecast another on 20 March, caused by a "moon-shot straight through the centre of the earth". Rattled residents fled the city. Predicting quakes is highly controversial, says Brian Baptie, head of seismology at the British Geological Survey. Many scientists believe it is impossible because of the quasi-random nature of earthquakes. "Despite huge efforts and great advances in our understanding of earthquakes, there are no good examples of an earthquake being successfully predicted in terms of where, when and how big," he says. Many of the methods previously applied to earthquake prediction have been discredited, he says, adding that predictions such as that in Rome "have little basis and merely cause public alarm". Woman holding pet cat in a tsunami devastated street in Japan Can animals pick up quake signals? Seismologists do monitor rock movements around fault lines to gauge where pressure is building up, and this can provide a last-minute warning in the literal sense, says BBC science correspondent Jonathan Amos.

Text id	Source	Date	Author	Text
				"In Japan and California, there are scientists looking for pre-cursor signals in rocks. It is possible to get a warning up to 30 seconds before an earthquake strikes your location. That's enough time to get the doors open on a fire station, so the engines can get out as soon as it is over." But any longer-range prediction is much harder. "It's like pouring sand on to a pile, and trying to predict which grain of sand on which side of the pile will cause it to collapse. It is a classic non-linear system, and people have been trying to model it for centuries," says Amos. In Japan, all eyes are on the faults that lace its shaky islands…