

Acquisition and Use of Multilingual Name Dictionaries

Bruno Pouliquen, Ralf Steinberger & Camelia Ignat
European Commission – Joint Research Centre
21020 Ispra (VA), Italy
{Name.Surname}@jrc.it, <http://langtech.jrc.it>

Abstract

We are presenting a method and a working system that automatically builds up a large multilingual dictionary of person and organisation names through daily news analysis and that makes use of this name dictionary – together with a gazetteer of location names and other means – to link related news articles across languages for 19 languages. Prominent features of the system are the simplicity of the approach (required to extend the functionality to so many languages), the fact that monolingual and cross-lingual name variants are automatically merged with the name's base form, and the fact that the system aggregates information about persons independently of the spelling of their name. The system, accessible online at <http://press.jrc.it/NewsExplorer/>, has currently collected over 630,000 different names with up to 140 variants for the same name from real life news, plus their inflections. We will put this work into the wider context of other text-related activities carried out at the European Commission's Joint Research Centre (JRC).

Keywords

Dictionaries; Named Entity Recognition; Name Variants; Multilinguality.

1. Introduction

It is known that multilingual machine-readable dictionaries – both generic and specialist – are needed for many Language Technology applications. These applications include Machine Translation, cross-lingual Information Retrieval, cross-lingual glossing (Ignat et al. 2005), term-highlighting in specialist texts, relevance-ranking of documents for specific subject areas, and more. The EU-funded projects *Multext*¹ and *Multext-East*² had the purpose of addressing this need and to create various basic multilingual resources.

Proper names are not normally included in dictionaries. The reason for this is that, firstly, there is a seemingly infinite number of person, organisation and location names and new names are being created all the time. Secondly, it is often implicitly assumed that proper names are not normally translated, except probably across writing systems (Roman, Cyrillic, Arabic, Greek, etc.). This assumption does indeed hold for unknown places (for example, there is no foreign name equivalent for the small Northern-Italian town of *Ispra*) and the exact same spelling of a name like

George Bush can be found across many languages. However, for larger or historically known places like the Italian city of *Venezia*, many translations exist: This Italian name has the English translation *Venice*, German *Venedig*, French *Venise*, Greek *Βενετία*, Spanish *Venecia*, Russian *Венеция*, Dutch *Venetië*, Czech *Benátky*, Arabic البندقية [Al Bunduqī], etc. In a small number of European languages, person names are transliterated (e.g. the name *George Bush* is *Džordž Buš* in Serbian, *Džordžs Bušs* in Latvian and *Corc Buş* in Azerbaijani). However, there are many more reasons why multiple variants of person names can be found, even within the same language and sometimes even within the same document. These reasons include (Steinberger & Pouliquen 2007):

- Morphological variants such as added suffixes (e.g., in Slovene, *Tonyem Blairem* may be found for the name *Tony Blair*);
- Spelling mistakes (e.g. *Condoleza Rice*, *Condaleezza Rice*, *Condollezza Rice*, *Condeleezza Rice*, all found more than once in real news texts);
- Adaptation of names to local spelling rules (e.g. the German name *Schröder* is frequently found as *Schroder* in English language press because the letter 'ö' does not exist in English);
- Transliteration differences due to different transliteration rules or different target languages (e.g. the same Russian name *Владимир Устинов* may be transliterated as *Wladimir Ustinow* in German and as *Vladimir Ustinov* in English);
- In the specific case of Arabic, where short vowels are usually not written, vowels need to be inserted during transliteration, which can cause large numbers of variants (e.g. the Arabic name محمد consists of only the consonants *Mhmd*, which explains the different Romanised variants *Mohammed*, *Mahmoud*, *Muhamad*, etc.);
- The reuse of name parts to avoid repetition (e.g. *Condoleezza Rice*, *Ms. Rice* and *Secretary of State Rice*): these are not part of the name dictionary.

In this paper, we present the publicly accessible NewsExplorer system, which automatically acquires large, multilingual lists of person and organisation names plus their variants from texts in 19 languages and uses these person dictionaries, together with multilingual gazetteers, to link related news across 19 languages and to collect and aggregate

¹ <http://aune.lpl.univ-aix.fr/projects/multext/>

² <http://nl.ijs.si/ME/>

gate information about people and organisations independently of the name spelling.

The following sections will present the Europe Media Monitor (EMM) news data we work with (Sect. 2), describe the method to acquire name dictionaries in many languages automatically (Sect. 3). We will focus on the fact that the methods used need to be simple and easily extendable to many languages and on the specific process of identifying whether newly found names are actually new persons or whether they are mere variants of a known name. Section 4 then shows how the dictionaries of person, organisation and location names are used to compute the similarity of news articles across many languages. Section 5 will put this work into the context of other Language Technology-related work going on at the European Commission's (EC) Joint Research Centre (JRC). Section 6 concludes and points to future work.

2. The Europe Media Monitor multilingual news data

NewsExplorer (Steinberger et al. 2005) is part of the *Europe Media Monitor* (EMM) family of applications (Best et al. 2005). NewsExplorer receives from EMM an average of 35,000 news articles in 32 languages, scraped from more than 1,000 news portals in Europe and around the world. EMM converts all articles into a standard UTF-8-encoded RSS format and classifies them into a given set of several hundred categories. The articles received within the last few minutes and hours are displayed on the live NewsBrief website (<http://press.jrc.it>), which is updated every ten min-

utes. NewsExplorer clusters related articles once per day, separately for each language, in order to group news about the same event or subject. From each of these clusters, person, organisation and location names are extracted (see Section 3). The information on the entities found in each cluster, combined with the database information on names and name variants and further information, will later be used to link equivalent news clusters across languages (see Section 4). A database keeps the information where and when each name was found, which other names were found in the same cluster, and which name attributes were found next to the name.

3. Building up the multilingual name dictionary

We will now describe the method used to recognise known names and their variants in each cluster (3.1), and how to guess potential new names (3.2 and 3.3). We will show how online sources are exploited to extend the known names with additional variants (3.5) and how we identify whether a newly found name is a variant to a known name already stored in the name database (3.6).

3.1 Lookup of known named entities

The NewsExplorer database currently contains over 630,000 names plus about 135,000 name variants. We feed a FLEX finite state automaton (Paxson 1995) with the 50,000 most frequent known names and their variants. Additionally, we generate regular expressions so that frequent spelling variants will be recognised, as well. These include:

<p>156. Mahmoud Ahmadinejad 2007-08-28T23:10+0200</p> <p>Mahmoud Ahmadinejad / Mahmoud Ahmadinedschad / محمود احمدی نژاد / Mahmoud Ahmadinejad / Mahmud Ahmadinejad / Mahmoud Ahmadinejad / محمود احمدی نژاد / محمود احمدی نژاد / محمود احمدی نژاد / Mahmoud Ahmadinejad / Махмуд Ахмадинеджад / Mahmud Ahmadeschad / محمود احمدی نژاد / احمدی نژاد / احمدی نژاد / Mahmoud Ahmadinejad / Mahmoud Ahmadinejad / Mahmoudas Ahmadinejadas / Mahmud Ahmedinejad</p> 	<p>2007-08-29T14:07+0200 . M. Bush invoque le spectre d'un "holocauste nucléaire"... LeMonde (fr) ...heures auparavant, le président iranien, [Mahmoud Ahmadinejad], avait donné corps à certaines des mena... Devant des anciens combattants, mardi 28 août, à Reno (Nevada), le président américain a déclaré que l'Iran est le "premier Etat du monde en matière de soutien au terrorisme". Le président iranien propose, lui, de "comblér le vide" en Irak".</p> <p>2007-08-29T14:02+0200 . القوات الأمريكية تفرح عن 7 إيرانيين اعتنقهم اللبنة الماسية [bbc-arabic] (ar) ... في العراق، أما نظيره الإيراني [محمود أحمدی نژاد] فقال في وقت سابق إنه يدعو الأمريكيين في الـ "عراق" و... تزايد التوتر بين واشنطن وطهران بخصوص الوضع في العراق، وطهران تقول إن القوات الأمريكية اعتنقت 7 خبراء إيرانيين يعملون في إصلاح محطات الكهرباء العراقية، وأقرحت عندهم لاحقاً.</p> <p>2007-08-29T09:51+0200 . G.W.Bushas perspėja dėl branduolinio holokausto Artimuosiuose Rytuose [rytas] (lt) ... pat G.W.Busho kalbą Irano prezidentas [Mahmoudas Ahmadinejadas] šaipėsi iš JAV atakos idėjos, o naujo... JAV prezidentas George'as W.Bushas antradienį perspėjo dėl branduolinio holokausto Artimuosiuose Rytuose, jei Izraelio priešas Iranas įsigys branduolinių ginklų, ir pažadėjo neleisti, kad taip nutiktų.</p> <p>2007-08-29T13:35+0200 . Rüstung: Bush zieht Parallele zwischen Holocaust und Irans Atomprogramm [spiegel] (de) ...erhängen&quot;. Der iranische Präsident [Mahmud Ahmadinedschad] hatte mehrfach mit der Vernichtung Isr... Reno - Irans Streben nach der Atombombe "droht eine ohnehin schon für Instabilität und Gewalt bekannte Region in den Schatten eines nuklearen Holocausts zu stellen", sagte Bush vor Veteranen in Reno im US-Bundesstaat Nevada. "Wir werden gegen diese Gefahr angehen, bevor es zu spät ist.</p> <p>2007-08-29T13:27+0200 . Irán da por zanjado su expediente ante la AIEA [jornada] (es) ...O Teherán. [Mahmud Ahmadinejad], mandatario iraní, afirmó hoy que el expediente atómico de la república Islámica "está cerrado". El presidente George W. Bush, por su parte, pidió una vez más a Irán que detenga de inmediato sus actividades nucleares.</p> <p>2007-08-29T13:21+0200 . Махмуд Ахмадинеджад изпрати поздравление до Гюл [actualno] (bg) ... Иранският президент [Махмуд Ахмадинеджад] изпрати поздравително послание до Абдуллах Гюл... Иранският президент Махмуд Ахмадинеджад изпрати поздравително послание до Абдуллах Гюл по повод избирането му за президент на Република Турция, предаде ДПА, като се позовава на иранската информационна агенция ИРНА. Ахмадинеджад посочи в посланието си, че е сигурен, че отношенията между двете съседни...</p>
<p>143. Moghtada Sadr 2007-08-28T23:05+0200</p> <p>Moghtada al-Sadr / Muqtada al-Sadr / Muqtada Sadr / مقتدى الصدر / Muqtada al Sadr / Al Sadr / Moktada Sadr / Muqtada Al Sadr / Muqtada</p> 	

Figure 1. Screenshot from the publicly accessible live site <http://langtech.jrc.it/entities/>, showing the person names and their variants plus the text snippets in which the name was found since midnight CET. The example of *Mahmoud Ahmadinejad* shows that – even in such a short time period – a large number of spelling variants and morphological variants can be found. The screenshot shows texts with different orthographies in French, Arabic, Lithuanian, German, Spanish and Bulgarian.

- Hyphen/space alternations: for hyphenated names such as *Jean-Marie* or *Nawaf al-Ahmad al-Jaber al-Sabah*, we generate alternation patterns (`Jean[\ \-]Marie`);
- Diacritic alternations: Words with diacritics are frequently also found without their diacritics, e.g. *Émile Lahoud* may be found as *Emile Lahoud*; we thus generate the pattern (`É|E`)`mile[\]Lahoud`;
- Declensions of names: we pre-generate morphological variants for all known names in the languages that decline person names (e.g., for Balto-Slavonic and Finno-Ugric languages, see Przepiórkowski 2007). In Slovene for example, we can find the following declensions of the name *Javier Solana*: *Javierja Solane*, *Javiera Solane*, *Javierom Solano*, *Javierjem Solano*, *Javierja Solano*. The simple rules to pre-generate morphological variants are hand-written, but inspired by empirical evidence, i.e. the most frequently found suffix variants are used (for details, see Pouliquen et al. 2005). The suffix replacement rules are rather simplistic and they may over-generate (produce inexistent word forms), but this is not usually a problem because these will simply not be found in real text.

The complete regular expression generated for the name *José Ramos-Horta* is then:

```
Jos(é|e)(e|a|o|u|om|em|m|ju|jem|ja)?[\ ]
Ramos(e|a|o|u|om|em|m|ju|jem|ja)?[\ \-]
Hort(e|a|o|u|om|em|m|ju|jem|ja)?
```

Figure 1 shows how large a variety of names and their declensions can be found within a few hours.

3.2 Guessing unknown named entities

While many names can be found in the news every day, there are always new names. In order to identify these unknown names, a relatively light-weight language-independent procedure was developed. It is relatively easy to extend to new languages by providing language-specific resources for this language. These resources contain lists of known first names, frequent titles (Mr., Chancellor, etc.) and some other lexical patterns. The first names include the most frequent names found for various languages (e.g. *John*, *Jean*, *Hans*, *Giovanni*, *Johan*, etc.). The patterns can consist of titles (e.g. *Minister*), words indicating nationality (e.g. *German*), age (e.g. *32-year old*), occupation (e.g. *playboy*), a significant verbal phrase (e.g. *has declared*), and more. We refer to these patterns generically as *trigger words*. The trigger words in the language-specific resource file are listed as strings (Minister, Head of State, American) or patterns (“death of”, [0-9]+-year-old, etc.).

The name guessing software will identify any sequence of at least two uppercase words as a name if either one of the name parts is a known first name (*John XXX*) or if the word sequence is accompanied (to the left or to the right) by one

or more trigger words. Combinations of trigger are frequent and are also captured by the system, e.g.:

Somali-born Dutch politician Ayaan Hirsi Ali.

A number of additional features have been implemented in order to raise the performance of the system. For instance, *name stop words* are used to avoid identifying frequent uppercase words (e.g. *Also*, *Friday*) as part of the name. Frequent name parts such as *von*, *van der*, *de la*, *bin*, *abu*, etc., which can be written in lowercase, are also accepted as part of the name. For languages that do not distinguish case (Arabic, for example) we add additional rules to recognise the sequence of consecutive words that are most likely to be a name and to determine where the name stops. For that purpose, we make use of frequent surnames and of frequent verbs, etc. as name stop words. The process does not make use of part-of-speech or other linguistic information in order to keep the rules and the process simple and so that it can easily be extended to many languages. For a more detailed description of the name guessing software, see Steinberger & Pouliquen (2007).

Language-specific resources are currently in use for twelve languages (Danish, German, English, Spanish, Estonian, French, Italian, Dutch, Norwegian, Portuguese, Slovene and Swedish). Resources also exist for Arabic, Bulgarian, Farsi, Polish, Turkish, Romanian and Russian, but are not yet fully integrated in the system.

3.3 Empirically enlarging trigger word lists

Building first lists of trigger words and expressions is not difficult because lists of first names and of professions can

Table 1. Some of the top-ranking items (log-likelihood-ranked) of an automatically generated list of trigger word candidates for Romanian. The first column indicates the frequency of this word combination in the context of known names.

<i>Occurrences</i>	<i>Title</i>	<i>Translation</i>
77	PSD	Romanian Social Democratic Party
91	premier	Premier
57	israelian	Israeli
64	britanic	British
62	premierului	Premier
61	german	German
55	rus	Russian
33	palestinian	Palestinian
60	preşedintele	President
31	Sir	Sir
	...	
34	Premierul	premier
	...	
41	a declarat	has declared
33	Preşedintele american	American president
13	Liderul PSD	PSD leader
9	ministrul de interne	Interior minister

Figure 2. NewsExplorer entry showing some of the name variants (first column) collected for the Afghan president Hamid Karzai. The associated trigger words (column 2) give the user additional information about this person (in this case mostly: *Afghan President*). Further information displayed on this dedicated person page includes: the last news clusters this person was mentioned in, quotations by and about this person, as well as other people frequently mentioned together with this person. The information has been collected in the course of years from news articles in many different languages. Hamid Karzai’s page is available at <http://press.jrc.it/NewsExplorer/entities/en/49.html>.

Names	Key Titles and Phrases
Hamid Karzai (Eu,vi)	afghan president (en - 1066)
Hamid Karsai (de)	président afghan (fr - 411)
Hamid Karzai (en,pt)	presidente afgano (es,it - 322)
Hamid Karzaj (sl)	president (de,sv - 1745)
حامد کرزاي (ar,fa)	afghaanse president (nl - 235)
Hamida Karzaja (sl)	presidente afegão (pt - 283)
Hamidom Karzajem (sl)	presidente afghano (it - 241)
Hamed Karzai (en,it)	präsident (de - 625)
Ahmid Karzai (it)	präsidenten (de - 280)
حامد کرزاي (ar)	presidente (es,pt - 353)
Hamid Karza (es,pt)	homologue afghan (fr - 46)
Hamid Kharzai (en,it)	afganistanski predsednik (sl - 41)
Хамид Карзай (bg,ru)	président (fr - 245)
Jamid Karzai (es)	afganistanskim predsednikom (sl - 24)
Hamid Karzai (en)	afghanske præsident (da - 22)
حامد قرزاي (ar)	afgano (es,it - 33)
חמיד קרזאי (he)	afghan (en,fr - 36)
Ahmid Karzai (fr)	- afghan president (en - 19)
Хомид Карзай (tg)	afganistanskega predsednika (sl - 46)
Hamis Karzai (es)	
Hámid Karzai (sk)	

Latest Clusters - English	
[it] [es] [fr] [nl] [pt] [de] [da] [sl] [tr] [sv]	German hostage frx cnn 20-AUG-07
Iran seeks foreign oil investment fr 20-AUG-07	Suicide attack kills RTERadio 18-AUG-07
German woman kidnapped in Kabul bbc 19-AUG-07	US pledges to work fr 16-AUG-07
Bomber kills Afghan district head bbc 17-AUG-07	

be compiled easily from open source documents (see for example <http://en.wikipedia.org/wiki/Category:Occupations> and especially the links to other languages).

To enlarge such manually compiled language-specific resources, we use a bootstrapping technique to exploit news corpora to capture the most frequent patterns: First, we look up all known names in a news corpus. We then produce frequency lists of left and right-hand-side contexts of these known names and manually accept or reject the patterns found. Table 1 lists the top-ranking proposals for Romanian trigger words found in a corpus of one year of Romanian news. The expert judgment on acceptance or refusal is necessary to avoid entering terms such as *PSD* (abbreviation of a political party) as a trigger word and to replace it by a compound regular expression such as *Liderul [A-Z]+* instead.

In many languages, there are gender variants or morphological variants of trigger words, such as *fireman/firewoman* in English, *sénateur/sénatrice* in French,

senator/senatorja/senatorju/senatorjom/senatorjem in Slovene, etc. As our tool does not make use of lemmatisation or any other linguistic process, we list all the word forms as regular expressions. In the mentioned Slovene case, this would be *senator(ja|ju|jom|jem)?*. In the automatically proposed list in Table 1, the expert can easily recognise *premierul* and *premierului* as morphological variants of *premier*.

An obvious alternative way to improve trigger words for other languages would be to translate a trigger word list from one language (or even from several) into another, either automatically or manually. However, when translating these words out of context, this could generate some errors. For instance, the English trigger word *Palestinian* can be translated into German as the adjective *palästinensisch* (not a trigger word) or as the noun *Palästinenser* (correct trigger word).

The bottom-up bootstrapping technique has the advantage that the most frequent trigger words will be found. For instance, in Spanish, the term *lendakari* (leader of the autonomous government of the Basque Country) was found.

Yet another alternative to find name recognition patterns or to recognise unknown names would be to use Machine Learning (ML) approaches. Although we use ML for other tasks, we have opted not to use it for person name recognition. The reason is that we feel that we will produce the language-specific resources for a new language faster with the bootstrapping procedure than with a ML approach.

3.4 Name knowledge base

A database keeps track of all the historical information for each name occurrence: date and language of the text, information on the cluster where it was found, the name variants used and the trigger words that appear next to the recognised name. Some of this information is then aggregated into a NewsExplorer person page (one webpage for each of the hundreds of thousands of persons of the knowledge base). Figure 2 shows part of such a dedicated person page, showing the name variants for Afghan President Hamid Karzai, the multilingual trigger words (mostly titles) and the latest clusters in which he was mentioned. The information displayed was collected in the course of years and from news in many different languages.

3.5 Adding name variants from web sources

For known names, the system periodically checks whether a corresponding entry exists in the Wikipedia online encyclopaedia (Wikipedia 2007). If it exists, the URL of the Wikipedia page will be added to the NewsExplorer page to provide readers with more information on the person, if required. Additionally, the system downloads the picture for the person or organisation and checks whether Wikipedia lists name variants not yet known to the NewsExplorer

database. Additional variants found are then added to the knowledge base. This is particularly useful for name variants in foreign scripts (Arabic, Russian, Chinese, etc.). Figure 3 shows that some interesting name variants can be found for the Afghan President Hamid Karzai.



Figure 3. Examples of name variants automatically collected from the relevant Wikipedia page.

3.6 Merging Name Variants

Every day, NewsExplorer detects an average of 450 unknown person names. For each of these, we need to check whether they are really new names – in which case a new database entry should be created – or whether they are mere variants of a known name – in which case they should be added as a variant to the known name. On average, 50 of the 450 new names can safely be added as variants to a known name. An additional average of 42 names per day are kept for manual verification. The dedicated person pages on NewsExplorer are updated every day.

To compare each new name with almost one million known names and their variants in the database, an approximate matching algorithm is used. In order to make this process computationally tractable, we first pre-select merger candidates (see Section 3.6.1) and then we use an approximate name matching algorithm to compare the pre-selected names with all known names (see 3.6.2). The pre-selection is done by first *normalising* (simplifying) each name. If the normalised forms of the new name and that of any known name (or any of its variants) are identical, the new name will be considered to be a merger candidate.

3.6.1 Normalisation of person names

The first normalisation step consists of applying standard transliteration rules for names written in scripts other than

the Roman alphabet. This will allow us to compute the similarity between names even if they are not written in the same script:

- Cyrillic - Russian, Bulgarian and Ukrainian - (Симеон Маринов => Simeon Marinov);
- Greek (Κώστας Καραμανλής => Kostas Karamanlis);
- Arabic (جلال طالباني => jlal tlbani - “Jalal Talabani”) including some additional transliteration rules for Farsi;
- Devanagari - Hindi, Nepalese - (सोनिया गांधी => sonya gandhi).

The second step consists of normalising the name orthography. We lowercase the name, eliminate diacritics (*François Chèreque* will be replaced by *francois chereque*) and reduce two neighbouring identical consonants to single consonants (*Mohammed Atta* becomes *mohamed ata*). We then apply a further set of about thirty manually compiled normalisation rules. These rules are motivated by observations on frequent variations between names collected from the multilingual corpus, e.g.:

- the German name-initial ‘Wl’ and the name-final ‘ow’ for Russian names (as in Wladimir Ustinow) will get replaced by ‘Vl’ and ‘ov’;
- the Slovene ‘š’, the German ‘sch’, the French ‘ch’ will get replaced by ‘sh’ (as in Bašar al Assad, Baschar al Assad, Bachar al Assad);
- the French ‘ou’ (as in Oustinov) will get replaced by ‘u’;
- the ‘x’ gets replaced by ‘ks’, etc.

The third step consists of deleting the vowels. This is compulsory when dealing with languages that do not write all vowels (Arabic, Farsi etc.) or with names originating in such a language.

We would like to stress that these normalisation rules are exclusively driven by pragmatic needs and have no claim to represent any underlying linguistic concept.

As a result the names are represented by the main consonants forming their names. The names *ألكسندر سلطانوف* [alksndr sltanuf], *Александр Салтанов* [Aleksandr Saltanov], *Alexander Saltanow*, *Alexandr Saltanov* and *Alexander Saltanov* will all have the same normalised form: *lksndr sltnv*. The normalised form for each of the known names and their variants is stored in the database and will be used for the comparison with any new name.

3.6.2 Similarity measure

All new names whose normalised form is identical with any of the normalised forms of a known name or its variants are pre-selected name variant candidates. For these candidates only, we use a similarity measure to identify whether each candidate is possibly a variant of any of the

Table 2. Some variants of person names automatically merged during our daily process (where the similarity is above 0.94) and merger candidates waiting for evaluation by an expert.

<i>Name 1</i>	<i>Name 2</i>	<i>Similarity</i>	<i>Merged?</i>	<i>Same? person?</i>
<i>Barzan al-Tikriti</i>	<i>Barzan al Tikriti</i>	<i>0.99</i>	<i>Yes</i>	<i>Yes</i>
<i>Ismail Hanieh</i>	<i>Ismail Hanyieh</i>	<i>0.98</i>	<i>Yes</i>	<i>Yes</i>
<i>Farouq al-Qaddoumi</i>	<i>Farouk al-Kadoumi</i>	<i>0.97</i>	<i>Yes</i>	<i>Yes</i>
<i>Abdullah bin Abdul Aziz</i>	<i>Abdullah bin Abdel Aziz</i>	<i>0.96</i>	<i>Yes</i>	<i>Yes</i>
<i>Barzan al-Tikriti</i>	<i>Barazan al-Takriti</i>	<i>0.94</i>	<i>Yes</i>	<i>Yes</i>
<i>Manfred Wörner</i>	<i>Manfred Werner</i>	<i>0.93</i>	<i>No</i>	<i>No</i>
<i>Michel Ancel</i>	<i>Michael Ancel</i>	<i>0.92</i>	<i>No</i>	<i>Yes</i>
<i>Jorge Costa</i>	<i>Jorge Acosta</i>	<i>0.92</i>	<i>No</i>	<i>No</i>
<i>Falon Gong</i>	<i>Falun Gong</i>	<i>0.90</i>	<i>No</i>	<i>Yes</i>
<i>Roberto Panella</i>	<i>Roberto Pianelli</i>	<i>0.87</i>	<i>No</i>	<i>No</i>
<i>Peter Struck</i>	<i>Peter Starck</i>	<i>0.82</i>	<i>No</i>	<i>No</i>
<i>Jamie Foxx</i>	<i>Jaime Foxx</i>	<i>0.77</i>	<i>No</i>	<i>Yes</i>

known names. This process is explained in detail in Steinberger & Pouliquen (2007).

For each candidate, the edit distance (Zobel & Dart 1995) is applied twice, each time to a different representation of the same name pair: First it is applied to the normalised form *with vowels* and then to the lower-cased non-normalised name, as it was found in the text (or its transliteration in case of foreign script). The first similarity has a relative weight of 0.8, the second of 0.2. If the combined similarity value is above the threshold of 0.94, the candidates are automatically merged. Otherwise the new name is entered into the name database with a new identifier. Note that the newly identified name is merged with existing names if it is similar enough to *any* of the stored name variants. The threshold of 0.94 was determined empirically by looking at large numbers of merger candidates. In this test set, all name variant candidates with a similarity above 0.94 were indeed good candidates (i.e. the precision was 100%). By lowering the threshold, some more good candidates could have been found, but we would merge some name variants that do not belong to the same person. The focus of the merging is thus on high precision rather than on good recall. Table 2 shows a few name merger candidates and their combined similarity values. The shown example set includes candidates above the threshold (automatically merged – the expert agreed that they were all correct) and below (kept separate – the expert sometimes confirmed that they were two different persons, sometimes that they should be merged).

At irregular intervals, the topmost similar merger candidates *below* the threshold of 0.94 are being looked at to determine whether they should be marked as variants of a known name.

4. Using recognised entities to compute similarity across languages

In addition to displaying the collected and aggregated information on names in NewsExplorer, the name dictionaries are used for a second, more original purpose: it is used – together with other dictionaries and means – to link related news clusters across languages. This will be explained in 4.2, following a short overview of cross-lingual document similarity calculation in general (4.1).

4.1 Cross-lingual document similarity calculation

There is a restricted number of options to identify automatically whether two documents written in different languages are equivalent or not (in the case of news: whether they talk about the same theme or event). The most common approach is to use Machine Translation (MT) to translate one document into the language of the other and to apply monolingual similarity measures such as those based on a vector space representation (e.g. Leek et al. 1999). An alternative, but similar method consists of using bilingual dictionaries and to use dictionary entries found in both documents as anchors (Wactlar 1999). A fundamentally different approach consists of generating a bilingual vector space by using chunks of parallel text to achieve cross-lingual word associations. This has been tried using Lexical Semantic Analysis (LSA, Landauer & Littman 1989) and Kernel Canonical Correlation Analysis (KCCA, Vinokourov et al. 2002). All of these approaches have in common that they require bilingual resources (MT systems, dictionaries, bilingual lists of lexical associations). When dealing not with two languages, but with many, there is an explosion of required resources. In the case of NewsExplorer, for example, which currently processes news in 19 languages, there are 171 language pairs: $((n^2 - n)/2)$ with n being 19.

4.2 Using name dictionaries as anchors for cross-lingual linking

To avoid this complexity when linking news in so many languages with each other, we are using yet another approach, which works particularly well for news. News is mostly about people, organisations, places, subjects or events at a given time. If news articles in different languages are published on the same day and talk about roughly the same persons, places and subjects, one can reasonably assume that they are equivalent or at least related. Names of persons, organisations and locations can thus act as *anchors* to link news articles or clusters across

languages. As the surface forms of names can differ across languages (e.g. *Bush* and *Buš*, *Venezia* vs. *Venice*, etc.), the names first need to be standardised. For person and organisation names, the standardisation can easily be done via the multilingual name dictionary described earlier: each name and its variants is represented by a unique identifier. For locations and subject domains, additional work is necessary. We will now briefly summarise how to extract place and subject domain anchors from news clusters and we will then describe how they are used to link documents across languages.

In order to use the mentions of *geographical place names* in the multilingual news articles as anchors, a multilingual gazetteer is needed. We compiled such a gazetteer from various sources: the Global Discovery database (2007), the Estonian multilingual name database KNAB (2007), and some European Commission-internal documents. This gazetteer is not complete (there are at least half a million places world-wide), but it includes the biggest and most frequently mentioned places in most of the languages of interest. However, even when such a gazetteer is available, the cross-lingual comparison is not straight-forward, because many place names are homographic with other places (there are 32 places called *Washington*, 15 places each called *Berlin* and *Paris*, 244 places called *Aleksandrovka*, etc.). They can also be homographic with person names (there are places called *Victoria*, *Tony*, *Blair*, *Annan*, etc.) or with common words (*And*, *To*, *Be*, *By*, etc. are all names of places). Pouliquen et al. (2006b) describe an approach

that attempts to solve these ambiguities. The result is a non-ambiguous list of places represented by their numerical place name identifier, as well as by their Latitude and Longitude values. We use this disambiguated list as input for the cross-lingual linking process.

Unlike persons, organisations and locations, events or *themes of news* cannot be represented by entities. In order to get a language-independent representation of the subjects of each news cluster, we classify news clusters according to the multilingual Eurovoc (2007) thesaurus. Eurovoc exists in over twenty languages and distinguishes about 6,000 different subject areas. It was developed for parliaments to categorise mostly legal documents manually, but it can still be used to generate an abstract representation of the themes of a news cluster. Pouliquen et al. (2003) describes how an existing collection of manually categorised parliamentary documents was exploited to build automatic multi-label classifiers and how that software can be used to produce a ranked list of the most relevant Eurovoc classes for a text. In NewsExplorer, we use the 100 most relevant Eurovoc classes to produce a subject domain vector.

For the cross-lingual similarity calculation, we then use a cluster representation consisting of four ingredients (see Figure 4). The overall similarity is a linear combination of the weights of each feature overlap (cosine of vectors).

- Subject domain representation via the vector of Eurovoc classes;

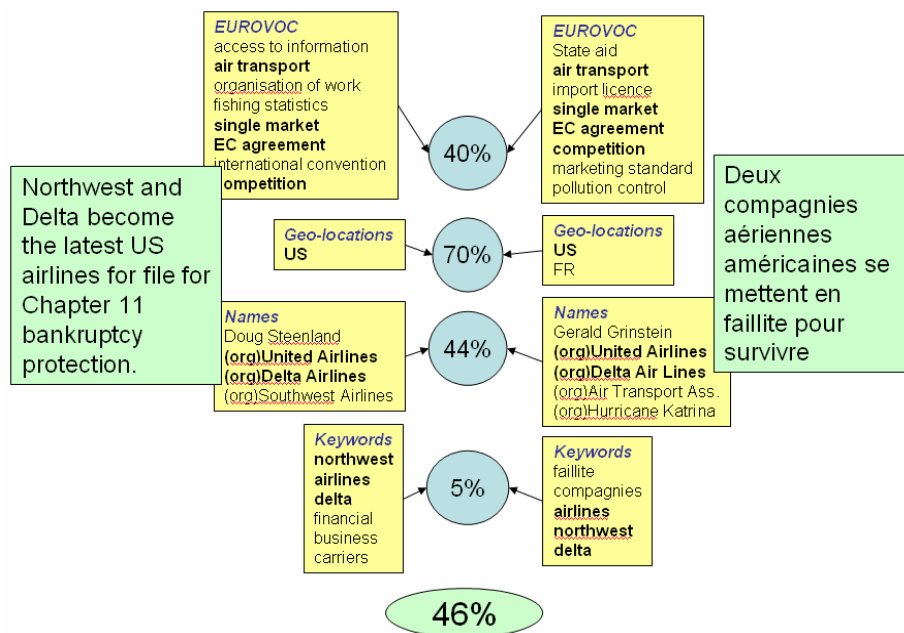


Figure 4. Combination of four ingredients to calculate the document similarity between an English and a French document.

- Normalised sum of references to countries in the text: each direct (country name) or indirect (city name) reference to a country is summed up and log-likelihood-normalised by the average mention of that country in a large reference corpus of news);
- Combined frequency list of persons and organisations;
- Monolingual weighted list of words (it is common that texts in different languages share words, such as street or project names, cognates such as *Tsunami*, etc.).

These four ingredients are combined with relative weights of 40%, 30%, 20% and 10%. Half of the weight for the cross-lingual linking of news clusters is thus based on name dictionaries (persons, organisations and names).

In the approach to cross-lingual document similarity calculation adopted in NewsExplorer, similarity also needs to be established separately for each language pair. However – unlike in the approaches mentioned in Section 4.1 – no language pair-specific resources are needed. Instead, the news clusters of all languages are represented in the same way.

5. Further text-related JRC activities

On request, we will now summarise further related work going on at the European Commission’s Joint Research Centre. Above, we described technology to acquire and use name dictionaries in the EMM **NewsExplorer** application, which takes as input the 35,000 news articles per day in 33 languages gathered by the EMM **NewsBrief** application. NewsBrief (Best et al. 2005) additionally classifies the news into about 600 categories, displays statistics on the categories most active at any moment, clusters live news, detects breaking news and sends email and SMS alerts to notify subscribed users in case a big event happens and is picked up by the media. The third product in the EMM family of applications is the *Medical Intelligence System MedISys* (Fuart et al. 2007; Yangarber et al. 2007), of which a restricted public version is available at <http://medusa.jrc.it>. MedISys takes all EMM articles plus documents from about 150 specialist medical websites as input, selects only the articles of relevance to Public Health authorities, classifies them according to hundreds of health-related categories such as diseases and disease sub-types (e.g., respiratory infections), bioterrorism-related issues, toxins, bacteria (e.g., anthrax), viral hemorrhagic fevers (e.g., Ebola), viruses, medicines, water contaminations, etc. The MedISys website presents a quantitative summary of latest reports visually and informs subscribed users via instant emails or via daily, automatically generated summary reports.

In addition to these products, the EMM team has developed a number of further applications, which do not make use of Language Technology, but which are much appreciated by the users inside the European Commission and elsewhere.

These applications, described briefly below, are summarised in Best et al. (2005). The European Commission maintains a network of representations and delegations who review the media of their respective host countries and manually submit daily categorised summaries using the **PressReview** system. **Rapid News Service** (RNS) is a tool that allows users to view all manually submitted or automatically collected news documents in a single interface, to filter, group and edit them, to forward them to key personnel via email or SMS in case of major news reports, and to easily compile printable twice-daily newsletters. These are called **EMM Panorama** and are factually in-house up-to-date newspapers. For EMM news, a **WAP service** has been set up that allows stakeholders to read any urgent news item on their mobile phones.

EMM’s Language Technology group has downloaded a collection of legal and other EU documents from the EUR-Lex websites (<http://eur-lex.europa.eu/>) and has used it to construct the **JRC-Acquis** (Steinberger et al. 2006). The JRC-Acquis is a multilingual parallel corpus in currently 22 languages with paragraph or sentence alignments for all 231 language pairs. With a total of over 1 Billion words, the JRC-Acquis is the largest parallel corpus, when taking into consideration the number of languages. It is freely available for download at <http://langtech.jrc.it/JRC-Acquis.html>. In collaboration with the EC’s *Directorate General for Translation*³ (DGT), the JRC is furthermore preparing a dump of DGT’s Translation Memory – for a similar document collection and the same language pairs. This resource will soon be distributed for research purposes via the JRC’s website under the name of **DGT-TM**.

In NewsExplorer, as of August 2007, links between persons or organisations are calculated on the basis of *co-occurrence* of their names in the same news clusters (Pouliquen et al. 2006). Various **relation extraction** efforts are currently being undertaken to *specify* the type of relationship that holds between persons. In Pouliquen et al. (2007a), we describe the effort to produce social networks on the basis of co-occurrence of person names in selected news collections restricted by time, language or subject area. The effort described in Pouliquen et al (2007b) focuses on identifying *quotations* in the news in eleven languages and to detect who makes reference to whom in direct speech. Tanev (2007) detects the specific binary relations *contact* and *support* that hold between persons, as found in English language news. *Criticise*, *family relationship* and other relations are under development. For that purpose, he has developed a Machine Learning method that learns linguistic patterns in a bootstrapping method.

In a parallel effort, lexical patterns for **event extraction** are being learnt by bootstrapping from English language news

³ <http://ec.europa.eu/dgs/translation/>

in the specific field of violent events (Best et al., forthcoming). A working, but not yet online system detects violent events in incoming news, detects event type, number and type of victims, actors, time and place of the event (where available), and visualises the events on a continuously updated map. These automatically learnt patterns are currently being re-written in **ExPRESS** (Piskorski 2007), a fast JRC-developed formalism to write linguistic extraction rules. Finally, the usage of an **ontology** is currently being explored to derive additional information from event templates that have been automatically extracted and manually verified (Oezden Wennerberg et al. 2007). Finally, various approaches are being undertaken to visualise information extracted from the news.

6. Conclusion & Future work

We have tried to show the usefulness of name dictionaries (for persons, organisations and locations) for at least two purposes: (1) the aggregation of information on persons and organisations from news sources in currently 19 languages, and (2) the linking of related news clusters across many languages by using person, organisation and location names. We have also shown how large-scale multilingual name dictionaries can be acquired automatically both from multilingual news collections and from external sources such as Wikipedia.

In the future, we would like to work on exploiting our multilingual name database with Machine Learning methods to *predict* name variants of known names in various target languages.

In the news of different languages, the same entities are usually mentioned in the same time period. Shinyama & Sekine (2004) refer to this phenomenon as the “synchronicity of names”. We plan to use this distribution of the time line as an additional important feature to merge names. We should additionally be able to exploit the existing cross-lingual cluster links.

Our Named Entity Extractor is still quite weak (mixed-language F-measure is about 87.5% for person names - see Steinberger 2007). We would like to use Machine Learning techniques to improve the name recognition performance.

Until recently we have concentrated on merging name variants. A challenging problem is now to disambiguate various homographic names (*John Adams* is a common name referring to various different persons⁴) or names with very similar spelling such as *Yasser Arafat* and *Yasir Arafat* (one is the Palestinian ex-leader, the other is a famous Pakistani cricketer). Solutions exist (see for example

Pedersen 2006) but the accuracy is currently too low to disambiguate fully automatically.

7. Acknowledgements

The development of NewsExplorer, which creates and uses the name dictionaries discussed in this paper, is a group effort, to which many developers have contributed in the course of the years. We would like to thank the entire EMM team and especially the group leader Clive Best and the lead developer Erik van der Goot for providing the raw multilingual news data and for making a robust system available to tens of thousands of users every day. We also thank the many computational linguists who helped us develop the language-specific resources for the many different languages over time.

8. References

- [1] Best Clive, Erik van der Goot, Ken Blackler, Teofilo Garcia, & David Horby (2005). *Europe media monitor—system description*. Technical Report EUR 22173 EN.
- [2] Best Clive, Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger & Hristo Tanev (forthcoming). *Automatic Event Extraction for the Security Domain: Techniques and Applications*. In: Security and Intelligence Informatics.
- [3] Fuat Flavio, David Horby & Clive Best (2007). *Disease Outbreak Surveillance Through the Internet – the MedISys Project*. Proceedings of European Federation for Medical Informatics Special Topic Conference Medical Informatics in Enlarged Europe. Brijuni Islands, Croatia, 30.5.-1.6.2007.
- [4] Global Discovery (2006). *Digital world reference map*. Europa Technologies. Available at <http://www.europa-tech.com/gd.htm> (last visited 28.03.2007).
- [5] Ignat Camelia, Bruno Pouliquen, Ralf Steinberger & Tomaž Erjavec (2005). *A tool set for the quick and efficient exploration of large document collections*. Proceedings of the Symposium on Safeguards and Nuclear Material Management. 27th Annual Meeting of the European Safeguards Research and Development Association (ESARDA-2005). London, UK, 10-12 May 2005.
- [6] KNAB (2006). *Place Name Database of EKI*. Institute of the Estonian language, Tallinn, available at <http://www.eki.ee/knab/knab.htm> (last visited 28.03.2007).
- [7] Landauer Thomas & Michael Littman (1991). *A Statistical Method for Language-Independent Representation of the Topical Content of Text Segments*. 11th International Conference Expert Systems and Their Applications, vol. 8: 77-85. Avignon, France.
- [8] Leek Tim, Hubert Jin, Sreenivasa Sista & Richard Schwartz (1999). *The BBN Crosslingual Topic Detection and Tracking System*. In 1999 TDT Evaluation System Summary Papers. <http://www.nist.gov/speech/tests/tdt/tdt99/papers> [7.04.2006]
- [9] Oezden Wennerberg Pinar, Hristo Tanev, Jakub Piskorski & Clive Best (2007). *Ontology-based Analysis of Violent Events*. In: Proceedings of Intelligence and Security Infor-

⁴ http://en.wikipedia.org/wiki/John_Adams_%28disambiguation%29 on Wikipedia lists 34 persons sharing this name (last visited 28/08/2007)

- matics (ISI'2007). New Brunswick, New Jersey, USA, 23-24 May 2007.
- [10] Paxson Vern (1995). *Flex – Fast Lexical Analyzer Generator*. Lawrence Berkeley Laboratory, Berkeley, CA. Available on <http://flex.sourceforge.net/#downloads> (last visited 28.08.2007).
- [11] Pedersen Ted, Anagha Kulkarni, Roxana Angheluta, Zornitsa Kozareva & Tamar Solorio (2006). *An Unsupervised Language Independent Method of Name Discrimination Using Second Order Co-occurrence Features*. Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), pp. 208-222. February 19-25, 2006, Mexico City.
- [12] Piskorski Jakub (2007). *EXPRESS - Extraction pattern recognition engine and specification suite*. Proceedings of the 6th International Workshop on Finite-State Methods and Natural Language Processing (FSMNL'2007). Potsdam, Germany, 14-16 September 2007.
- [13] Pouliquen Bruno, Ralf Steinberger, Camelia Ignat, Irina Temnikova, Anna Widiger, Wajdi Zaghouni & Jan Žižka (2005). *Multilingual person name recognition and transliteration*. Journal CORELA - Cognition, Représentation, Langage. Numéros spéciaux, Le traitement lexicographique des noms propres. Available online at: <http://edel.univ-poitiers.fr/corela/document.php?id=490>.
- [14] Pouliquen Bruno, Ralf Steinberger, Camelia Ignat & Tamara Oellinger (2006a). *Building and displaying name relations using automatic unsupervised analysis of newspaper articles*. Proceedings of the 8th International Conference on the Statistical Analysis of Textual Data (JADT'2006). Besançon, 19-21 April 2006.
- [15] Pouliquen Bruno, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuat, Wajdi Zaghouni, Anna Widiger, Ann-Charlotte Forslund, Clive Best (2006b). *Geocoding multilingual texts: Recognition, Disambiguation and Visualisation*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 53-58. Genoa, Italy, 24-26 May 2006.
- [16] Pouliquen Bruno, Ralf Steinberger, Jenya Belyaeva (2007a). *Multilingual multi-document continuously updated social networks*. Proceedings of the Workshop Multi-source Multilingual Information Extraction and Summarization (MMIES'2007) held at RANLP'2007. Borovets, Bulgaria, 26 September 2007.
- [17] Pouliquen Bruno, Ralf Steinberger, Clive Best (2007b). *Automatic detection of quotations in multilingual news*. Proceedings of the International Conference *Recent Advances in Natural Language Processing* (RANLP'2007). Borovets, Bulgaria, 27-29 September 2007.
- [18] Przepiórkowski Adam (2007). *Slavic Information Extraction and Partial Parsing*. Proceedings of the ACL Workshop on Balto-Slavonic Natural Language Processing. Prague, June 2007
- [19] Shinyama Yusuke & Satoshi Sekine (2004). *Named Entity Discovery Using Comparable News Articles*. 20th International Conference on Computational Linguistics (CoLing). Geneva, Switzerland.
- [20] Steinberger Ralf & Bruno Pouliquen (2007). *Cross-lingual Named Entity Recognition*. In: Satoshi Sekine & Elisabete Ranchhod (eds.). *Linguisticae Investigationes* LI 30:1, pp. 135-162. Special Issue Named Entities: Recognition, Classification and Use.
- [21] Steinberger Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga (2006). *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006), pp. 2142-2147. Genoa, Italy, 24-26 May 2006.
- [22] Tanev Hristo (2007). *Unsupervised Learning of Social Networks from a Multiple-Source News Corpus*. Proceedings of the Workshop Multi-source Multilingual Information Extraction and Summarization (MMIES'2007) held at RANLP'2007. Borovets, Bulgaria, 26 September 2007.
- [23] Vinokourov, A., Shawe-Taylor, J., Cristianini, N. (2002). *Inferring a semantic representation of text via cross-language correlation analysis*. *Advances of Neural Information Processing Systems* 15, 2002.
- [24] Wactlar H.D. (1999). *New Directions in Video Information Extraction and Summarization*. In Proceedings of the 10th DELOS Workshop, Sanorini, Greece, 24-25 June 1999.
- [25] Wikipedia: The free encyclopaedia (2007). FL: Wikimedia Foundation, Inc. Retrieved August 22, 2007, from <http://www.wikipedia.org>.
- [26] Yangarber Roman, Clive Best, Peter van Etter, Flavio Fuat, David Horby & Ralf Steinberger (2007). *Combining Information about Epidemic Threats from Multiple Sources*. Proceedings of the Workshop Multi-source Multilingual Information Extraction and Summarization (MMIES'2007) held at RANLP'2007. Borovets, Bulgaria, 26 September 2007.
- [27] Zobel Justin & Philip Dart (1995). *Finding approximate matches in large lexicons*. *Software – Practice and Experience*, Vol. 25(3), pp. 331-345.