

Multilingual multi-document continuously-updated social networks

Bruno Pouliquen, Ralf Steinberger & Jenya Belyaeva

European Commission – Joint Research Centre

Via Enrico Fermi 1, 21020 Ispra (VA), Italy

{Bruno.Pouliquen, Ralf.Steinberger}@jrc.it, Jenya.Belyaeva@ext.jrc.it

Abstract

We are presenting a fully-automatic live online system (accessible at <http://langtech.jrc.it/SocNet>) that produces monolingual or mixed-language social network graphs showing which groups of persons are being mentioned together in the world news of the last few hours. The basis for this system are name mentions extracted automatically from an average of 35,000 news articles per day in 32 languages. For any given person on the graph, hyperlinks lead to the list of text snippets and to the original texts where the person was mentioned, plus to a dedicated webpage containing additional information about this person gathered in the course of several years. For any link between persons, hyperlinks lead to the list of text snippets and to the full texts where both persons are mentioned. Building multilingual social networks that even cross writing systems (Arabic, Greek, Chinese, etc.) is made possible by exploiting the name database built up by the multilingual online NewsExplorer system (Steinberger et al. 2005), which automatically associates name variants to the same person identifier. We also discuss differences between live social networks generated from the news in different languages for the same time period.

Keywords

Social Networks, multilinguality, multi-document summarisation, Named Entity Recognition, name variant merging, visualisation.

1. Introduction

To a large extent, the factual part of news is about themes or events (taking place at certain locations at a certain time) and about persons or organisations. The news analysis system NewsExplorer (Steinberger et al. 2005; accessible at <http://press.jrc.it/NewsExplorer>) tries to give views of the news from the axes *events* (news clusters), *locations*, *named entities* (mainly persons and organisations) and *time* (via time lines, i.e. historical linking of news). In addition to linking news via these entities and axes, news items in NewsExplorer are also linked across languages. In this paper, we present an additional way of allowing users access to news: we present live social networks, i.e. graphs displaying groups of persons that are frequently mentioned together in the news of the last few hours and up to 1 day. Probably the most interesting aspect of the presented approach is the high multilinguality of the system (32 languages) and the fact that names are linked across languages (and writing systems) even if spelt differently and when the names have been inflected. Users can view the multi-document, multilingual and cross-language live system at the site <http://langtech.jrc.it/SocNet>. Additionally to the most

recent multilingual social networks displayed at that site, it is also possible to produce social networks separately by language or by the country of origin of the news, as well as for documents covering a specific theme. These customised social networks are not accessible to the public, but in this paper we compare the multilingual networks with monolingual networks in four languages (section 5).

This social network generation tool takes as input the *Europe Media Monitor* (EMM; Best et al. 2005) news data and makes furthermore use of the following technology: (a) multilingual name recognition software, (b) approximate name matching software that identifies name variants for the same person, (c) multilingual language-dependent morphological name inflection generation software, and (d) network generation and visualisation software. Tools (a), (b) and (c) are part of the NewsExplorer system, which analyses news every day, links news over time (topic detection and tracking) and across languages (cross-lingual topic tracking), extracts new and known names, collects information about people and visualises the results in various ways.

The 12 co-occurrence graphs visible at the above-mentioned site are updated every two hours. Graph production starts completely anew every 24 hours at midnight so that users will always see the social network graphs of world-wide news of today. Information found in the news of all 32 languages are fully aggregated and all the results are visualised together.

Section 2 points to work with a similar focus. Section 3 summarises the text analysis technology underlying the social network generation. Section 4 focuses on the network generation, size reduction and visualisation. In Section 5, we discuss the network generation results, comparing the mixed language network with various monolingual networks for a sample 8-hour snapshot for Friday 13 July. Section 6 concludes the paper and points to future work.

2. Related Work

Due to the large volume of various types of information on the internet, there are now various applications that try to produce person profiles and to exploit similarities for various purposes (e.g. to provide focused advertising, to provide meeting forums, etc.). Some social network services like *LinkedIn* (LinkedIn 2007) or *MySpace* (MySpace 2007) build and verify online social networks, connecting registered users by different types of interests

(company, country, research interests, etc.). The features used for the linking are typically user-provided. To our knowledge, the only tools that extract the underlying linking features fully automatically are called *Connivence Maps* by *Pertinence Mining* (Connivence 2007; based on English and French news) and *Silobreaker*, based on *Elucidon* software (Silobreaker 2007, English only), but the producers do not say how their technology works and it is not even clear whether the networks are manually edited.

For related work on individual components of the presented system (Named Entity Recognition, name variant matching, dealing with highly inflected languages, etc.), see Steinberger & Pouliquen (2007).

3. The underlying news data and text analysis technology

The social networks under discussion are extracted from live news, using resources on person names and their spelling variants. In this section, we briefly summarise where the news data comes from (section 3.1), how person names have been extracted across many languages and over years to build a name database of currently 615,000 names (3.2), how spelling variants for the same name have been gathered and merged automatically (3.3) and how morphological inflections of known names are being recognised in Balto-Slavonic and other highly inflected languages (3.4). Section 4 will then explain how this data is used to produce live social networks.

3.1 Gathering the news data

The JRC's *Europe Media Monitor* system (Best et al. 2005) gathers an average of 35,000 news article per day in 32 languages, by continuously monitoring about 1,100 public news sites from around the world for newly published information. All new articles are downloaded, converted to the standard UTF-8-encoded XML news format RSS, full-text indexed and classified according to themes and the countries mentioned in the text. The result is published in the *EMM-NewsBrief* site (<http://press.jrc.it>), which is updated every ten minutes.

3.2 Multilingual Named Entity Recognition

For 19 of the 32 languages, the related *EMM-NewsExplorer* application (<http://press.jrc.it/NewsExplorer>, Steinberger et al. 2005) clusters all articles gathered during the previous day by similarity in order to group all articles about the same subject or event. For all clusters, references to geographical places, to persons and organisations are identified, using finite state automata to recognise known names and regular expressions to recognise new names or name variants (recognition of *new* names in 14 languages only: Da, De, En, Es, Et, Fr, It, Nl, No, Pt, Ro, Sl, Sv, Tr). Sequences of uppercase words are identified as being a name if they contain known first names or if they are surrounded by empirically collected lexical patterns consisting 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*. Name *stop words* are used to exclude identifying frequent uppercase words (e.g. *Monday*) as part of the name. For a detailed description of this process, see Steinberger & Pouliquen (2007). The process does not make use of part-of-speech or other linguistic information in order to keep the process simple and so that it can easily be extended to many languages.

3.3 Name variant matching and merging

For all unknown names found during the daily analysis, an approximate string matching algorithm checks whether the name is likely to be a variant of a known name or whether it is a new name. New names are added to the database with a new identifier. Names found in at least five different news clusters are added to the list of known names. Periodically, a search on Wikipedia (Wikipedia 2007) is carried out to gather name translations that can be found there, as well as photographs. Wikipedia is especially useful to find name transliterations in languages using different scripts, such as Asian languages or languages using the Cyrillic, Arabic or Hellenic scripts.

The approximate string matching algorithm to compare newly found names with the 615,000 known names and their 143,000 known variants (status July 2007) is a multi-step process, details of which are described in Steinberger & Pouliquen (2007). To avoid a performance bottleneck when comparing each of several hundred new names per day with close to a million known names and name variants, we first apply a name normalisation step. Only if the normalised new name is identical with a normalised name (or any of its variants) in the database, we apply the edit distance approximate matching algorithm (Zobel & Dart 1995) to two different name representations: once to the normalised name form and once to the normalised name form with the vowels removed. If the average similarity for the new and the known name are above an empirically set threshold, the two names will be classified as variants of each other. Otherwise, the new name will be added to the database as a new name.

The name normalisation rules eliminate diacritics, reduce two neighbouring identical consonants to single consonants, unify frequent spelling variants across languages, etc. For instance, 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' and the German 'sch' will get replaced by 'sh'; French 'ou' (as in *Oustinov*) will get replaced by 'u', etc. These normalisation rules are exclusively driven by pragmatic needs and have no claim to represent any underlying linguistic concept.

An average of 400 new person names are automatically recognised as part of the NewsExplorer text analysis every day. The NewsExplorer database keeps track of all name mentions plus the list of trigger words (the titles and phrases) they are associated with.

Lang	NewsPaper	Snippet
sl	vecer	glavnega osumljenca za umor Aleksandra Litvinenka v Londonu postavili pred
sl	vecer	v ponedeljek zavrnil izrocitev Andreja Lugovoja , da bi ga kot glavnega
tr	sabah	öldürülen eski KGB ajani Alexander Litvinenko 'nun davasi, İngiltere-Rusya
tr	sabah	cinayetin zanlisi olarak istedigini Andrei Lugovoy 'u Rusya'nin iade etmemesi
en	dailytimesPK	suspected of killing Kremlin critic Alexander Litvinenko in London last year,
en	dailytimesPK	when British prosecutors alleged that Andrei Lugovoi used a rare radioactive
pt	DiariodeNoticias	assassinio do ex-oficial do KGB Alexander Litvinenko . A revelação foi feita
pt	DiariodeNoticias	acederia ao pedido de extradição de Andrei Lugovoi (outro ex-agente do KGB)
en	taipeitimes	Kremlin following its refusal to extradite Andrei Lugovoi , the former KGB....
en	taipeitimes	KGB agent suspected of murdering Alexander Litvinenko last November.
en	eirepost	Lugovoi over the murder of Alexander Litvinenko , describing the decision
en	eirepost	Russia's refusal to extradite Andrei Lugovoi over the murder of Alexander
sl	delo	in nekdanjega tajnega agenta KGB Andreja Lugovoja . London - Britanija in
sl	delo	in ostrega Putinovega kritika Aleksandra Litvinenka , ki je bil nekoc prav tako
en	rian	- Russia considers the Alexander Litvinenko case a purely criminal matter,
en	rian	Moscow has refused to extradite Andrei Lugovoi , a former Kremlin bodyguard,

Table 1. Text snippets in newspapers of various languages showing both the names *Alexandre Litvinenko* and *Andrei Lugovoi*.

3.4 Dealing with morphological inflection

The current list of *known names*, i.e. the names that were found in at least five independent news clusters, consists of approximately 50,000 names plus 135,000 variants. These names can be identified in text of any language through a simple lookup procedure, i.e. no lexical patterns are required. This works well for languages with little morphological proper noun variation (e.g. most Western European languages, Arabic, Bulgarian, etc.). However, for Balto-Slavonic, Finno-Ugric and other languages, looking up the base form of a name will yield poor results as the names will not be found when they are inflected. For instance, Estonian *Bushiga* and Slovene *Bushom* are both inflections of *Bush*. Table 1 shows some morphological variants of the names *Alexander Litvinenko* (e.g. *Litvinenka*, *Litvinenko'nun*) and *Andrei Lugovoi* (*Lugovoja*, *Lugovoy'u*). As acquiring or developing morphological resources for all 32 EMM languages is out of our reach, we use relatively simple, hand-crafted paradigm expansion rules that generate – for each of the known names and their variants – a number of morphological variants. These rules, described in more detail in Pouliquen et al. (2005), either add various name endings to the same name or they substitute endings to generate a set of new endings. For the name of the Secretary-General of the Council of the European Union *Javier Solana*, for instance, we generate various inflection forms so that the strings *Javierja Solane* (sl), *Javierom Solanom* (sk), *Javierem Solana* (pl), *Javierjem Solano* (sl), *Javiera Solany* (pl) will all be found and identified as variants of *Javier Solana*.

These morphological paradigm extension heuristics do not solve all problems, but the most frequent morphological variants can normally be captured and over-generated (wrong) variants are not harmful as they will

simply not be found. For the lookup procedure, we use FLEX (Paxson 1995) to produce a finite state automaton. This tool is useful for the efficient lookup of large name lists including character-level regular expressions for suffixes, etc. It also allows looking up person names in languages that do not use white space to separate words, such as Chinese.

4. Social network generation and visualisation

The input for the work on social network generation consists of a stream of all incoming EMM news articles (32 languages), in which references to known persons have been marked up using the finite state automaton described in Section 3.4. Names not previously known are not recognised in the live system, but the list of known names is updated every day.

For efficiency purposes, we build a constantly updated index that records, for each recognised name and for each pair of names, all 300-character text snippets around the names. Table 1 shows multilingual text snippets for the names *Alexander Litvinenko* and *Andrei Lugovoi*. The index is reset at midnight every day so that it always contains the latest news and name mentions. We plan to turn this index into a 24-hour rolling window from which articles older than 24 hours will get deleted. This will give more consistency to the networks shown and will be more useful for users living in different time zones.

The index is read every two hours to update the graphs.

4.1 Turning links into a network

When two names are mentioned in the same article, a *link* is created between these two names. The more frequently the two names are mentioned, the stronger the link. In the input example shown in Table 1, the tool will thus build a

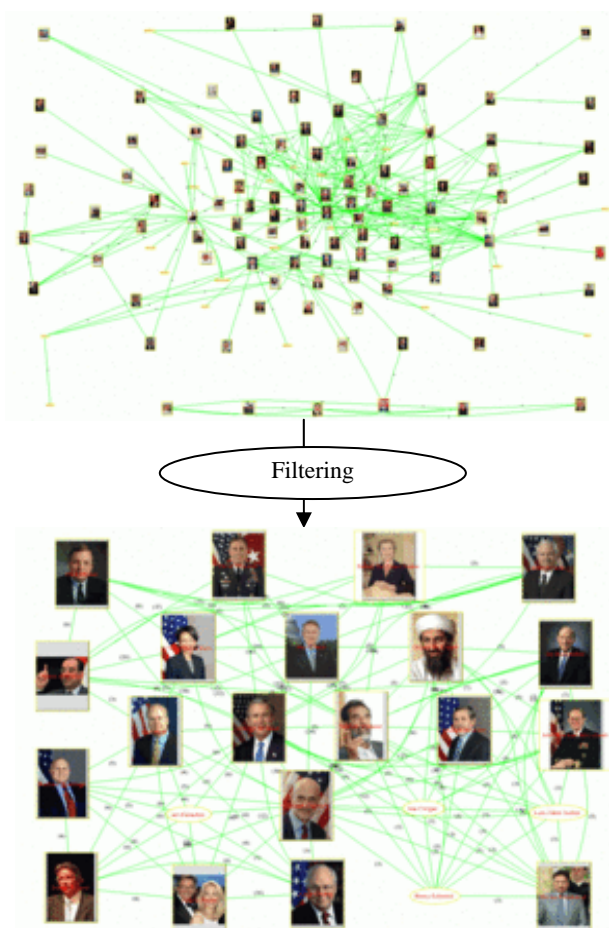


Figure 1. Example of the result of graph filtering, we retain only persons having lots of links to others.

link of weight (strength) 8 between the two persons Alexander Litvinenko and Andrei Lugovoi because they are mentioned together in eight different documents. We also have the possibility to filter graphs by language, country or subject area. This can be useful for users wanting to analyse news articles for a very specific domain. When considering only English documents, the link between Litvinenko and Lugovoi in Table 1 would be 4.

Any set or subset of links can be used to build a graph with edges (links) and vertices (persons, nodes). When considering the co-occurrence relationship under discussion, links are obviously non-directional, but in other types of relationships, it may be necessary to show the direction of relations. In the *criticism* and *support* relationship extracted by Tanev (2007), for instance, links need to be directed because such relationships are not necessarily bidirectional, and in the quotation relationship (person A refers to person B in direct speech; Pouliquen et al. 2007), users may also want to see who makes reference to whom. These more specific types of relationships and the resulting social networks are not yet incorporated in the public version of NewsExplorer.

Each sub-graph contains connected persons, but these persons may be connected indirectly, i.e. if person A is

linked to person B and person B to person C, A and C will occur in the same sub-graph. The presence or non-presence of drawn edges indicates whether the link is direct or indirect.

4.2 Reducing the size of a network

The network of links can grow rather big in any given 24-hour period. For all languages together, we find approximately 50,000 links per day. In order to reduce the amount of information to the stronger links, we reduce the size by setting a threshold on the number of links required for each name pair. Depending on the number of articles, we set this threshold to between 1 (for instance for graphs fed by single languages with less articles) and 4 (for the multilingual graph fed by all 32 EMM languages).

We use a simple algorithm to divide the graphs into sets of sub-graphs, which may be connected or not. Using the previously mentioned threshold, the network of all links can be cut down into sub-graphs, i.e. graphs are automatically separated if they either have no links or if the link strength is lower than the threshold. However, if there are links above the threshold, both graphs will be joined into one.

Towards the end of a 24-hour period, the graphs are often illegibly large (see, for instance, the first graph in Figure 1). This might be an indication that 12-hour windows may be more appropriate.

For practical reasons, we display only the first 12 biggest sub-graphs. For visualization purposes, we further reduce the size of graphs if the total number of persons (nodes, or vertices) is larger than 120. We do this by deleting those persons having only one link and loop until no more vertices can be deleted. If the remaining number of edges is above 140, we remove the persons having more than two edges. If the remaining number of edges is more than 160, we delete the ones having three edges, and so on, according to the following algorithm:

```

threshold=1; min=100;
While (numberOfVertices > min+20*threshold)
{
  Do {
    DeleteVertexHavingLessThan(threshold);
  } until (noVertexDeleted);
  threshold += 1;
};

```

4.3 Visualising the network

The graphs are displayed using *GraphViz* (GraphViz 2007), and more specifically the *Neato* tool which computes automatically the best place for each vertex to be displayed. We additionally make use of our database of images downloaded from Wikipedia. GraphViz allows for various output formats. We have chosen *gif + html ImageMap* as the output format, as it allows us to visualize the network on any web browser.

This simple visualisation does not allow to zoom or to change the angle of view, but it allows us to provide two types of hyperlinks: (1) When clicking on the link between two persons, the user is shown the latest text snippets in which the related persons have been mentioned

Table 2. Evaluation of person name recognition in various languages. The second column shows the number of manually identified person names.

Language	No. of Names	Precision	Recall	F-1
English	405	92	84	88
French	329	96	95	95
German	327	90	96	93
Spanish	274	85	84	84
Italian	298	92	90	91
Russian	157	81	69	74

together. Such a list of text snippets is shown in Figure 3. (2) When clicking on a node (a photograph or a person name), the user sees the most recent text snippets surrounding the name. Figure 2 shows the text snippets surrounding the name *Alexander Litvinenko* (the Russian spy who was murdered in London). This allows users to see in what context the person was mentioned. From this page, two additional hyperlinks lead to the original article where the text snippet was found and to the dedicated *NewsExplorer* person page where the user can see additional background information about this person collected over the last few years: a list of name variants, name attributes (lists of trigger words), lists of associated persons, the latest articles about this person, quotations by and about this person, and more.

5. Evaluation of the results, discussion

There is no obvious method to evaluate our automatically generated social networks quantitatively. What can be evaluated quantitatively are Precision and Recall of the Named Entity Recognition (NER) and – as we merge name variants for the same person – the results of the merging algorithm. Such results have been described in Steinberger & Pouliquen (2007) for the languages English, French, German, Spanish, Italian and Russian, repeated in Table 2 for the reader’s convenience.

All persons that have been found to be mentioned together in a news article are linked somehow. The useful-

ness of such a link cannot be evaluated quantitatively, but is by definition of qualitative nature and depends heavily on the user interests. An example will make this clearer: In the case where one of the two linked names is that of a politician and the other that of the journalist writing the article, people who are interested in finding out about the politician may consider that the occurrence of the journalist’s name is unwanted noise. People looking for the journalist, however, may very well want to know who and what the journalist writes about. For this reason, we consider correct all social network links found in the news of the day, i.e. if two persons have been mentioned together in the news, they are linked. In that sense, all social networks summarised in Table 3 are correct so that we could claim near to 100% Precision. Incorrect links would then only be those where two unrelated articles are accidentally merged into one. In order to give at least some idea of the performance and usefulness of our system, we will first discuss some errors made by the system and highlight some observations made during the analysis (Section 5.2). We will then compare the social networks across languages (5.3).

5.1 The test set

Our social networks are generated live and change continuously. For evaluation purposes, we froze a time snapshot of 7 hours and 45 minutes, starting at midnight on Friday 13 July 2007. The following numbers give an idea of the relative impact of the various languages in this time snapshot.: EMM gathered 4358 news articles in 28 languages during this time, with the most prominent languages being En: 1688, Es: 800, Pt: 274, Ni: 221, Fr: 214, De: 189, Tr: 98, Hu: 96, Da: 95, It: 91. In these articles, 12,415 name mentions of 3,417 different persons and 16,954 links between persons were found. The following number of person name mentions were found for the various languages: En: 5971, Es: 1953, Fr: 849, Pt: 659, De: 491, Ni: 443, Ar: 247, Tr: 229, Da: 209, Ro: 197, It: 190, Hu: 175, Sv: 116. The networks and links for this test set have been frozen and can be found at <http://langtech.jrc.it/entities/socNet/test/last.html>.

2007-07-13T11:47+0200 . [Diplomati i kris](#) [SvenskaDagbladet]

...Vi har hela tiden hävdats att mordet på [Alexander Litvinenko](#) är ett allvarligt brott. Hundratals bri...
Det rysk-brittiska relationerna är mycket spända efter giftmordet på den avhoppade ryska agenten Litvinenko.

2007-07-13T11:41+0200 . [Бразилия подозревает Березовского в создании преступной группы и отмывании денег](#) [polit]

...попытка России отвлечь внимание от смерти [Александра Литвиненко](#), в отравлении которого Березовский обе...
Теперь Бориса Березовского хочет заполнить не только российское, но и бразильское правосудие. В четверг, 12 июля, федеральный суд Бразилии выдал ордер на арест беглого российского олигарха и нескольких руководителей футбольного клуба Corinthians из Сан-Паулу по подозрению в создании преступной группировки и отмывании денег.

2007-07-13T11:33+0200 . [Affäre Litwinenko: London droht Moskau mit Sanktionen](#) [FrankfurterRundschau]

...eheimdienstler und späteren Dissidenten [Alexander Litwinenko](#) im November 2006 hochradioaktives Polon...
Russland lehnt die Auslieferung eines Verdächtigen ab - und antwortet auf die Anfrage mit einer Propagandaoffensive.

2007-07-13T11:30+0200 . [Crispation entre Londres et Moscou](#) [rtbf]

...empoisonnement de l'ancien agent russe [Alexandre Litvinenko](#). Du coup, Londres pourrait expulser des...
L'affaire Litvinenko menace de plus en plus de faire dérailler les relations diplomatiques entre le Royaume-Uni et la Russie. Moscou a refusé l'extradition du principal suspect de l'empoisonnement de l'ancien agent russe Alexandre Litvinenko. Du coup, Londres pourrait expulser des diplomates russes.

Figure 2. Clicking on a person (here *Alexander Litvinenko*), shows the context (article, description and snippet). Clicking on the document title leads to the original document. At the top of the page, another hyperlink leads to the dedicated *NewsExplorer* page about this person.

5.2 Observations on the extraction of links

Links between persons are either not shown because their names were not recognised in the text (NER Recall) or because they were filtered out when reducing the network size (see Section 4.2). We found two examples where the system missed important persons: the British Queen Elizabeth II was not recognised because she was only referred to by the term “The Queen”, and the US president George W. Bush was missed because his known name variants were never mentioned. For disambiguation purposes, in NewsExplorer, we only recognise person names if at least two name parts (such as first and last name) are mentioned at least once in the article. In the case of the US president George Bush, his mention was not detected because he was only referred to by the strings “the Bush administration”, “the President”, and similar references.

In one case, *John F. Kennedy* was wrongly identified although the text referred to the airport with the same name. This disambiguation would have been difficult to solve as the word ‘airport’ was not mentioned: “... John F. Kennedy terror suspects ...”.

We also found one example of erroneous name merging: The news text referred to the Pakistani doctor *Mohammed Anif*. When following the hyperlink to the NewsExplorer page, we found that this person was merged with two other persons with the same name: a prisoner and a cricket player. In this case, the social network link was thus correct, but the additional NewsExplorer background information was partly wrong.

Another problem we came across was due to partially duplicate news articles as it happened that several identical text snippets led the system to create a strong link between two persons. The tenth English cluster and the second German cluster in Table 3 are such examples: The links are based on identical text snippets coming from two different newspapers, who even chose different titles for their stories.

Entirely or partially identical news articles are a frequent

phenomenon as a small number of news agencies provide information to many newspapers. Newspapers often either copy the whole article or large parts of it. Furthermore, they sometimes publish news updates with articles that change only slightly from one to the other. In NewsExplorer, these duplicate or near-duplicate articles are automatically eliminated as part of the clustering procedure (see Section 3.2). The social network analysis, however, operates on single articles so that every duplicate will be counted as one. One possibility to reduce the impact of duplicate articles at least to some extent would be to count only those name pairs that were found in different news sources.

5.3 Social networks across languages

The page <http://langtech.jrc.it/SocNet> shows the live social networks as identified in world-wide news in 32 different languages. As the number of articles per language and per country differs, the relations found in news articles of some languages and countries will clearly have a bigger impact than the relations found in other languages. For this reason, we list the number of articles per language separately for each of the 12 largest live social networks as part of the graph. This gives the user an idea how much each language contributed to each graph.

English is by far the most prominent EMM-NewsExplorer language, with approximately 7,000 English language news articles from around the world per day, followed by Spanish (3000), German (2500), Dutch (2000), Portuguese (1800), etc. EMM sources are continuously updated and changed so that the relative importance of the languages can change.

News articles in some of the languages clearly come from one country (e.g. Bulgarian, Farsi and Polish news are exclusively from Bulgaria, Iran and Poland). News in some other languages, however, represents various countries: English news may be dominated by British and US-American news sources, but comes from all around the world. German news comes from Germany, Austria and Switzerland, Dutch news from Holland and Belgium, etc.

2007-07-13T06:38+0200 . [UN nuke delegation arrives in Iran](#) [iranmania]

...to meet with Iran's top nuclear negotiator, **Ali Larjani**, later in the day, the report said. Accordiated Press (AP), Larjani and IAEA Chief **Mohammad ElBaradei** met last month in Vienna, Austria. Earli...

LONDON, July 12 (IranMania) - Iran's President Mahmoud Ahmadinejad said that the West should not expect his country to suspend uranium enrichment activities, the official Islamic Republic News Agency reported.

2007-07-13T06:37+0200 . [OIEA asegura haber logrado acuerdos Irán](#) [HoyDigital-DO]

...Internacional de Energía Atómica (OIEA), **Mohamad el Baradei**, hizo esta declaración al término de la (...) i, el asesor del principal negociador iraní **Ali Larjani**, que preside la parte iraní en las negociac...

TEHERAN, (EFE).- El jefe de la delegación del OIEA que visita Irán, Olli Heinonen, afirmó ayer que su equipo ha alcanzado un acuerdo con los dirigentes iraníes sobre "algunas cuestiones" en las negociaciones entre las dos partes sobre el caso nuclear iraní.

2007-07-13T02:13+0200 . [R E G I O N : Iran, UN team hold talks on nuclear issues](#) [dailytimesPK]

... deputy to Iran's chief nuclear negotiator, **Ali Larjani**. President Mahmoud Ahmadinejad said on Wedn (...) unclil. The UN watchdog's Director General **Mohammed ElBaradei** has said Iran's transparency offer combi...

TEHRAN: Iranian nuclear officials and a visiting team from the UN nuclear watchdog held a second round of talks on Thursday to discuss ways to remove outstanding questions about Iran's disputed nuclear programme. Iran has offered to draw up an "action plan" to address Western suspicions that its nuclear programme is a front to obtain nuclear arms.

2007-07-13T04:31+0200 [نتائج بناءة بين إيران والوكالة الدولية للطاقة](#) [alrai]

... في حين بناءة سيامة الاجراء...

طهران - وكالات - اعلنت ايران امس انه تم التوصل الى نتائج جيدة بعد ثلاث جولات من المحادثات مع وفد من الوكالة الدولية للطاقة الذرية. واختتمت الجولة الثالثة من المحادثات حول البرنامج النووي الإيراني بين المسؤولين الإيرانيين ووفد الوكالة الدولية للطاقة الذرية برئاسة أولي مابونزين نائب مدير الوكالة الدولية للطاقة الذرية.

Figure 3. Context of a link: Here *Ali Larjani* and *Mohammed ElBaradei* are highlighted in articles in which they appear together.

Table 3 summarises the main contents of the twelve biggest social network of the first eight hours of Friday 13 July 2007. The first two columns show the main mixed language networks with two different link thresholds (see Section 4.2). In brackets, we show which languages mainly contributed to each of the networks. The four remaining columns of Table 3 show the main monolingual networks for the four languages English, French, German and Arabic. Where appropriate and useful, we also show the most central names of each of the networks.

Table 3 shows that some social networks are clearly related across different languages. These are mainly linked to international politics and to various types of sports. Other networks are more country-specific and are not shared across languages. Whether or not these national networks make their way into the multilingual graph largely depends on the relative impact of each language.

The networks for international politics across languages mainly show the same persons, but different people take the more central roles. The position of the persons on the GraphViz network is automatically determined depending on the number of links to other persons. For instance, in the English network for international politics, the most prominent persons are G. Bush, H. Clinton, M. McConnell, M. Chertoff, Bin Laden and S. Hussein. The same persons can broadly be found in the related Arabic network, but the most central roles are taken by M. Jamil, Bin Laden, T. Blair, G. Bush, A. al-Zawahiri and M. Abbas. In the French related network, the most central persons are G. Bush, S. Hussein, N. Al-Maliki, C. Rice and H. Clinton. Interestingly, during the evaluation period, international politics scored only second in the French network, giving place to a French network of names, whereas for all other languages international politics took the first position.

For all languages but English, the number of articles is not large enough to fill all 12 social networks (indicated by the dashes ---), at least for the 8 hour test period. The fact that the first network is usually extremely tightly-knit whereas other networks are often scarce (consisting sometimes of only two or three names) indicates that the minimum link threshold should probably differ depending on the number of names per network.

6. Conclusion - Outlook

Altogether, the live social networks provide a lot of food for thought and they do show who is in the news right now across languages and internationally. They may also show how different countries present the same themes from different angles, depending on the people they mainly mention. It goes without saying that these live social networks are not a ready-made socio-political analysis of current events across languages and countries, but that they should be seen as a tool and one of the types of input that may help analysts do their work. However, they clearly also give observers a good first impression of current events world-wide and across countries. We

reckon that the user group that could reap the biggest benefit of our technology are analysts or researchers when investigating social networks produced for a selection of documents of their interest.

Work necessary to make the social networks more useful includes the adaptation of the link threshold according to the number of nodes on the graph. In addition to displaying the multilingual social networks, it would be useful to also give access to the monolingual social networks. We would like to work on improving the weight of links by boosting the link if it is fed from different sources or even from several languages and if the names are mentioned close to another in the text. We will eventually use different visualisation software and we plan to display thematic information for each social network, by either providing a list of keywords or by displaying the medoid article(s) for the documents that fed the network. Finally, we would like to experiment with graph theory algorithms to infer additional information from our graphs including cliques, walks and sub-graphs, as well as to highlight those paths that link different social networks.

7. Acknowledgement

We thank the entire EMM team and especially group leader Clive Best and chief developer Erik van der Goot for providing the valuable EMM news data and a very reliable and robust large-scale system. Martin Atkinson has started to develop a customised visualisation tool that will be used in the future.

8. References

- [1] Best, Clive, Erik van der Goot, Ken Blackler, Teofilo Garcia, David Horby (2005). *Europe Media Monitor – System Description*. Report No. EUR 22173 EN.
- [2] Connivence. 2007. See <http://www.connivences.info/> (last visited 28.03.2007).
- [3] GraphViz (2007). See <http://www.graphviz.org/> (last visited 13.07.2007)
- [4] LinkedIn. 2007. <http://www.linkedin.com/> (last visited 28.03.2007).
- [5] MySpace. 2007. <http://www.myspace.com/> (last visited 28.03.2007).
- [6] Paxson, Vern. 1995. *Flex – Fast Lexical Analyzer Generator*. Lawrence Berkeley Laboratory, Berkeley, CA. Available at <ftp://ftp.ee.lbl.gov/flex-2.5.4.tar.gz> (last visited 28.03.2007).
- [7] Pouliquen Bruno, Ralf Steinberger, Camelia Ignat, Irina Temnikova, Anna Widiger, Wajdi Zaghouni, Jan Žižka (2005). *Multilingual person name recognition and transliteration*. Journal CORELA. Numéros spéciaux, Le traitement lexicographique des noms propres.
- [8] Pouliquen Bruno, Ralf Steinberger, Camelia Ignat & Tamara Oellinger (2006). *Building and displaying name relations using automatic unsupervised analysis of newspaper articles*. Proceedings of JADT'2006. Besançon, France.
- [9] Pouliquen Bruno, Ralf Steinberger & Clive Best (2007). *Automatic detection of quotations in multilingual news*. Proceedings of RANLP'2007.
- [10] Silobreaker. 2007. <http://www.silobreaker.com/Corporate/> (last visited 28.03.2007).

- [11] Steinberger Ralf, Bruno Pouliquen, Camelia Ignat (2005). *Navigating multilingual news collections using automatically extracted information*. Journal of Computing and Information Technology - CIT:13.4, pp. 257-264.
- [12] Steinberger Ralf & Bruno Pouliquen (2007). *Cross-lingual Named Entity Recognition*. In: Satoshi Sekine & Elisabete Ranchhod (eds.). *Linguisticae Investigations* LI 30:1, pp. 135-162. Special Issue *Named Entities: Recognition, Classification and Use*.
- [13] Tanev Hristo (2007). *Unsupervised Learning of Social Networks from a Multiple-Source News Corpus*. Proceedings of RANLP'2007.
- [14] Wikipedia. 2007. <http://www.wikipedia.org/> (last visited 13.07.2007).
- [15] Zobel Justin & Philip Dart, 1995. *Finding approximate matches in large lexicons*. Software – Practice and Experience, Vol. 25:3, pp. 331-345

	All (3)	All (4)	En (2)	Fr (2)	De (2)	Ar (1)
1	International Politics (En-Fr-Es-De-Pt): Bush, Rice, Hussein, Musharraf, Chertoff, Petraeus, Boucher, ...	International Politics (En-Fr-Es-De-Pt)	International Politics (Bush, H. Clinton, McConnell, Chertoff, Bin Laden, Hussein)	French politics: Sarkozy + Strauss-Kahn + Jospin + Barroso, ...	US politics + Irak (Bush + various senators + al Maliki)	International Politics: M. Jamil, Bin Laden, Blair, Bush, al-Zawahiri, Abbas, ...
2	Cycling (En-Nl-De-Da-Fr)	Cycling (En-De-Fr-Nl-Da)	Cycling	International Politics: Bush + Hussein + Al-Maliki + Rice + H. Clinton	German calendar (birthdays etc. for this date)	Saad, Hariri, Amine + Pierre Gemayel, ...
3	German calendar (De)	German calendar (birthdays etc. for this date) (De)	UK politics + public life: Churchill, cook J. Oliver, actor A. Hepburn)	Canadian politics	Libyan HIV-scandal: Nicolas + Cécilia Sarkozy + Ferrero-Waldner + Gaddafi	Iran politics: Rafsandjani, Khamenei, Khatami, ...
4	JFK terror suspects (En): Abdul Qadir + ...	JFK terror suspects (En): Abdul Qadir + ...	Football + public life; D. + V. Beckham, Paris Hilton	Cycling	German politics: Stoiber + Huber + Seehofer + Beckstein	Sadr + Abdul Aziz-al-Hakim
5	Argentinean politics (Es-Pt)	Football (En-Fr-Tr-Sk-Pt-Es)	New Zealand rugby + public life	Hariri assassination: Brammertz + Eido	German-Turkish politics: Merkel + Köhler + Necdet Sezer + Kolat	Chomsky + Fisk + Miller
6	Spanish politics (Es-Ca-Ro)	Dutch historical (Hitler, Stalin, Gandhi, ...) (Nl-En)	Nigerian politics	European Union + European Parliament: Solana + Pötering	Cycling	Iran nuclear conflict: El Baradei + Larijani + Mohammed Said
7	Dutch historical (Hitler, Stalin, Gandhi, ...) (Nl-En)	Argentinean politics (Es-Pt)	Cricket	French politics: Carrez + de Courzon	UK politics: Blair + Brown	Annan + Miró + ...
8	Japanese politics (En)	Baseball (En)	Japanese politics	---	---	Public Life: David Beckham, J. Lennon + ...
9	Iran nuclear (En-Pt-Ar-Es)	Public Life, Stars (En-Da-Pt-Sv-Pl-De)	JFK terror suspects	---	---	Pakistan politics: Musharraf + Youssef Mohamad
10	Baseball (En)	Spanish politics (Es-Ca-Ro)	Tennis	---	---	---
11	Cricket (En)	Turkish politics (Tr-Sv)	Iran nuclear conflict	---	---	---
12	Canadian politics (Fr-En)	Slovene politics (Sl)	Kosovo news + UN resolution + Russia: Lavrov + Miliband + Litvinenko + ...	---	---	---

Table 3. Main subjects or actors in the 12 top social networks generated for the first eight hours of Friday 13 July 2007. The first two columns show mixed-language networks, with the languages in brackets indicating the dominant languages contributing to each of the social networks. Columns three to six show monolingual English, French, German and Arabic networks. The number in brackets in the header row indicates the threshold (minimum number of links required for links to be displayed). Dashes (---) indicate that there were no more social networks with a link strength above the threshold.

en [1507] pt [644] de [482] es [441] fr [297] sl [154] da [129] ro [104] ru [103] nl [79] sv [50] hu [39] it [38] no [31] bg [31] tr [29] ar [24] lt [19] cs [15] fi [13] et [12] pl [8] sk [5]

