Challenges and methods for multilingual text mining

Ralf Steinberger
European Commission – Joint Research Centre (JRC)
Via Fermi 2749, 21027 Ispra (VA), Italy
E-mail: Ralf.Steinberger@jrc.ec.europa.eu

Abstract

Multilingual text processing is useful because the information content found in different languages is complementary, both regarding facts and opinions. While Information Extraction and other text mining software can, in principle, be developed for many languages, most text analysis tools have only been applied to small sets of languages because the development effort per language is large. Self-training tools obviously alleviate the problem, but even the effort of providing training data and of manually tuning the results is usually considerable. In this paper, we gather insights by various multilingual system developers on how to minimise the effort of developing natural language processing applications for many languages. We also explain the main guidelines underlying our own effort to develop complex text mining software for tens of languages. While these guidelines – most of all: extreme simplicity – can be very restrictive and limiting, we believe to have shown the feasibility of the approach through the development of the *Europe Media Monitor* (EMM) family of applications (http://press.jrc.it/overview.html). EMM is a set of complex media monitoring tools that process and analyse up to 100,000 online news articles per day in between twenty and fifty languages. We will also touch upon the kind of language resources that would make it easier for all to develop highly multilingual text mining applications. We will argue that – to achieve this – the most needed resources would be *simple, parallel and uniform* multilingual dictionaries, corpora and software tools.

1. Introduction

The share of non-English documents on the internet is rising continuously. While many private users will only be interested in finding monolingual information in their own language, the need for multilingual information retrieval, information extraction and cross-lingual information access for professionals, organisations and businesses is rising steadily. Starting from the premise that we need multilingual text mining tools, the question we would like to ask here is: *How can we avoid that the development of (any) text mining tool for N languages takes N times the effort of developing them for one language*. It is generally acknowledged that developers benefit from the experience of having produced tools in one or more languages before, and that the existence of an efficient implementation infrastructure is extremely important (e.g. Maynard et al. 2002). Such software building blocks can include, for instance, a grammar implementation formalism, tools for marking up text, debugging tools, automatic evaluation tools and procedures, etc. Furthermore, simple applications like sentence splitters are typically so similar for different languages that – once one exists – the same tool is usually quickly applied to new languages. We will thus try to take the effort of developing the infrastructure out of the equation. The question should thus be reformulated: Assuming that you have already developed text mining tools for some languages, how can you limit the effort to develop such tools for several other languages.

In the next section, we will try to demonstrate the need for multilingual text processing and to show that most application providers offer monolingual tools or tools covering a few commonly spoken languages. In Section (3), we will describe the type of data we work with (mostly news) and give a short overview of the functionality of the *Europe Media Monitor* family of applications. In Section (4), we will then try to answer the main question asked here. First, we will summarise insights by other multilingual system developers (4.1) and discuss the contribution of Machine Learning methods (4.2) – in our view an extremely promising approach to go highly multilingual. We will then present our own guidelines on how to minimise the effort of multilingual tool development (4.3), which – of course – largely overlap with those proposed by others. In Section 5, we will give some examples of what these insights and guidelines concretely mean for the development of a small selection of natural language processing tools. One obvious bottleneck for the development of multilingual tools is the lack of linguistic resources. In Section 6, we thus share our view on which kind of resources would be particularly beneficial to achieve highly multilingual text mining applications. Section 7 summarises and concludes.

2. Motivation for multilingual text mining

The *Joint Research Centre* (JRC) is the scientific-technical arm of the European Commission (EC). The European Union (EU) institution EC is a multinational organisation with strong links also to countries outside the EU. It is thus natural that multilinguality plays a big role inside the organisation. However, experience with the many partners and customers of the JRC shows clearly that even many national organisations have a need for highly multilingual text processing applications.

The JRC receives frequent requests to monitor media reports in dozens of languages, involving news gathering, classification, information extraction and analysis. The JRC's users consist of EU institutions, state organisations inside its 27 Member States, institutions of partners outside the EU (e.g. in the USA, Canada, China, etc.), as well as international organisations (including various United Nations and pan-African sub-organisations). These users have a wide range of interests so that not only media reports in the 22 official EU languages need to be monitored, but also, for instance, those in the languages of

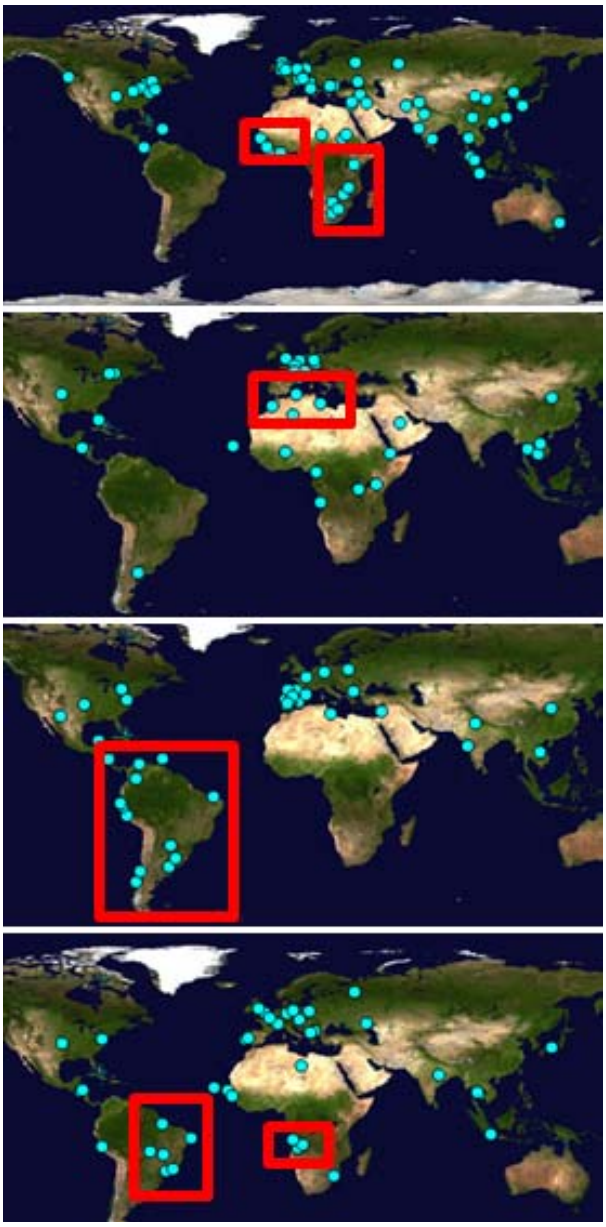

Figure 1. The four maps show the complementary locations mentioned in health-related news published in the same time window in the four world languages English, French, Spanish and Portuguese (from top left to bottom right).

the EU's neighbouring countries, of the world's crisis areas and of political partner countries around the world.

To give a concrete example: Public Health organisations around the world monitor any threats to the populations of their counties – be they chemical, biological, radiological or nuclear (CBRN). For that purpose, they not only gather information on communicable diseases, etc. from their hospitals (*indicator-based* risk monitoring), but they also scan online news articles and government websites to find out about the outbreak of communicable diseases, etc. (*event-based* risk monitoring; Linge et al. 2009). In the era of high mobility and mass long-distance travel, the risk of contracting a disease (e.g. the human influenza virus, also referred to as 'swine flu' or H1N1), taking it home and passing it on to others is so big that the Public Health community follows the situation around major tourist destinations and locations for international religious and sports-related mass gatherings thoroughly, by monitoring international media reports published around the world.

It is our experience that *multilingual* media monitoring is not only a luxury, but – due to the *information complementarity* in the news across different languages – an urgent requirement. Large events and events that are in the focus of the world media (e.g. reports from conflict areas such as Iraq or Israel, or reports about human bird flu cases) will usually be translated into English and other world languages. However, many smaller events rarely make it into the international news, including local reports on the outbreak of more common diseases (e.g. tuberculosis or malaria), or reports about pastoral conflicts in Africa, although this type of report may be important to organisations monitoring Public Health or country stability. **Figure 1** gives a good indication of cross-lingual information complementarity occurring in domain-filtered real-life news.

Information complementarity not only applies to contents, but also to opinions: by considering points of view from around the world, readers will get a *less biased*, and more balanced, view on world events. To give only one simple example: Daily and long-term social network analysis across various countries and languages (Pouliquen et al. 2007b) has shown that the most central personalities are usually the respective leaders of state. When only reading English language news, readers will thus get an inflated impression of the importance of the US President and the British Prime Minister, while the readers of Russian, Arabic or Spanish language news will get quite a different impression.

The most common approach to capturing information published in foreign languages is the use of Machine Translation into one target language (e.g. English) and to apply information filtering and extraction tools in that target language. A limitation of this approach is that proper names and specialist terms are frequently badly translated so that information can easily get lost. Our own insight (supported by the *native language hypothesis* observed by Larkey et al. 2004) is that information filtering in the source language is more efficient than filtering machine-translated text. In the USA, Machine Translation is nevertheless an attractive solution, as there is only one official national language. However, when looking at Europe, Asia and other parts of the world, it becomes clear that the situation in the US is an exception rather than the rule, as there is no agreement on one common language.

News aggregators such as Google News[1], Yahoo News[2] and EMM[3] already gather and cluster news in many languages (currently 45, 35 and 50 languages, respectively – status March 2010), but most of the more complex systems carrying out some level of analysis of the gathered texts are monolingual, including *SiloBreaker*[4], *NewsVine*[5] and *DayLife*[6]. The news analysis

[1] See http://news.google.com . All websites mentioned here were last visited on 19 March 2010 or later.
[2] See http://news.yahoo.com/ .
[3] See http://emm.newsbrief.eu/ .
[4] See http://www.silobreaker.com/ .

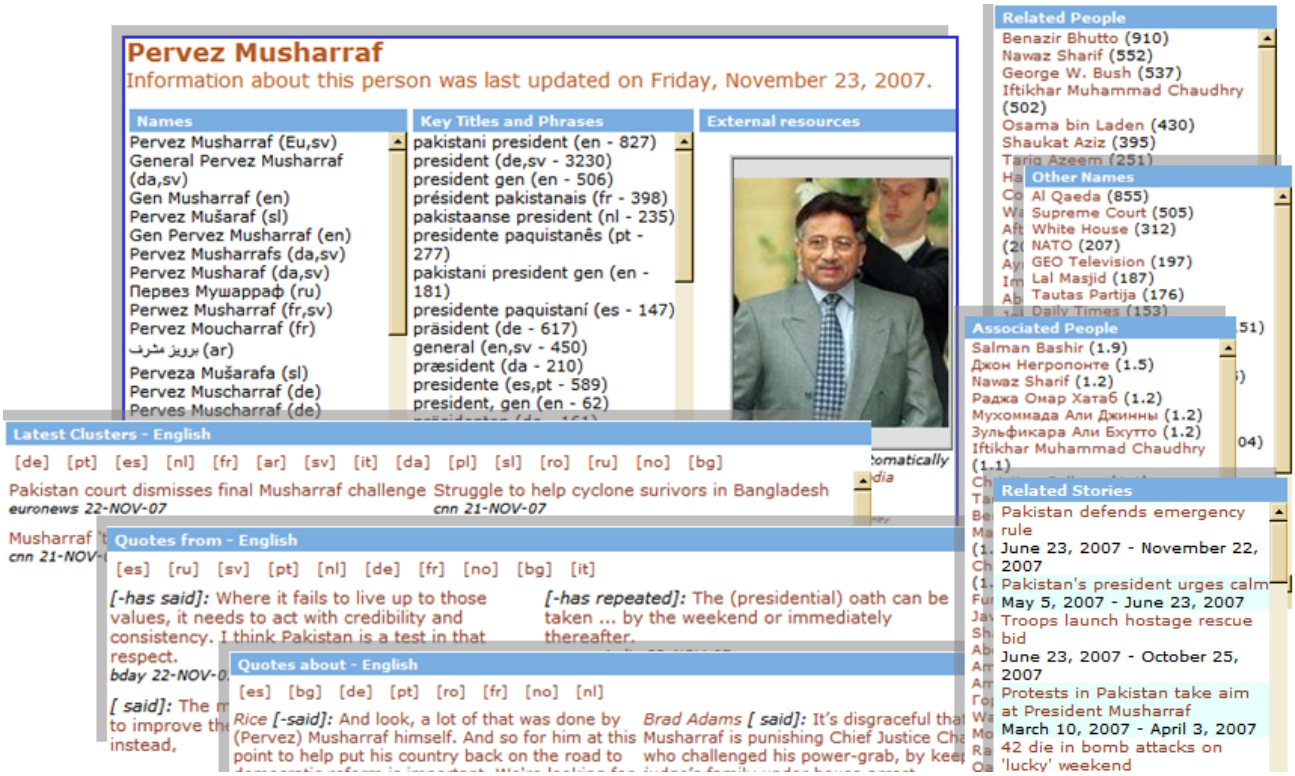

Figure 2. Named entity-related information extracted and aggregated by the EMM application *NewsExplorer* from news in 20 languages, including: name variants, titles, latest clusters and 'stories' where mentioned, quotes by and about that person, ranked lists of persons and other entities mentioned historically in the same clusters.

systems *NewsTin*[7] and the EMM product *NewsExplorer*[8] are notable exceptions, covering 11 and 19 languages, respectively.

We believe that the main reason for the existence of monolingual analysis systems is the large effort required to produce text processing software for new languages. In the worst case, the effort required to develop tools in N languages is N times the effort of developing monolingual software, but various multilingual system developers have found methods to minimise this effort. These insights will be the main focus of the rest of the paper.

3. The Europe Media Monitor family of applications

The *Europe Media Monitor* (EMM, Steinberger et al. 2009) is the basic engine that gathers an average of about 100,000 news articles per day in approximately 50 languages (status March 2010), from about 2,500 hand-selected web news sources, from a couple of hundred specialist and government websites, as well as from about twenty commercial news providers. EMM visits the news web sites up to every five minutes to search for the latest articles. When news sites offer RSS feeds, EMM makes use of these, otherwise it extracts the news text from the often complex HTML pages. All news items are converted to Unicode. They are processed in a pipeline structure, where each module adds additional information. Whenever files are written, the system uses UTF-8-encoded RSS format.

The EMM news gathering engine feeds its articles into the four fully-automatic public news analysis systems (accessible via http://emm.jrc.it/overview.html), and to their non-public sister applications (Steinberger et al. 2009). The major concern of *NewsBrief* and *MedISys* is breaking news and short-term trend detection (topic tracking), early alerting and up-to-date category-specific news display. *NewsExplorer* focuses on daily overviews, long-term trends (topic tracking), linking of related news across languages, in-depth analysis and extraction of information about people and organizations (see **Figure 2**). *EMM-Labs* is a collection of more recent developments and includes various tools to visualize the extracted news data. For NewsBrief and MedISys, there are different access levels, distinguishing the entirely public web sites from an EC-internal website. The public websites do not contain commercial sources and may have slightly reduced functionality.

The following text mining methods and tools are applied and closely integrated in EMM; if not mentioned otherwise, they work for 20 languages: document clustering and Boolean classification (up to 50 languages); breaking news detection and automatic user notification (50 languages); Named Entity Recognition (persons,

[5] See http://www.newsvine.com/ .
[6] See http://www.daylife.com/ .
[7] See http://www.newstin.com/ .
[8] See http://emm.newsexplorer.eu/. NewsExplorer processes news articles in Arabic, Bulgarian, Danish, Dutch, English, Estonian, Farsi, French, German, Italian, Norwegian, Polish, Portuguese, Romanian, Russian, Slovene, Spanish, Swedish and Turkish. Text mining tools for the 20th language, the African Bantu language Swahili, have been developed and tested and are currently being integrated.

organisations); name variant matching (i.e. string distance calculation, including across scripts); geo-tagging (recognition and grounding for map-display); quotation recognition (reported speech by and about named entities; 14 languages); multi-label classification using the thousands of categories from the Eurovoc[9] thesaurus; multi-monolingual topic tracking (to detect 'stories') and aggregation of information per 'story'; cross-lingual news cluster linking (available for the majority of the 171 possible language pairs); social network generation based on information extracted from multilingual news (based on co-occurrence, and also on who mentions whom in reported speech); detailed scenario template filling for events causing victims (violence, natural disasters, accidents, etc.; six languages); visualisation (using geographical maps, trends, social networks, etc.).

EMM was mostly developed to serve the interests of the European Institutions and their international partners, but the public web pages are also visited by an average of 30,000 anonymous users per day.

4. How to achieve multilinguality

Many individual natural language processing applications have been developed for two or more languages. We have not found many publications directly addressing the issue on how to minimise the effort of multilingual tool development, but several that describe the efforts of adapting a certain tool to a new language. Typically, these applications are named entity recognition systems or syntactic parsers. Section 4.1 contains a list of ideas found in such publications. Section 4.2 addresses the role of Machine Learning approaches, which seem to be particularly useful to achieve multilinguality. Section 4.3 summarises our own approach which, obviously, in many cases, overlaps with that of other developers.

4.1 Related work: Insights by other multilingual developers

Multiple authors have described work on developing resources and tools for a number of different languages. This was typically done by reusing the resources from a first language and adapting them to new languages (e.g. Gamon et al 1997; Rayner & Bouillon 1996; Pastra et al. 2002; Carenini et al. 2007; Maynard et al. 2003). Practical tips from various system developers for achieving multilinguality include the use of **Unicode** and of the usage of **virtual keyboards** to enter foreign language script (Maynard et al. 2002); **modularity** (Pastra et al. 2002; Maynard et al. 2002); **simplicity** of rules and the lexicon (Carenini et al. 2007; Vergne 2002); **uniform input and output structures** (Carenini et al. 2007; Bering et al. 2003); and the use of **shared token classes** that are ideally based on surface-oriented features such as case, hyphenation, and includes-number (Bering et al. 2003). SProUT grammar developers took the interesting approach of using shared resources between languages (lexica, gazetteers, grammar rules) for named entity recognition in seven languages, and of splitting the multilingual grammar rule files (Bering et al. 2003): some files contain rules that are applicable to several languages (e.g. to recognise dates of the format *20.10.2010*) while others contain language-specific rules (e.g. to cover *20th of October 2010*). The fact that this latter date format, and others, can also be captured by using language-independent patterns was shown by Ignat et al. (2003).

Both Maynard et al. (2002) and Pastra et al. (2002) point out that the usage of **theory-neutral data types** is an advantage for the Language Engineering architecture GATE because it facilitates reuse. This does make sense for a platform that is meant to be used by many groups for many purposes. However, there are several grammar developers who point out that adhering to **grammar theories** is very efficient because they separate universal rules from language-specific parameters and differences. For instance, Bender & Flickinger (2005) highlight the benefits of adhering to Head-Driven Phrase Structure Grammar (HPSG) for writing multilingual general-purpose grammars. They even propose to *generate* starter grammars automatically, based on a number of linguistic features of a language. Gamon et al. (1997) report that the framework of Universal Grammar allows them to create a generic grammar that "can easily be parameterized to handle many languages". Interestingly, they provide detailed information on the percentage of grammar rule overlap between their original English general-purpose Microsoft-NLP grammar and the German, French and Spanish grammars they derived from the English version. Wehrli (2007), using Chomsky's generative grammar to build parsers for six languages, stipulates that the design he adopts "makes it possible to 'plug' an additional language without any change or any recompilation of the system. It is sufficient to add the language-specific modules and lexical databases". Ranta (2009, e.g. pp. 47ff), having worked within the Grammatical Framework on fourteen languages, also addresses the degree of grammar sharing across languages, as well as within language families. He highlights that the mere existence of an abstract syntax implies grammar sharing and he shows that some linguistic phenomena can be treated in a systematic way.

Vergne (2002) does not adhere to a grammar theory, but tries to reach language-independence by using an extremely **simple, minimalistic and radical approach** to building multilingual chunkers and (partial) parsers, without using full dictionaries. He shows the feasibility of his approach by building a tool that extracts subject-verb combinations for five languages, using dictionaries of only about 200 elements per language, case information and regular expressions matching certain combinations of word endings. More recently, Vergne (2009) proposed a chunker using only string length and word frequency, and applied it to 23 languages. The basic idea, which we share, is thus to limit the used resources to a bare minimum, i.e. to those elements that are required for a specific task.

It goes without saying that simple applications can more easily be achieved with simple means and that more complex applications are likely to benefit from a deeper linguistic analysis. There is thus not one solution for all tools and applications. However, we observed – for the information extraction tasks we are targeting – that even simple means can take you relatively far, and that minimalism and simplicity paid off for us.

[9] See http://europa.eu/eurovoc/. Automatic Eurovoc indexing has been trained for 22 EU languages.

4.2 Related work: Machine Learning

Machine Learning (ML) approaches have become very popular. The technology has advanced a lot over the last years, and the trend is likely to continue. The obvious appeal of self-learning software is that it will by itself take care of learning rules and vocabulary, and that it can be optimised for real-life data by training it on such data. ML is thus a very promising solution to achieve high multilinguality.

In the field of Machine Translation (MT), statistical (i.e. self-learning) methods are currently the major trend, i.e. systems that learn automatically from texts that have previously been translated manually. *Google translate*[10] now offers all language pair combinations for the impressive number of 52 languages, i.e. 1326 language pairs (status: March 2010). Never before has any translation software been available for so many languages. A current trend is to combine purely statistical MT with symbolic MT, e.g. by integrating the processing of syntactic rules (e.g. Goutte et al. 2009). When doing this, the question arises again how this can be done with minimal effort for many languages, but presumably the rules will be rather language or language pair-dependent.

In the field of NER, ML techniques have been widely used (Nadeau & Sekine 2009). The most common approach is to use *supervised ML*, i.e. training a system on previously annotated corpora. While the idea is attractive, the de-facto limitation is the fact that producing such annotated corpora (e.g. for new languages) is labour-intensive and expensive. Alternatives are to use semi-supervised or unsupervised learning methods. *Semi-supervised* learning involves a set of seeds to start the learning process and boot-strapping methods to gradually increase the number of patterns and resources. *Unsupervised* learning makes use of external lexical resources and large corpora. Typically, known entities from the lexical resource are first looked up in the corpus in order to derive frequent lexical patterns around these entities. These patterns are then used to detect new entities. An open issue is how to combine ML methods with manual intervention, e.g. if one wants to manually correct recognition mistakes.

ML methods, especially semi-supervised and unsupervised, are clearly very promising when attempting to achieve high multilinguality. In the context of EMM, however, we decided for ourselves to use hand-crafted rules, and to enhance manually produced dictionaries and word lists by using bootstrapping and Machine Learning methods. Doing this allows us to keep control over the recognition performance. Information redundancy is high in EMM, so that we aim at high precision and accept lower recall, assuming that, if we miss some information in one article, we are likely to find it in another.

We believe that our approach requires less time per language than when using pure Machine Learning methods. We typically invest a maximum of three person months to add a new language to the tool set, as this is the average time of having a native speaker trainee available for us. In this time period, the person can discover and add news sources, translate the Boolean category definitions, provide the linguistic IE resources for the new language, and test the performance. However, it is also possible to produce reasonable initial linguistic resources to recognise named entities and quotations in a new language within one working week.

[10] http://translate.google.com/

4.3 Insights by EMM developers

Due to the strict requirement of having to analyse documents in many languages (ideally, all 22 official EU languages, plus more) while working in a small team (three computational linguists during most of the years, but currently seven), we always had to use minimalistic methods and try to achieve with them as much as possible. Basically, we were reduced to *not* using parsers, part-of-speech taggers, morphological analysers and full dictionaries for any of the languages, and we had to keep the effort of adding a new language to the tool set to a maximum of three months, including testing. While good linguistic resources are available freely for some languages, we could not make use of them as we needed to keep the work parallel for all languages. The kind of resources we *do use* are targeted word lists (name titles; gazetteers of place names; sentiment words; reporting verbs and – very important – different types of stop words, etc.); mixed-language Boolean combinations of category-defining words; the output of our own NER tools; statistics, heuristics, boot-strapping methods and machine learning.

Regarding methods to keep the development effort per language down, we basically had the same insights other groups identified (i.e. those mentioned in the 1st paragraph in section 4.1). The most important ones for us are *modularity* and *simplicity*. Another principle we often applied, closely linked to simplicity, is **under-specification**. The idea is: don't formulate constraints if you don't urgently need them, as they are time-consuming to produce and they may hinder you in your analysis of other languages. For instance, if it is not strictly necessary in local patterns to specify the morphological agreement and the order of words or word groups (e.g. modifiers for titles in person name recognition), simply leave them unspecified (see also Section 5.1).

Another difference to the work presented in 4.1 is that we developed further the *modularity* idea, by using mostly **language-independent rules** that make reference to language-specific resource files containing application-focused word lists. For applications such as person and organisation name recognition, quotation recognition, and for geo-tagging and grounding (distinguishing, e.g., which of the 15 locations world-wide with the name of 'Paris' is being referred to in the text), this principle was adhered to quite closely. In exceptional cases, such as person name recognition in Arabic (which does not distinguish upper and lower case), separate recognition patterns were added and located in the file containing the language-specific information (Zaghouani et al. 2010). That way, the resulting system is entirely modular. When adding a new language, it is normally sufficient to plug in the language-specific parameter file. For person name recognition, this file

includes long lists of words, phrases and regular expressions that are typically found next to person names and that help determine whether some uppercase words are a name or not. The resulting patterns can also identify and store names and titles in more complex expressions such as: *the recently elected chairperson of LREC, Nicoletta Calzolari*, or *Tony Blair, 56-year old former British Prime Minister*. The required word lists are usually produced using seed patterns, machine learning and knowledge discovery, and boot-strapping, but external knowledge sources such as Wikipedia are of course also used, when available.

Highly inflected languages are a challenge for simple methods that rely a lot on matching expressions in a text against word lists. To solve the problem, we either apply some simple suffix stripping and suffix replacement rules (e.g. to recognise *New Yorgile* as an Estonian inflection of the name *New York*), or we pre-generate many variants of known names so as to facilitate their recognition in text, using finite state tools. Our data base contains over 1 million known entities (plus additional hundreds of thousands of known name variants), collected through multi-annual multilingual information extraction. For example, for the name part (Tony) *Blair* and the Slovenian language, inflections such as the following are thus automatically generated: *Blairom, Blairju, Blairjem*, etc.

For the more complex task of event scenario template filling in six languages, we did not entirely adhere to the principle of language-independent grammars (Tanev et al. 2009). However, the approach still is minimalistic in the sense that no part-of-speech taggers or syntactic parsers are used and that we do not use complete dictionaries. Instead, the system uses local grammars to identify the information for the individual slots, such as event type; number, status and type of victims; perpetrator, location and time. This information is then combined to produce the entire event description.[11]

The approach for the development of multilingual text mining applications in EMM is described in more detail in Steinberger et al. (2008), where we also give an overview of how these generic principles work in practice, for seven different text mining applications. In Steinberger et al. (submitted), we describe the concrete effort of adding a new language to the tool set (the African Bantu language Swahili).

EMM-NewsExplorer also offers some *cross-lingual* functionality for its nineteen languages, i.e. cross-lingual cluster linking, name variant matching (including across scripts), and merging the information extracted about entities in all monitored languages. As there are 171 language pairs for 19 languages, the use of bilingual resources and methods needed to be strictly avoided. Another guideline we follow is thus: for cross-lingual applications, **avoid the usage of bilingual resources** and favour (more or less) language pair-independent methods.

It should be clear by now that EMM tools do not adhere to a grammar theory or any other theoretical framework.

[11] The event extraction results are accessible at http://emm.newsbrief.eu/geo?type=event&format=html&language=all .

5. Some examples

The means imposed by the multilinguality requirement, presented in Section 4.3, are very restrictive. While they make extending to many languages easier, they also represent a challenge for most text mining applications. In the previous section, it already became clear how we solved the challenge for person name recognition and event scenario filling. We will now try to sketch solutions for another application we have developed already (quotation recognition; Section 5.1), and for others we are currently working on (Sentiment Analysis, 5.2; and Multi-document Summarisation, 5.3).

5.1 Quotation Recognition

The quotation recognition tool, currently covering 14 languages, aims to detect occurrences of direct reported speech if the speaker can be unambiguously identified (for display on the person pages in NewsExplorer[12]). If the quotation makes reference to another known entity, this will be recorded, as well (quotation *about* an entity). Details on this tool can be found in Pouliquen et al. (2007a). The patterns consist of quotation markers (e.g. “, ‘, «), previously identified person or organisation names, reporting verbs (e.g. *said, reported, argues*, etc.) and a range of modifiers that can be found between any of the other elements (e.g. *yesterday*, *on TV*, etc.). The sample rule below would successfully identify the quotation, the speaker (Angela Merkel) and the entity referred to in the quotation (Barack Obama) in the following string: *Merkel said yesterday on TV “…Obama …”*.

NAME REPORTING_VERB MODIFIER “QUOTE”

Note that the co-reference between *the US President* or *President Obama* and the known entity *Barack Obama* will be established if the full name is mentioned at least once in the document and if either at least one name part and/or one of the many previously identified titles for that name are found.

To comply with the *simplicity* and *under-specification* requirement, the order of modifiers and any morphological agreement (e.g. in number or gender) will not be specified. It is furthermore possible to allow any combination of individual modifier words (e.g. *TV yesterday on*) without much risk as we focus on recognition (and not generation) and the ungrammatical combinations will simply not be found in real-life text.

5.2 Sentiment analysis

EMM users are not only interested in factual content, but also in opinions on certain entities and issues (such as the EU constitution). Questions asked concern the (positive or negative) attitude of media sources in certain countries towards these *targets*, and of changes across languages and over time. Approaches to opinion mining vary widely

[12] See, for example, Barack Obama’s page at http://emm.newsexplorer.eu/NewsExplorer/entities/en/1510.html

regarding the methods and the depth of analysis (see, e.g. Pang & Lee 2008). Due to our multilinguality requirement, we again need to use the simplest possible methods and we have to avoid using parsers, part-of-speech taggers and full dictionaries. Instead, we restrict ourselves to the usage of word lists (positive and negative words, polarity inverters and enhancers) and previously recognised named entities). To avoid negative news content (e.g. in news on natural disasters) having an impact on the detected sentiment towards any entity mentioned in these news items, we decided not to consider sentiment words that are also part of EMM's category-defining words, such as *disaster, tsunami* and *flood* for the EMM category 'Natural Disasters'. These word lists are not ideal for the task, but they are readily available for all languages. To ensure furthermore that the sentiment words actually apply to the entity we are interested in, we use word windows around the entities and their titles. Experiments with various English language sentiment vocabularies showed that the best-performing results were achieved with a window size of six words to either side of the entity and its titles. See Balahur et al. (2010) for details.

Many English language sentiment dictionaries are freely available, but such vocabulary lists are scarce for other languages. Having identified a reasonably performing language-independent method for sentiment analysis, our next challenge is to semi-automatically generate large non-English sentiment vocabularies.

5.3 Multilingual multi-document summarisation

Due to the high redundancy of EMM's news content (100,000 news articles per day collected from about 2,500 different media sources), a major task performed by the EMM systems is to group related articles into clusters, and to track the development of these news clusters over time. Currently, EMM displays the title and description of each cluster's centroid article, but a proper summary per cluster, and update summaries for clusters related over time would be very useful. Hence our motivation to work on multilingual multi-document summarisation.

As abstractive summarisation would require many linguistic resources, our multilingual environment restricts us to extractive methods, not considering syntax. The proposed solution consists of using latent semantic analysis (LSA) to select the most informative sentences from the whole cluster (similar to Gong & Liu 2002). In addition to a list of words and word-ngrams per sentence, the LSA input in our system consists of previously identified entity mentions, and of (non-disambiguated) mentions of terms from the multilingual MeSH thesaurus (*Medical Subject Headings*[13]). The idea behind this approach is to (a) give higher weight to entities and (b) to capture some synonymy and hyponymy relations, both to select the most important sentences and to avoid information redundancy in the selected sentences. Due to our historical collection of multilingual name variants and a list of previously found titles for each entity, our lookup recognises name mentions even if the spelling varies. The approach was successful at the TAC'2009 competition, achieving second place for the most important category 'overall responsiveness', out of 54 submissions (see Josef Steinberger et al. 2009).

6. Required language resources

In the previous sections, we tried to summarise the constraints we imposed on ourselves when developing multilingual text mining applications. We also tried to sketch simple solutions that allowed us to avoid using too many linguistic resources. If linguistic resources had been freely available for all the languages we are trying to cover, development time would have been reduced drastically and it is likely that the results achieved would be better. In this section we thus want to give an idea of tools and resources that – we believe – would enable the community to build multilingual text mining applications better and more quickly.

The major – probably banal – statement we would like to make is that the community would strongly benefit from *freely available, simple, parallel and uniform multilingual dictionaries, corpora and software tools*.

The resources should ideally be *free* because universities and research organisations in many countries would otherwise not get access to these resources. This is particularly true for lesser-used languages, meaning: the majority of languages. The current situation leads to a scientific brain drain because students and researchers around the world have to work on (mostly) English language applications because this is one of the very few languages for which tools are readily available. If working on their own languages, they would be reduced to developing basic tools and resources such as corpora, dictionaries and morphological analysers.

The tools and resources should be *simple* because they would otherwise never be built for many languages. We believe this to be true because of the associated cost, the time required for the development, and the limitations on available qualified manpower. At a recent FLaReNet event[14], Grefenstette (2010) presented the idea of a community-based Web 2.0 effort to build simple dictionaries for many languages. The basic idea is to ask native speakers to provide lemma, main part(s)-of-speech and English translation(s) for a list of (possibly frequency-sorted) word surface forms. The usual Web 2.0 incentives and control mechanisms could be applied. Even non-linguists can provide this type of information. Usability would be limited for more complex applications requiring, for instance, sub-categorisation frames, but applications like those developed as part of EMM would certainly benefit. The pragmatic proposal of also providing the English translation is, of course, the most arguable feature.

[13] See http://www.nlm.nih.gov/mesh/. The multilingual MeSH term recognition software was developed by Health-on-the-Net (HON, http://www.hon.ch/).

[14] See http://www.flarenet.eu/?q=node/347.

The tools and resources should be *parallel and uniform*, i.e. input and output format should be the same for all languages, the same set of parts-of-speech and syntactic categories should be used for all, etc. Ideally, resources should also be linked across languages. Uniform and parallel dictionaries would allow, for instance, to write multilingual rules and patterns much more easily. Successful efforts that produced such lexical resources in the past were Multext[15], Multext-East[16], EuroWordNet[17] and the GeoNet Names Server[18]. Unfortunately, the parallelism for the WordNet family ended with version 1.6, when the project ended. The *Eurovoc thesaurus*[19], a multilingual categorisation scheme used by parliaments in Europe, was not developed for machine use, but it is still very useful because it covers almost thirty languages and it has been used to manually classify large numbers of documents. Using such uniform lexical resources, multilingual grammars are likely to be much more comparable and the effort of adapting one to another language would be minimised. Parallel corpora are much more useful than multi-monolingual corpora. Apart from their usefulness to train statistical machine translation and to construct multilingual dictionaries, they can be exploited to train and evaluate systems for information extraction, alignment, document categorisation, etc. with minimal effort. For instance, one can assume that the mentioned entities are the same across languages so that marking up a documents in one language would allow using the resource for many. In spite of its limited subject domain, the parallel corpus (22 languages) *JRC-Acquis* (Steinberger et al. 2006) has therefore been useful for various multilingual tasks. In the CoNLL shared tasks 2006 and 2007 (Nivre et al. 2007), dependency parsers were trained and tested for 13 and 10 languages, respectively. This was a very useful effort for promoting multilinguality, and more. However, as the training corpora used different grammatical features and labels (e.g. for part-of-speech and syntactic phrases), the output for the same parsing system is not homogeneous across languages. Any rules reading the dependency tree output would thus need to be written differently for each language. This limits the usability of the otherwise very useful multilingual tool enormously. Software tools trained or built with uniform and parallel resources are likely to be parallel, or at least very similar, themselves. They would minimise any effort of building upon their output considerably. We were thus happy to find out about LDC's *Less Commonly Taught Languages* project[20], which produces parallel resources for many languages.

Finally, it is important to have a *single access point for licensing* issues (such as ELDA[21] and LDC[22]) to avoid having to contact many different content providers when building a highly multilingual system, although the usage entirely without licences would, of course, allow even more flexibility.

It goes without saying that building resources and tools with these specifications is expensive and time-consuming. The number of highly multilingual parallel texts is limited and copyright issues may make it difficult to use them. The existence of the resources and tools described here may remain a dream. However, we feel that such resources would be a big step towards developing highly multilingual text mining applications, and awareness may be the first step towards achieving this goal.

There has been a lot of progress recently in the field of multilinguality and multilingual resources, which gives us hope that – also from a linguistic point of view – this world will soon be much smaller. Past and present initiatives such as ELRA[23], FLaReNet[24], CLARIN[25], CLEF[26], ENABLER[27] and TELRI[28] are very promising and encouraging.

7. Summary and Conclusion

We have tried to show that there is a strong need for highly multilingual text mining applications (10, 20 or more languages), but that most available and operational systems cover only one or a small number of languages. Assuming that this is mostly due to the fact that the development of natural language processing tools for each language is time-consuming and expensive, we asked the question how the development effort per language can be minimised. The major tips and ideas we found in publications and personal discussions with multilingual system developers (see *Acknowledgements*, below) are: (a) keep your system modular; (b) keep the system simple, not only from a user's point of view, but also from that of the developer; (c) try to use uniform input and output structures; (d) use shared token classes, ideally based on surface-oriented features; (e) try to share grammar rules and lexical resources between languages; and (f) try to be minimalistic by providing and using only the type of information really needed for the application, rather than filling the whole paradigm (e.g. use partial dictionaries rather than trying to produce a complete lexicon for a language). Several developers of multilingual parsers furthermore pointed out the advantage of (g) adhering to grammar theories, as these allow to stipulate general principles that apply to whole groups of languages, i.e. another type of grammar sharing. From an architectural point of view, however, the point was made that a theory-neutral approach is more flexible and lends itself more to a reuse of resources. While developing various text mining tools in up to twenty

[15] http://www.issco.unige.ch/en/research/projects/MULTEXT.html
[16] See http://nl.ijs.si/ME/
[17] See http://www.illc.uva.nl/EuroWordNet/
[18] See http://earth-info.nga.mil/gns/html/
[19] See http://europa.eu/eurovoc/
[20] See http://projects.ldc.upenn.edu/LCTL/
[21] See http://www.elda.org/
[22] See http://www.ldc.upenn.edu/
[23] See http://www.elra.info/
[24] See http://www.flarenet.eu/
[25] See http://www.clarin.eu/
[26] See http://www.clef-campaign.org/
[27] See http://www.enabler-network.org/
[28] See http://telri.nytud.hu/

languages for the *Europe Media Monitor* (EMM) family of applications, we furthermore got convinced that it is useful and efficient (h) to write language-independent rules that make use of information stored in language-specific parameter files; (i) to under-specify wherever possible, in order to save time and not to use restrictions that may get in the way in another language.

In the case of EMM tools, these requirements basically mean that as little language-specific linguistic resources and tools as possible should be used. Instead, we limited ourselves to work with restricted word lists, lookup procedures, machine learning and bootstrapping methods. Such simple means are rather restrictive and challenging. To show what can and what cannot be done adhering to these restrictions, we sketched the solutions adopted in a few of our own multilingual text mining applications.

We saw that machine learning solutions are particularly promising to achieve high multilinguality, but that the need for pre-tagged training data limits at least supervised learning methods to those few languages for which tagged corpora are available. Semi-supervised or unsupervised methods can more easily be adapted to lesser-used languages, for which few linguistic resources exist.

We finally presented our own – probably unrealistic – opinion regarding the types of linguistic resources that would be useful to allow the computational linguistics community to develop more highly multilingual text mining applications more quickly, and why. These resources can be described as *freely available, simple, parallel and uniform multilingual dictionaries, corpora and software tools*. The number of current efforts and projects to produce multilingual resources shows the positive trend and the willingness to produce multilingual language resources.

There is more than one possible solution to overcome the multilinguality barrier, and each application has its own specific requirements. The proposed ideas may thus not suit all needs. We hope, though, that this collection and discussion of ideas and insights may nevertheless be useful for multilingual system developers.

8. Acknowledgements

I would like to thank the following persons for having shared their own multilingual grammar writing experience with us, or their views on linguistic resources: Kalina Bontcheva (Sheffield University) on *GATE;* Frédérique Segond, Caroline Hagège and Claude Roux (Xerox Research Centre Europe) on the *Xerox Incremental Parser*; Aarne Ranta (Gothenburg University) on the *Grammatical Framework*; Jacques Vergne (Caen University) on sentence chunking using extremely light-weight methods; Eric Wehrli (Geneva University) on his deep-linguistic parser; Gregory Grefenstette (Exalead) and Gregor Thurmair (Linguatec) on their respective multilingual products; Khalid Choukri (ELRA/ELDA) and Gregory Grefenstette on linguistic resources; and my JRC colleagues Maud Ehrmann, Vanni Zavarella and Hristo Tanev for sharing their experiences and for their feedback on earlier versions of the paper. The ultimate responsibility for any errors, however, lies with me.

I would furthermore like to thank my superiors Erik van der Goot and Delilah Al Khudhairy for their support, and my colleagues in the OPTIMA group at the JRC for the fruitful and efficient collaboration over the past years, and for so reliably providing large amounts of clean multilingual news data, which allowed us to run many multilingual experiments. My special thanks go to my former colleague Bruno Pouliquen (now at WIPO in Geneva). We developed most ideas together, and he very efficiently implemented many ideas and integrated the many tools with each other.

9. References

Balahur-Dobrescu Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen & Jenya Belyaeva (2010). Sentiment Analysis in the News. *Proceedings of LREC*. Valletta, Malta.

Bender Emily & Dan Flickinger (2005). Rapid prototyping of scalable grammars: Towards modularity in extensions to a language-independent core. In *Proceedings of IJCNLP*, Jeju Island, Korea.

Bering Christian, Witold Drożdżyński, Gregor Erbach, Lara Guasch, Peter Homola, Sabine Lehmann, Hong Li, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, Atsuko Shimada, Melanie Siegel Feiyu Xu & Dorothée Ziegler-Eisele (2003). Corpora and evaluation tools for multilingual named entity grammar development. *Proceedings of the Multilingual Corpora Workshop at Corpus Linguistics*, pp. 42-52, Lancaster, UK.

Carenini Michele, Angus Whyte, Lorenzo Bertorello & Massimo Vanocchi (2007). Improving Communication in E-democracy Using Natural Language Processing. In: *IEEE Intelligent Systems 22:1*, pp 20-27.

Gamon Michael, Carmen Lozano, Jessie Pinkham & Tom Reutter (1997). Practical Experience with Grammar Sharing in Multilingual NLP. In *Proceedings of ACL/EACL*, Madrid, Spain.

Gong Y. and X. Liu (2002). Generic text summarization using relevance measure and latent semantic analysis. In: *Proceedings of ACM SIGIR*, New Orleans, US.

Goutte Cyril, Nicola Cancedda, Marc Dymetman & George Foster (2009). *Learning Machine Translation*. MIT Press, Cambridge, USA.

Grefenstette Gregory (2010). *Proposition for a web 2.0 version of linguistic resource creation*. Presentation at FLaReNet Forum 2010 in Barcelona on 12.02.2010.

Ignat Camelia, Bruno Pouliquen, António Ribeiro & Ralf Steinberger (2003). Extending an Information Extraction Tool Set to Central and Eastern European Languages. In: *Proceedings of the Workshop Information Extraction for Slavonic and other Central and Eastern European Languages* (IESL'2003), held at RANLP'2003. Borovets, Bulgaria, 8 - 9 September 2003.

Larkey Leah, Fangfang Feng, Margaret Connell, Victor Lavrenko (2004). Language-specific Models in Multilingual Topic Tracking. *Proceedings of the 27th annual international ACM SIGIR conference on*

Research and development in information retrieval, pp 402-409.

Linge Jens, Ralf Steinberger, Thomas Weber, Roman Yangarber, Erik van der Goot, Delilah Al Khudhairy & Nikolaos Stilianakis (2009). Internet Surveillance Systems for Early Alerting of Health Threats. *EuroSurveillance Vol. 14, Issue 13*. Stockholm, 2 April 2009.

Maynard Diana, Valentin Tablan & Hamish Cunningham (2003). NE Recognition without training data on a language you don't speak. In: *Proceedings of the ACL Workshop on Multilingual and Mixed-Language NER: Combining Statistical and Symbolic Methods*. Sapporo, Japan.

Maynard Diana, Valentin Tablan, Hamish Cunningham, Christian Ursu, Horacio Saggion, Kalina Bontcheva & Yorick Wilks (2002). Architectural Elements of Language Engineering Robustness. *Journal of Natural Language Engineering 8:3*, pp 257-274. Special Issue on Robust Methods in Analysis of Natural Language Data.

Nadeau David & Satoshi Sekine (2009). A survey of entity recognition and classification. In: Satoshi Sekine & Elisabete Ranchhod (eds.): Named Entities – Recognition, classification and use. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Nivre Joakim, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel & Deniz Yuret (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pp. 915-932, Prague, Czech Republic.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. In: *Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1-2*, pp. 1-135, 2008.

Pastra Katerina, Diana Maynard, Oana Hamza, Hamish Cunningham & Yorick Wilks (2002). How feasible is the reuse of grammars for Named Entity Recognition? *Proceedings of LREC*, Las Palmas, Spain.

Pouliquen Bruno, Ralf Steinberger & Clive Best (2007a). Automatic Detection of Quotations in Multilingual News. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing* (RANLP'2007), pp. 487-492. Borovets, Bulgaria, 27-29.09.2007.

Pouliquen Bruno, Ralf Steinberger, Jenya Belyaeva (2007b). Multilingual multi-document continuously updated social networks. *Proceedings of the Workshop* Multi-source Multilingual Information Extraction and Summarization (MMIES'2007) held at RANLP'2007, pp. 25-32. Borovets, Bulgaria, 26 September 2007.

Ranta Aarne (2009). The GF Resource Grammar Library. In: *Linguistic Issues in Language Technology, LiLT 2:2*. December 2009

Rayner Manny & Pierrette Bouillon (1996). Adapting the Core Language Engine to French and Spanish. *Proceedings of the International Conference NLP+IA*, pp. 224-232, Mouncton, Canada.

Steinberger Josef, Mijail Kabadjov, Bruno Pouliquen, Ralf Steinberger & Massimo Poesio (2009). WB-JRC-UT's Participation in TAC 2009: Update Summarization and AESOP Tasks. In: *Proceedings of the Text Analysis Conference 2009* (TAC'2009). National Institute of Standards and Technology, Gaithersburg, Maryland USA, 16-17 November 2009.

Steinberger Ralf, Bruno Pouliquen & Camelia Ignat (2008). Using language-independent rules to achieve high multilinguality in Text Mining. In: Fogelman-Soulié Françoise, Domenico Perrotta, Jakub Piskorski & Ralf Steinberger (eds.): *Mining Massive Data Sets for Security*. pp. 217-240. IOS Press, Amsterdam, The Netherlands.

Steinberger Ralf, Bruno Pouliquen & Erik van der Goot (2009). An Introduction to the Europe Media Monitor Family of Applications. In: Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): *Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop* (SIGIR-CLIR'2009), pp. 1-8. Boston, USA. 23 July 2009.

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.

Steinberger Ralf, Sylvia Ombuya, Mijail Kabadjov, Bruno Pouliquen, Leo Della Rocca, Jenya Belyaeva, Monica de Paola & Erik van der Goot (submitted). Expanding a multilingual media monitoring and information extraction tool to a new language: Swahili. In *Language Resources and Evaluation – Special Issue on African Language Technology*. Springer, Netherlands.

Tanev Hristo, Vanni Zavarella, Jens Linge, Mijail Kabadjov, Jakub Piskorski, Martin Atkinson & Ralf Steinberger (2009). Exploiting Machine Learning Techniques to Build an Event Extraction System for Portuguese and Spanish. In: *linguaMÁTICA – Revista para o Processamento Automático das Línguas Ibéricas N.2*, pp. 55-66.

Vergne Jacques (2002). Une méthode pour l'analyse descendante et calculatoire de corpus multilingues : application au calcul des relations sujet-verbe. In : *Proceedings of TALN 2002*. Nancy, France.

Vergne Jacques (2009). Defining the chunk as the period of the functions length and frequency of words on the syntagmatic axis. In: *Proceedings of the Language Technology Conference LTC*. Poznan, Poland.

Wehrli Eric (2007). Fips, a "Deep" Linguistic Multilingual Parser. In *Proceedings of the ACL Workshop on Deep Linguistic Processing*, pp. 120-127. Prague, Czech Republic.

Zaghouani Wajdi, Bruno Pouliquen, Mohamed Ibrahim & Ralf Steinberger (2010). Adapting a resource-light highly multilingual Named Entity Recognition system to Arabic. In *Proceedings of LREC*, Valletta, Malta.