European Text Analytics

A white paper prepared for

European Text Analytics Summit 2007

Amsterdam, April 26-27, 2007

(*www.TextAnalyticsNews.com/europe07*)

Seth Grimes

Alta Plana Corporation

Alta Plana

Table of Contents

1 INTRODUCTION: EUROPE 2007

The interplay of languages and cultures drives European social and business interactions to an extent not found in any other region. For Europe, written and spoken words both unify and divide. They are key both to maintenance of national identities and to erasure of artificial barriers to social and economic integration, whether within official bodies such as the European Union, in cross-border business transactions, or in personal encounters.

Europe has a deep and rich scientific, literary, and legal heritage and a tradition of strong governmental support for national and cross-cutting pure and applied research. Given this background and given the role of multilingualism in everyday life – the European Union alone counts 27 member states and 23 official languages – it is no surprise that Europe has nurtured innovation in computational linguistics and similar language technologies.

European innovation has emerged as state-of-the-art text-analytics software solutions. These solutions compete internationally. Europe's contribution to text analytics is quite out of proportion to the continent's size and population, driven by national and international imperatives keenly felt by institutions and individuals.

This paper's author has in the past characterized text analytics as "technology and process both, a mechanism for knowledge discovery applied to documents, a means of finding value in text." Text analytics can be further described as facilitating machine processing of human communications. It enables a more natural human-computer interface. Now that computers are ubiquitous, text analytics offers enormous potential to transform interactions among individuals, business, and governments.

Text analytics opens new vistas for search, the Web's "killer app," and creates new categories of software for fields as diverse as biomedical research, intelligence and law enforcement, corporate operations and communications, and legal and intellectual-property analysis. Europe has led the way, not only in creation of basic text technologies, but also in their application to these and many other important, real-world problems.

This paper will describe text analytics applications of particular European interest. It will survey a number of technologies that originated in Europe. It will present examples of innovative European research projects, programs, and centers and will forecast coming technical and market developments.

As will be seen, the 2007 European text analytics scene is dynamic and robust. European text analytics is holding its own in the international research community and the global marketplace and is evolving to meet rapidly growing European and international demand.

2 TECHNOLOGY BASICS[1]

The term *text analytics* describes a set of linguistic, lexical, pattern recognition, extraction, tagging/structuring, visualization, and predictive techniques. The term also describes processes that apply these techniques, whether independently or in conjunction with query and analysis of fielded, numerical data, to solve business problems. These techniques and processes discover and present knowledge – facts, business rules, and relationships – that had been locked in textual form, impenetrable to automated processing.

Text analytics starts with document acquisition, either targeted retrieval of all material identified by a search or blanket intake of e-mail, Web pages, scientific papers, corporate reports, news articles, and the like. The next step is typically linguistic processing: determining sentence and phrase boundaries, stemming words, determining parts of speech. This step is followed by tagging and extraction of features – entities and their attributes, terms, concepts, sentiments, and relationships – with some form of term normalization and use of lexical analysis to provide frequency counts and the like. Use of taxonomies, lexicons and gazetteers, and machine-learning techniques facilitates this work.

Multilingual intake presents special challenges due to the wide variety of character sets, word morphologies, grammars and syntaxes, vocabularies and idioms, and cultural contexts from which the spectrum of human languages are built. Challenge, however, creates opportunity.

Text-mining tools extract, tag, and analyze associations among identified entities and concepts and the documents that contain them. They create categories or they may apply existing taxonomies – hierarchical knowledge representations – to classify documents, and extracted data may be used for other forms of analysis. They apply statistical techniques to cluster documents according to discovered characteristics. Lastly, they deliver both interactive exploratory capabilities and hooks to allow classification to be embedded in applications to add automated text processing.

The ability to stem words, identify phrases, and extract terms and entities is shared in degrees by search tools, which are, however, built for document retrieval rather than analysis and exploration of document sets. Information extraction, statistical analysis, visualization – none of these functions is present in typical search or content management offerings. Knowledge discovery – pattern recognition – via application of linguistic, statistical, and machine-learning techniques, and via data mining and visualization, is a key differentiator of text mining from those latter technologies.

Because text analytics looks at document sets and identifies interdocument relationships, it supplies context that enables far greater relevance to search results than is provided by search tools. Contextual relevance – the ability to apply domain knowledge to match patterns and cluster results – is a second key technology differentiator. Lastly, text-mining tools can be embedded in applications that produce and consume significant amounts of textual data and often pose real-time operational demands. Content management and enterprise-search tools do not offer the same potential for operational integration.

[1] Material in this section is drawn from The Developing Text Mining Market by Seth Grimes, Alta Plana, (*altaplana.com/TheDevelopingTextMiningMarket.pdf*), 2005.

3 Technology Adoption

Text, as written language, is closely linked with culture. This linkage creates a unique challenge for software that would automate text processing. The software must meet general, global business needs while simultaneously handling narrow, localized concerns particular to given languages and application domains.

Text analytics technical requirements include both internationalization and sufficient specificity to accept input in any human language (whether Western, Middle Eastern, African, or Asian) and input from both specialized domains such as biology, law, and manufacturing and from the Web. This need is typically met with modularized software architectures. Such architectures provide a generalized computing framework that accepts domain- and language-specific *plug-in* components and workflow capabilities that can be matched to application needs.

User adoption of text analytics, as of any specialized computing technology, varies greatly by region, business domain, and organizational profile. Uptake depends on perceptions of need, on availability of solutions, and on the existence of research and business climates that encourage and embrace innovation. For text analytics, Europe meets the three uptake criteria in varying degrees.

European government and industry have long recognized the role that technology can play in bridging language divides to promote social and economic integration and to create opportunities for regional and international business. The European Union sponsors many research programs, as do national and local economic-development agencies. These programs typically involve academia, private-sector commercialization partners, and government adopters.

Given European societal goals, the availability of research-program funding, and Europe's many strong academic institutions, text-analytics research in European academia is on a general par with academic research elsewhere. The Internet has only widened strong collaboration among European and international academics.

Within the government sector, Europe has aggressively adopted text analytics to meet critical national-security and law-enforcement goals. Criminal and terrorist networks are as likely to exploit national borders as they are to find them an impediment, but they do leave an electronic trail. The information they generate, while subject to privacy laws, is easy to exchange, and text analytics is uniquely capable of the data intake, integration, and analysis needed to fight these threats.

Government-prompted adoption of text analytics in business settings – for intellectual-property management, corporate governance, legal discovery, and the like – does appear to lag uptake outside Europe. European business is not subject to requirements directly comparable to U.S. Sarbanes-Oxley corporate governance mandates (which, n.b., do apply to all corporations with 300 or more U.S. shareholders) and eDiscovery information-preservation rules. Such needs have sped U.S. adoption of solutions that automate text processing but have not applied as widely in Europe.

Lastly, while Europe boasts world-class industrial computing-research centers, as well as start-ups that play in the text-rich, exploding social Web, corporate line-of-business adoption in Europe also appears to lag uptake elsewhere. The level of European interest appears very high however, particularly among media companies, publishers, and organizations conducting and consuming applied scientific research. Adoption rates should soon match international levels.

4 EUROPEAN INNOVATION

Europe boasts diverse and vibrant text-analytics research programs and centers, both governmental and academic. The European Commission sponsors noteworthy consortium and university initiatives throughout the continent. Other efforts are supported nationally and by industry. Many of these academic and industrial projects have been built out to create market-leading commercial software tools and services.

4.1 RESEARCH CENTERS

We will provide examples of notable European research centers that together hint at the breadth and orientation of efforts distributed around the continent.

Start with two Italian organizations: the Interuniversity Supercomputing Consortium, CINECA; and the Institute for Computational Linguistics of the Center for National Research, ILC-CNR. **CINECA** unites national-government and academic resources to further an applied-computing agenda in cooperation with industry, including IBM. It is comprised of 28 Italian universities and sponsors programs in many areas, not just text mining[2], serving as a "high-technology bridge between the academic world, research, and the world of industry and public administration." **ILC-CNR**[3] in Pisa has a national and pan-European focus, having participated in dozens of European Commission projects. They consider language to be "the privileged medium for the interactions through which most of social, economical and cultural activities take place and its applications have a fundamental impact in many fields." The institute's work is in design and construction of corpora and lexicons and related ontologies, and in development of tools for Italian-language text processing, information extraction, and domain application.

The European Commission's Joint Research Centre (JRC) is also located in Italy and similarly sponsors text analytics activities. The **JRC** states that they "have been using Language Technology since 1998 to fight the *information overflow* and to *overcome the language barrier*." This work is performed under the auspices of the Information Society Directorate General within the JRC's Institute for the Protection and Security of the Citizen (IPSC). The JRC reports[4] development of a language-technology toolkit with components for:

- Multilingual and cross-lingual retrieval of potentially user-relevant documents.
- Analysis of documents and information extraction with language-neutral representation of information.
- Visualization of documents and document-collection contents.

Their quite-interesting applications work includes the Europe Media Monitor NewsExplorer, which uses JRC-developed technology to automatically generate daily news summaries (*press.jrc.it/NewsExplorer/home/en/latest.html*).

Heading north and west, the National Centre for Text Mining, **NaCTeM**[5], located

[2] CINECA text mining, see: *www.cineca.it/gai/area/textmining.htm*
[3] ILC-CNR, see: *www.ilc.cnr.it/indexflash.html*
[4] JRC Language Technology Activities, see: *langtech.jrc.it/index.html*
[5] An overview of NaCTeM work may be found at *www.ariadne.ac.uk/issue42/ananiadou/*

at the Universities of Manchester and Liverpool, claims title to being the first publicly funded text mining centre in the world, supporting the UK academic community and participating in international research initiatives. NaCTeM was founded in the summer of 2004 and has focused initially on bioscience and mining of biomedical texts. They deliver analytical services via the TerMine term management system, AcroMine for expansion of Medline acronyms, and a TerMine integration with the Cheshire XML search engine; they provide and support software that includes toolkits for the University of Tokyo's GENIA software for information extraction from scientific literature.

Dr. Sophia Ananiadou of the University of Manchester explains that text mining is invaluable to systems biology researchers who need integrated approaches to generate hypotheses, and that the use of text mining technology is a must for facilitating scientific discovery given the amount of textual data generated every day[6]. She quotes an associate, Prof. Douglas Kell, who states that NaCTeM has tapped into this potential with great success:

> Of particular interest to system-biology researchers is synonym detection and linking textual knowledge with existing databases. This problem has been addressed by the TerMine system, which automatically discovers important concepts in text. One of the most impressive outcomes of the work of NaCTeM are the systems MEDIE and InfoPubMed, based on the GENIA tools, which perform semantic text mining based on full parsing. These outcomes are important for the discovery and classification of novel protein-protein interactions, which cannot be detected by high-throughput methods.

As a last example of an academically centered initiative, we cite **PASCAL**[7], *Pattern Analysis, Statistical modeling and ComputAtional Learning*, a Network of Excellence established under the European Commission's Sixth Framework Programme. (Framework Programme 7 started in 2007 and itself has quite interesting elements, for example, 4.2 Intelligent Content and Semantics[8].) The PASCAL objective is "to build a Europe-wide Distributed Institute" to pioneer "core enabling technologies for multimodal interfaces that are capable of natural and seamless interaction with and among individual human users."

4.2 Innovative Projects and Products

Europe is home to projects that have had very broad impact in the text-analytics world. Some of these projects are collaborative efforts that have produced and support software tools and resources that are used worldwide. Others are research efforts that have had considerable success as commercial products. Consider, in the latter category, Xerox's XeLDA[9].

XeLDA, variously the Xerox Engine for Linguistic Dependent Applications or the Xerox Linguistic Development Architecture, is a multilingual engine for transforming, normalizing, and extracting information from text. XeLDA is the product of natural-language processing research at the Xerox Research Centre Europe (XRCE) in Grenoble, done in cooperation with the Palo Alto Research Center (PARC). XeLDA is the linguistic engine within software vendor **TEMIS**'s Insight suite of text-analytics products, which integrates the engine

[6] Ananiadou, S., Kell, D.B. and Tsujii, J. (2006) Text Mining and its Potential Applications in Systems Biology, in *Trends in Biotechnology (TIBTECH), vol.24(12)*, pp.571-579.
[7] For PASCAL information, see: *www.pascal-network.org/*
[8] FP7 information is at *ec.europa.eu/information_society/istevent/koln2007/cf/network-detail.cfm?id=1081*
[9] For XeLDA background, visit: *www.xrce.xerox.com/competencies/past-projects/platforms/xelda.html*

within a framework that includes information-extraction, clustering, and classification tools. TEMIS further offers a product line called Luxid for annotation, integration, and analysis of information extracted from text. A variety of industry vertical applications take advantage of the products' modularized architecture, which supports use of plug-in "skill cartridges." TEMIS's presence is strongest in Europe, reflecting the company's origins, but is growing rapidly worldwide.

As stated, Europe is home to noteworthy collaborative projects. **NooJ** is a quite interesting even though it targets linguistics rather than broader text analytics[10]. According to author Max Silberztein, NooJ provides a development environment for "linguists who wish to formalize a set of linguistic phenomena (from the morphological level to the semantic level), corpus linguists who want to parse large corpora (e.g., for discourse analysis), or computer programmers who want to add NLP functionalities to their applications." Silberztein reports that NooJ, which is freeware, is used by a dozens of European research centers and companies.

With even broader capabilities and worldwide utilization, **GATE**[11], the General Architecture for Text Engineering from the Natural Language Processing Group at the University of Sheffield (UK), is made up of three elements:

- An *architecture* describing how language-processing systems are made up of components.
- A *framework* written in Java and tested on Linux, Windows and Solaris.
- A graphical *development environment* built on the framework.

A quotation provided by project lead Hamish Cunningham illustrates that free, open-source GATE is as much a philosophy as a technology. Cunningham states,

> GATE has made it easier to do good science with repeatable experiments and standardized measurement. It also helps collaboration between researchers sharing software and data, and has shown how public money can have a beneficial impact on a research field and significant commercial interests by being open and, dare I say it, honest. One problem with the research landscape, particularly in fields related to AI, is that measurement and repeatability come a distant second to obscurantist marketing speak, and this is part of a general trend towards universities as sources of "product" to be sold to student or business "clients". We need to recognize the human beings are better at understanding what needs to be done than the blind blunt instrument of profitability, and restructure our scientific programs accordingly.

YALE – Yet Another Learning Environment – is another groundbreaking open source tool[12]. The project is based in Dortmund, Germany, with contributors worldwide who have created a comprehensive knowledge-discovery environment supporting both data and text mining and a large variety of extensions that utilize a modularized, plug-in architecture. YALE project lead Ingo Mierswa and project initiator Ralf Klinkenberg have followed a path now familiar in the open-source world: They offer a commercially licensed version and related consulting services through a start-up company, Rapid-I.

[10] For NooJ information, see: *www.nooj4nlp.net/*

11 GATE pages: http://gate.ac.uk/

12 YALE information is at: *yale.sf.net/*

4.3 Technology and Service Vendors

A number of ground-breaking and market-leading text-analytics products have European roots. The previous section cites **TEMIS**, a French company that commercialized linguistic technology developed by Xerox and has built on a suite of analytical tools and industry plug-ins.

SPSS, a leader in the general analytics market with special strengths in market research, survey analysis, and predictive analytics, is another company whose text-analytics tools are of European origin. SPSS tools acquired those tools when it purchased a French company, LexiQuest, in 2002. LexiQuest's developed their natural language processing (NLP) technology with the support of the European Economic Community, now the European Community, and successfully licensed it to French agencies and prominent French and international companies.

SPSS's "predictive text analytics" technology folds LexiQuest tools into the company's Clementine data-mining workbench. These solutions allow users to analyze text-extracted concepts alongside data from other sources to create and solve integrated predictive models. The company states that the addition of attitudinal and interaction data discovered in text can provide *lift*, ranging from 10 to 50 percent, to results produced from only traditional sources. They claim over 2,000 text-analytics customers worldwide, many of them in Europe.

We next cite the two enterprise search market leaders, both European companies, Fast Search & Transfer (**FAST**) and **Autonomy**. Both companies go far beyond keyword search to support information extraction, content classification, document clustering, and creation and management of taxonomies: text-analytics functions. FAST was founded in 1997 in Oslo and has exhibited a very steep growth curve. Of late, FAST has been particularly aggressive in partnering with business intelligence (BI) vendors and in proposing a search-centric approach to BI, their Adaptive Information Warehouse Platform. Autonomy was founded just the year before and is headquartered in Cambridge. Their growth has also been rapid, fueled in part by acquisitions such as their 2005 takeover of enterprise-search rival Verity. They have a broad product line that includes tools for search of diverse types of media.

Lastly, we cite two European companies that analyze online social media. Both offer services rooted in text mining, delivering results to analyst users in packaged form. Brussels-based **Attentio** describes itself as providing European companies with market monitoring and analysis tools that continuously track a broad range of media for market intelligence about their brands, products, and fiercest competitors. Attentio services track newsgroups, blogs, Web sites that host reviews, and syndicated sources. They detect and report trends, clusters, and sentiments. And **Market Sentinel**, a London company, similarly monitors blogs and the Web for customers in a variety of sectors. One interesting point about these two companies is that they are selling solutions to end users – they are not selling technology.

European technology and service vendors have clearly found a ready market for products rooted in text analytics both at home in Europe and internationally. They are joined by foreign text-analytics software publishers who have established European offices and partner networks – Insightful, Inxight, Megaputer, Nstein, and SAS are among them – because they perceive the value and potential of the European market.

5 European Outlook

5.1 The Market for Software and Services

Technology drives economic development, and involvement with emerging technologies like text analytics can have a multifaceted effect of boosting both the academic and industrial sectors and of benefiting both society and business adopters. Witness already-cited research initiatives as well as regional and localized programs. ITI Life Sciences' text-mining R&D program provides an example of the latter, a £5.3 million, three-year investment in construction of software for information extraction from scientific and medical literature. The University of Edinburgh (UK) and a subsidiary of U.S.-based biological- and chemical-information management company Cognia are conducting the work.

Biomedicine, manufacturing (e.g., quality monitoring), and governmental applications for intelligence, counterterrorism, and law enforcement derived the earliest significant benefits from text-analytics technology. The last two years, 2005 and 2006, saw increased uptake in areas such as survey analysis, media monitoring, patents, and publishing. Publishers are interested in generating content databases for data mining and in delivering increasingly sophisticated information-retrieval capabilities. Current growth areas include monitoring and analysis of networked social media for applications such as reputation management. We should also expect significant growth in Europe as elsewhere in the integrated analysis of text and numerical data, the embedding of advanced text-processing capabilities in line-of-business applications, and the practical emergence of question-answering as a high-end search product.

Introduction of text analytics for users in most segments will continue to involve vendor professional-services and consulting engagements as a high proportion of overall implementation cost.

5.2 Market Estimation

The market for text-analytics software and services is substantial and growing rapidly. TEMIS cofounder Alessandro Zanasi of the Wessex of Institute of Technology (UK) offered in mid-2006 that "data and text mining is an expanding field and constitutes a market estimated to be more than US$12 billion."

The estimate of the author of this paper is that the worldwide market for text-analytics software from pure-play vendors for licenses, support, and professional services plays out to about US$200 million for 2005. Pure-plays include ClearForest, Inxight, Nstein, TEMIS, and Teragram. Annual growth is in the 25 to 40 percent range. We further allocate to text analytics a portion of overall revenue of UK-based Autonomy, SPSS, SAS, and other companies that sit part-way in the text-analytics market, applying a conservative multiplier of estimated license revenue to compute user/contractor labor, and we add the assumed value of in-house text analytics at companies such as Factiva, Thomson, and Reed Elsevier. The sum is a US$2 billion worldwide text-analytics market.

Europe is likely 30 to 40 percent of the worldwide text-analytics market. Interest is widespread in Asia but adoption lags that of North America and Europe. Therefore, the European demand-side (user) market proportion may dip in coming years, but Europe's share of the supply side of the market, or software sales and service delivery, should grow.

Seth Grimes

Seth Grimes is a business intelligence, data warehousing, and decision systems expert, a consultant and contributing editor for *Intelligent Enterprise*. He writes the magazine's Breakthrough Analysis column and contributes feature articles.

Seth founded Washington DC-based Alta Plana Corporation in 1997 and consults on information systems strategy for clients that include government statistical agencies, marketing firms, software publishers, and analytics users.

Seth writes and speaks on data management and analysis systems, industry trends, and emerging analytical technologies. He is chair of the European Text Analytics Summit and chaired the 2005 and 2006 summits in Boston and will chair the 2007 summit. His white paper for the 2005 summit, The Developing Text Mining Market, is available online at *altaplana.com/TheDevelopingTextMiningMarket.pdf*, and his 2006 summit paper, Finding Value in Text Analytics, is posted at *altaplana.com/FindingValueInTextAnalytics.pdf*.

Seth can be reached at grimes@altaplana.com, +1 301-270-0795.

European Text Analytics Summit 2007

The European Text Analytics Summit 2007 (*www.textanalyticsnews.com/europe07*), slated for April 26-27, 2007 in Amsterdam, is a mindshare event for the leading developers, researchers, vendors, tech-savvy users, and newcomers to the text-analytics space.

It is the first European Text Analytics Summit, following on the heels of two highly successful Text Analytics Summits held in Boston in June 2005 and June 2006. Analyst Curt Monash wrote in *ComputerWorld* that "the [2005] Text Mining Summit ... was one of the best conferences I've been to in a long time." SPSS Vice President Olivier Jouve called the 2006 summit "the best conference I attended last year." The 2007 Text Analytics Summit is scheduled for mid-June 2007, again in Boston.

The European and North American Text Analytics Summits both provide an opportunity for researchers and vendors to identify promising applications, size up technical challenges, and connect with users eager to keep up with market developments. Text-analytics users and prospective users in any application or industry find an unmissable opportunity to learn from peers and understand the bottom-line impact of the latest deployments. Developers and marketers benefit from the opportunity to engage end users and technologist to better understand market requirements, technology developments, and product directions.

Alta Plana Corporation

7300 Willow Avenue
Takoma Park, MD 20912
+1 301-270-0795
altaplana.com

