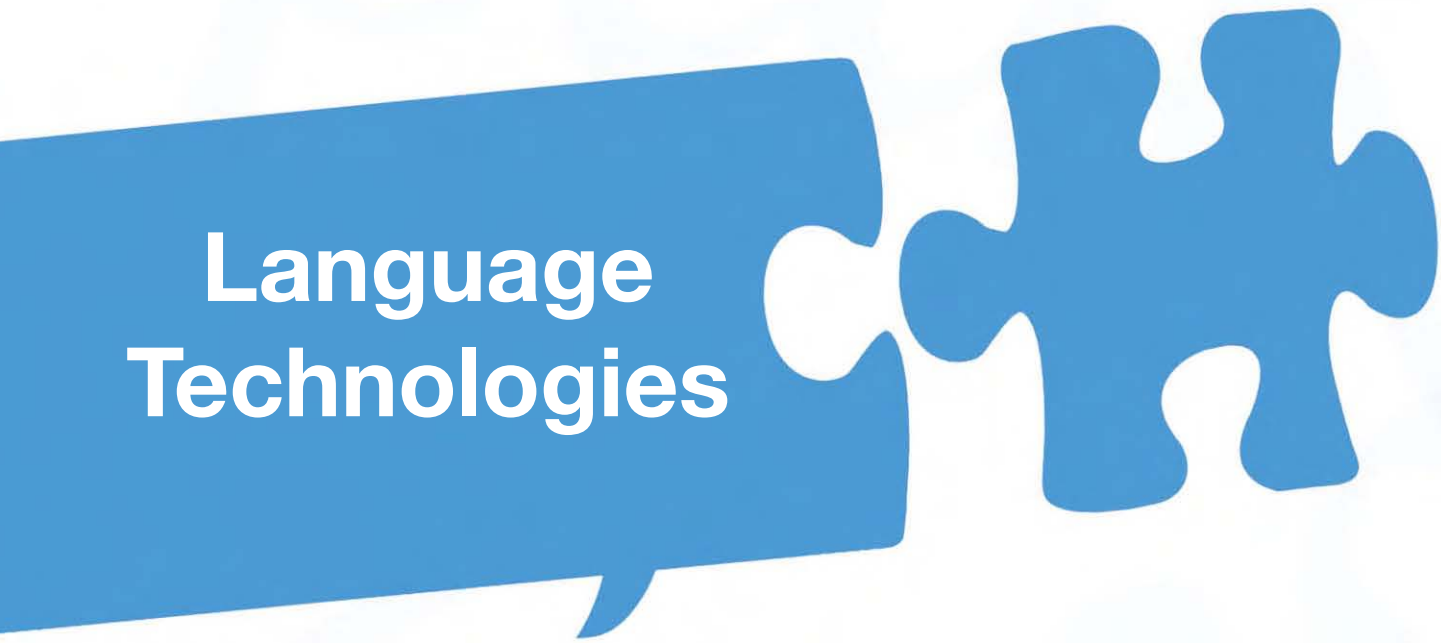
LT-INNOVATE.EU

THE FORUM FOR EUROPE'S LANGUAGE TECHNOLOGY INDUSTRY



Language Technologies

LT2013:

Status and Potential of the European Language Technology Markets

March 2013

www.lt-innovate.eu - contact@lt-innovate.eu

TABLE OF CONTENTS

LT-Innovate is the Forum for Europe's Language Technology Industry, a not-for-profit organisation representing mostly SMEs involved in developing products using intelligent content, speech and translation technologies. LT-Innovate was founded in January 2012. As of 1 November, it gathers 115 LT suppliers in 22 countries, as well as several dozens of other LT stakeholders. The European Language Technology industry generated an aggregate turnover of 19.3 billion € in 2011. LT is a very dynamic industry, with a yearly growth rate in excess of 10%.

Executive Summary

This LT-Innovate report provides a comprehensive survey of the state of the Language Technology (LT) market in Europe today and projections for the next five years. It is divided into six parts, covering global trends in the ICT ecosystem, an analysis of specific trends in the LT industry, an exploration of innovation options for European LT companies, and a detailed account of the three strategic technology segments of speech interaction, multilingual communication and translation, and intelligent content that make up the LT market.

Mobile communications, cloud service models and social media are transforming the way citizens, companies and public administrations act in the digital world. This report identifies three deep trends driving next-generation ICT that will open up significant opportunities for LT:

- **Unified Communication**: cross-platform, multimodal and multilingual. Mobile connectivity and service unification across devices and platforms will offer business and consumer users seamless communications

- **Unified Information Access**: in any language and across languages. This will remove barriers to content and enable integrated messaging, conferencing, collaboration, content - and data-sharing based on intelligent content and applications, multilingual and interactive systems and technologies.

- **Unified User Experience**, based on natural interaction with machines and processes, in any language. This will remove barriers to the access, use and understanding of information from large volumes of unstructured, semi-structured, and structured data.

LT is *the* critical enabling technology for each of these fundamental trends and stands to benefit from the emerging interconnections between interaction (speech), information processing (intelligent content) and automatic translation in a multilingual connected digital space. It is therefore vital for the European LT industry to embrace and foster the opportunities the rapidly evolving ICT eco-system offers and to pursue a dynamic innovation agenda ahead of its competitors globally.

LT-Innovate has developed a market model to estimate the size of LT market in terms of sales and services. The worldwide LT market is worth around €19.3B today and should grow to nearly €30B by 2015. The European speech technology market is growing by 9.7%; and should grow to €8.6B by 2015. The intelligent content market is set to grow to €6.2B. The translation technology market is worth some €8.6B and should grow to €14.9B. The growth rates in the “Rest of the World (ROW)” markets should be significantly higher than in Europe and the Americas as these emerging markets mature. The translation technology segment will continue to dominate the European LT market.

In terms of market participants, there are some 500 European companies actively developing or integrating LT, most of them still small companies and all too often, focus on niches in their national (language) markets.However, the European LT industry is gradually moving LT up the value chain into mainstream applications and markets. Furthermore, the gaps in language coverage for speech and content technology, and the potential to create a demand-driven dynamic holds significant potential for growth of the LT industry across Europe.

To facilitate this strategic growth, it is suggested that the pace of development could be accelerated through collaborative innovation bringing together LT companies with their peers and other corporate actors and buyers across the ICT value chain. Various scenarios for this process are explored in the Report

The final three sections analyse in detail the history, companies and product/service offerings in the key LT segments of speech, translation and intelligent content technology, providing a guide to key players and their role for the three different application areas.

Key Extracts

A new Ecosystem

New open paradigms, language-neutral development platforms and multilingual development resources could foster disruption, particularly in Europe. (p. 5)

The ability to manage and process the tsunami of data across the world's languages is one of the biggest challenges in the new ICT ecosystem, and one for which LT is a critical enabling technology. (p. 6)

LT is baked into the future of ICT in the mobile/social/global world of computing. (p.7)

In the era of semantics – when we need to know the meaning of the data that flows around the digital universe – Language Technology is essential for innovation. (p.8)

Although LT has been a commercial market for many years, only recently have technological conditions made it possible to exploit LT on a large scale. (p.9)

Markets for Language Technology

The fastest growth is in non-European languages, though Spanish and Portuguese gain significance because of Latin American markets. Aside from English, Spanish and Portuguese, only five other EU languages (German, French, Italian, Polish and Dutch), out of 60 or more spoken in the Union, are published on more than 1% of the top million sites. (p. 10)

While the potential is for a single European digital market with 500+ million customers, the reality is a series of fragmented linguistic markets, none bigger than possibly 70/80 million customers, most much smaller. (p. 11)

At present no company or website could be genuinely global using the localisation techniques currently at our disposal. Only with large-scale automation will the limited multilinguality of the web be transformed into a genuinely globally accessible medium. (p. 11)

Where language is the very stuff of our digital system – customer interactions, employee conversations, technical and scientific knowledge, cultural and social objects of all kinds – the era of the Lingua Franca is over. Interacting across the many languages of the digital world is no longer optional. (p. 12)

Europe's share of the worldwide market will increase slightly to 38% over the five year period. However, that share is significantly lower (24% in 2015) for the software portion of the market... Factors that could change this include:

- Faster and more extensive deployment of content applications in more European languages, in a coherent framework for all languages
- Development – and integration – of speech components (for recognition, generationand identification/verification) in more European languages, affordably available for European application and solution developers
- Large-scale deployment of open source machine translation in open environments using shared resources
- Large-scale sharing of resources (paid and free) throughout the European industry
- Development of vertical and industry-specific platforms for LT development and deployment, engaging

whole industries in cooperative initiatives (analogous to SWIFT in banking) (p. 21)

Collaboration between the industry and data owners will be needed. (p. 22)

Many IT managers are still relatively unaware of the benefits that LT can provide them... Suppliers, LT vendors and IT integrators should work closer and harder to identify killer business cases, increase market awareness and deploy market strategies understanding how economic return affects clients, developing modular/incremental products, and forging cross-industry alliances for to strengthen market channels. (p. 24)

LT-Innovate estimates that there are around 500 companies in Europe either actively developing Language Technology, or embedding its features in their products and services in an innovative way... The industry comprises mostly small companies, concentrated in the western and northern regions of the EU, with a mix of long-established players but also a significant number of new entrants. A quarter of the companies are micro-enterprises with fewer than 10 employees, while only 6% have more than 200 employees; almost the entire industry is composed of SMEs... Over half the industry comprises companies active for more than 10 years, many that remain small. The fact that so many companies fail to scale, even after years in business, is unusual in a technology industry, and indicative of the market context for LT in Europe: local/national companies with expertise in local languages serve local markets with services based on their own languages. This state of affairs is not likely to be sustainable, as cloud-based language-enabled services are launched on a large scale. At present, few European companies are in a position to compete in an ecosystem where access to technology, rather than narrow linguistic expertise, is the driving factor. (p. 25)

Innovation in the LT Industry

The dynamics of the general software market, and the limits of what is currently possible for niche LT SMEs, strongly suggest that a Digital Language Infrastructure for Europe could both unlock potential for the industry, and help meet the need for pervasive “multilinguality” in Europe’s digital economy. The industry itself should define the nature and content of the infrastructure, what features are appropriately shared and open, what should remain in the commercial IP realm. (p. 33)

The review of conditions in the LT industry suggests that collaborative approaches to the market could break through the fragmentation that is evident.(p. 33)

Asymmetric partnering for SMEs is a natural route to developing technologies in specialist areas with steep technical demands (heavy R&D), where domain expertise is key. Peer partnering takes the alternate route of creating new “breakout” categories of products or services through the collaborative combination of complementary technologies... Dominant markets are those where technical depth meets the greatest opportunity. (p. 38)

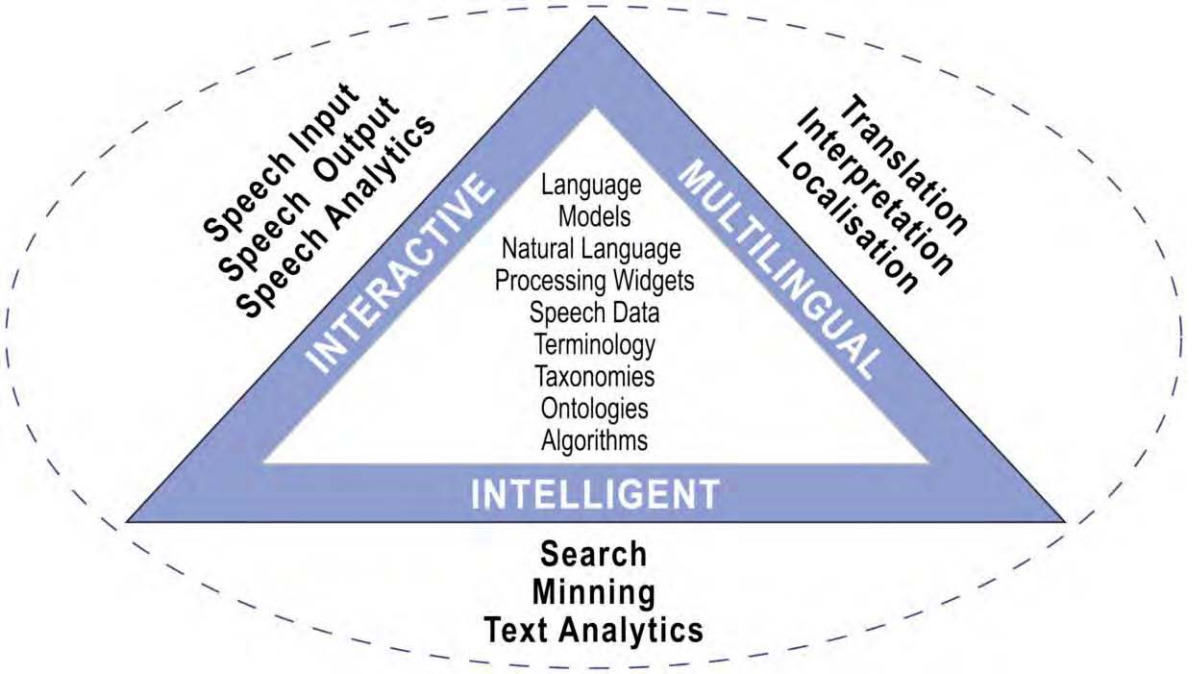
Interactive, Multilingual, Intelligent

The speech applications market shows immense potential, and it is expected to grow rapidly in the next few years... The market is mainly driven by the increased demand in the Mobile Devices segment. This segment is witnessing high demand for speech recognition applications because of the increase in the number of regulations on the use of mobile phones while driving. (p. 47)

We can debate whether the translation industry’s response to rapid globalisation and growth in content has been the right one. Has the industry made best use of technology to raise its capacity and stay profitable? Or has the content explosion marginalized an industry of artisans? (p. 55)

Intelligent content refers to content that is structurally rich and semantically aware, and is therefore discoverable,reusable, reconfigurable and adaptable and which is not limited to one purpose, technology or out-

put. These technologies rely on underlying techniques and tools such as natural language processing (NLP), categorization and clustering engines, and statistical approaches for processing the outputs of human language, such as written or spoken texts. (p. 83)



1. A New ICT Ecosystem

1.1 The deep transformation of the competitive landscape

A new ICT ecosystem is rapidly evolving, based on wide-scale connectivity for mobile access and social computing via remote cloud services, generating and using vast volumes of digital information end enabling the unification of the user experience on digital networks. These are the conditions in which LT companies operate, and they will influence the future direction of the industry.

In the next two years…

In 2012 digital content has grown to 2.837 zettabytes, up almost 50% from 2011, on its way to 8.5ZB by 2015. Big Data technologies[1], tools, and services that turn this information overload to information gains are the next opportunity for competitive advantage, and ***LT is a core Big Data technology.***

By 2014, the number of intelligent communicating devices on global networks will outnumber traditional computing devices by almost 2 to 1, with ***"intelligence" largely driven by the semantics of LT***. This will change the way people think about interacting with each other, and with their devices.

A growing number of ICT solutions will be built on the next-generation cloud platforms producing high-value, vertically-focused solutions. ***LT is premium enabling technology for vertical solutions***, as it employs semantically rich domain knowledge.

Over 80% of new software applications will be distributed/deployed on clouds, making the packaged software market increasingly obsolete. Legacy packaged Enterprise apps will start migrating to clouds, but the race for innovation is not necessarily dominated by legacy global software suppliers. ***Cloud platforms level the playing field for innovative LT suppliers***.

Cloud services spending will hit $60 billion, growing at 26% a year. This is still less than 10% of IT spending, but with over 50% of customers transitioning to the cloud, the huge strategic impact of cloud competencies is obvious. Amazon Web Services will join the $1 billion IT cloud services club and Google Enterprise is right behind. The opportunity for innovative cloud platforms is wide open. ***Cloud platforms unlock the potential for ubiquitous embedding of LT across the ICT landscape***.

Spending on mobile devices will grow 23%, driving 43% of IT growth; a mobile strategy is priority number one for all industry players in 2012. Mobile device spending will exceed PC spending, growing 4 times as fast. ***Intelligent interactivity using LT is a signal feature of the new mobile market.***

Mobile data services will exceed fixed-line data, growing over twice as fast. Mobile operators are on the front lines of mobile-device-driven opportunities; a sustainable model for funding network growth will demand new-generation services. ***Consumers have demonstrated their appetite for the personalised features LT delivers***.

Over 700 million smart phones and tablets will be shipped in 2012, a jump of 34%, nearing double PC shipments. This growing advantage in volumes is a major attraction for developers and a key factor in the PC-versus-mobile device war. Establishing ***LT-based dominance on mobile devices*** (speech interfaces, intelligent assistants, language translation, semantic awareness of location, context) is a strategic position in this war.

[1] Definition of Big Data Technology as well as its differences with Business analytics is provided in the report Annex.

In 2012, 1.5 million mobile apps will be available, over 15 times the number of PC apps. Amazon's Kindle Fire will gain a nearly 20% share of media tablets, and the number of Android apps will finally exceed those for Apple's iOS. The iPad will still dominate (62% share), but serious branded competition is arriving. This is a test of the power of «open» versus «closed» environments, still a close call in 2012. ***Open platforms for LT-enabling could help shift the power dynamics in this market.***

85 billion mobile apps will be downloaded in 2012, an increase of 122%; the 8% of mobile apps that are paid for will generate more revenue than mainframes. Competition between today's mobile app platforms and stores can be disrupted, as 15% of new mobile apps launched in 2015 will be based on HTML5 and developers try to bypass native OS fragmentation and distribution handcuffs. ***New open paradigms, language-neutral development platforms and multilingual development resources could foster disruption, particularly in Europe.***

In the next five years…

The hallmarks of next-generation ICT will be mobile connectivity and service unification across devices and platforms, with features common to both business and consumer users. Unified Communication will enable users to experience the social, communicative features of ICT through integrated messaging, conferencing, collaborating, content- and data-sharing based on "presence" information, intelligent applications, translation systems and speech technology. ***Unified Communication is cross-platform, multi-modal and multilingual.***

Unified Information Access will remove barriers to the access, use and understanding of information using highly scalable platforms that integrate large volumes of unstructured, semi-structured, and structured information into a unified environment for processing, analysis and decision-making. Unified Information Access will be built on hybrid architectures that combine database, search, reporting, visualisation and translation technologies. Intelligent multilingual tools support ***Unified Information Access in any language, and across languages.***

Standardised user interfaces will give way to a Unified User Experience that is customised, personalised, and "contextual"; online and offline experiences will converge through the use of mobile, Internet of things, Near Field Communications, social and hyper-connectivity, and ubiquitous connection to social networks, using social browsers or navigating from within applications. Intelligent personal assistants, with naturalistic voices and able to speak many languages, will mediate between users and their devices and data. ***The Unified User Experience is based on natural and convenient interaction with machines and processes, in any language.***

1.2 A Mobile/Social/Global Ecosystem

The robustness of the mobile market is driving the evolution of the new ecosystem. The always-on internet connection in consumers' pockets has spurred enormous growth in web usage, changed the way people shop, and vastly increased content consumption. Mobility is driving applications software and data into the Cloud, reducing location dependency, increasing the need for "presence" data (who is where, doing what, in what context).

Social media represents a paradigm shift with enormous consequences for all types of businesses and organisations, and especially for consumer-facing companies whose systems are shifting from channelled experiences and brand-controlled messages to empowered consumers in a channel-agnostic marketplace. This is what Geoffrey Moore calls the move from "systems of record" to "systems of engagement", another way to describe the transition from information-centric to interaction-centric ICT. While this is most visible in the consumer space, exactly the same tendencies are transforming internal business functions, sometimes called the "consumerisation" of the Enterprise.

Consumers who used to browse portals or interact with the Web using search (e.g. from Yahoo to Google) now engage via social media; the time spent online using search is hardly growing at all, while time spent on social networks increased 50% in 2011 (moving from Google to Facebook…or to Google+). For many users,

social media has already replaced email. Soon, more than half of all retail transactions will be influenced by the web. Ninety per cent of consumers now trust peer recommendations online, compared to only 14% who trust advertising. This is driving social commerce in all areas of online business. The same dynamics are at play in citizen-facing systems for the public sector, even in the delivery of healthcare and other social services.

Social media yields social data, where customer conversations create an insight-rich goldmine for businesses, or any organisation that interacts with the public. New transformational customer-centric organisations are emerging; capturing the full value of social data takes place across an entire organisation, often requiring cultural changes. Social data can drive change beyond marketing, impacting sales, customer service, and product development. Internal and external Enterprise social initiatives are designed and evaluated in relation to the larger context of business goals, and the new ICT ecosystem is affecting almost every type of business and organisation.

The rapid expansion of the Internet itself – with approximately 2.3 billion people online - combined with the mobile/social/global paradigm shift, are driving digital information volumes off the map, growing by a factor of 9 in just five years. While 75% of digital information is generated by individuals, Enterprises have some liability for 80% of that information at some point in its digital life. The number of «files» or containers that encapsulate digital information is growing even faster than the information itself as more and more embedded systems generate digital data. The amount of information individuals create themselves — writing documents, taking pictures, downloading music, etc. — is far less than the amount of information being created *about* them.

In this context of global growth in connectivity, the growth of digital information continues to outpace that of storage capacity, or at least of what is currently stored. A gigabyte of stored content can generate a petabyte or more of transient data that is typically not stored (e.g., digital TV signals we watch but don't record, voice calls that are made digital in the network backbone for the duration of a call). The cost of creating, capturing, managing, and storing information is one-sixth of what it was in 2005, yet since then the annual investment by Enterprises in digital technology has increased 50%, spent on hardware, software, services, and staff to create, manage, store, and derive revenues from digital data.

New capture, search, discovery, and analysis tools, almost all enabled with LT features, can help organizations gain insights from their unstructured data (text, voice recordings and transcriptions), which accounts for more than 90% of digital information. These tools can create data about data, or metadata, which is growing twice as fast as the digital universe as a whole. Business Intelligence tools (a $35B global market) increasingly are dealing with real-time data, from charging auto insurance premiums based on where people drive, to routing power through the intelligent grid, or changing marketing messages on the fly based on social networking responses.

The ability to manage and process this tsunami of data across the world's languages is one of the biggest challenges in the new ICT ecosystem, and one for which LT is a critical enabling technology.

The explosive growth of mobile and social platforms is driving the consumerisation of ICT, as new technologies replace old and companies use them to gain competitive and process/performance benefits. The proliferation of consumer devices is the biggest trigger for Enterprise adoption of unified communication solutions. Wireless connectivity is replacing wired; mobile, multi-modal devices are replacing desktop devices; "Dual Persona" mobile devices are replacing separate personal and business mobile devices; "BYOD" (bring your own device) is becoming the norm. Text messaging and automated proactive notification/alerts are replacing telephony. Process-to-person contacts are replacing person-to-person contacts; online self-service apps are replacing customer service phone calls; mobile apps are replacing online self-service apps.

New generation products and services look for engagement by adding intelligence to the ICT infrastructure, employing Language Technology as a core piece of the puzzle. Through intelligent interactivity using speech

technology, managing Big Data with analytics, new paradigms for linking users and information with semantics, and crossing the boundaries of language to do all these things, LT is baked into the future of ICT in the mobile/social/global world of computing.

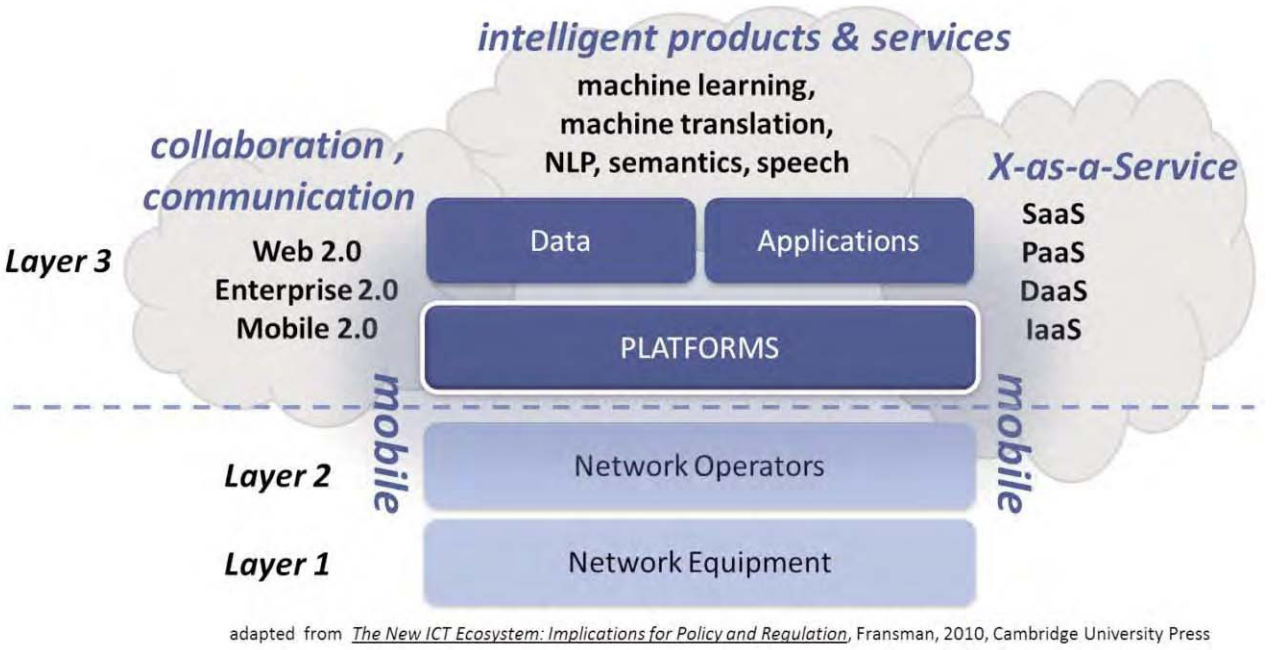


Figure 1:New ICT Ecosystem affecting Europe's ability to compete

The new ecosystem is centred on platforms, applications, and data as the maturing Internet makes the hardware/packaged-software ecosystem obsolete. Platforms are frameworks from which applications are developed and launched, with well-defined access points and rules and on which other players can build applications and services. Hardware Operating Systems created market dominance in the old ecosystem; cloud-based software platforms will have a similar weight in the new ecosystem.

As of 2012, the rising ecosystem platforms are social (e.g. Facebook, Twitter, LinkedIn), mobile (e.g. iPhone, Android phones) and increasingly imbued with services that make them global (e.g. Google and Bing Translate) to use the most dominant examples in the West. In China, similar social media platforms Sina Weibo, Renren, Douban, and many others, have hundreds of millions of users. The US-centric ecosystem is certainly powerful, even in Europe, but it does not control the global internet. Mobile operating systems do control the delivery of mobile apps (and the software platforms for selling them) at least for now, and an industry of Enterprise mobile app-development platforms has grown up around the explosion of mobility at work (not to mention the 1.5 million apps in the "long tail"). But new Web standards may return the centre of gravity to the more neutral Web, with apps that have the graphical and multimodal features needed on the new platforms.

Support for the global aspects of the new ecosystem, as a network of human engagement rather than pure information, includes platforms for language translation, and a vastly increased interest in such products in

consumer markets – as with everything else, driven by the availability of mobile apps that can translate text as well as speech.

Applications in the new ecosystem are emerging, many still at the consumer "long tail" stage of development, and the age of on-site IT using licensed Enterprise software is far from over. But the impact of the new ecosystem on Enterprise software markets is already clear. Social business processes are being defined for next-generation applications. All major global software vendors are developing cloud strategies, and being directly challenged by new entrants. Salesforce.com (whose SaaS CRM platform was the first major disruptor in the Enterprise space) has evolved into a complete Platform-as-a-Service (PaaS) company.

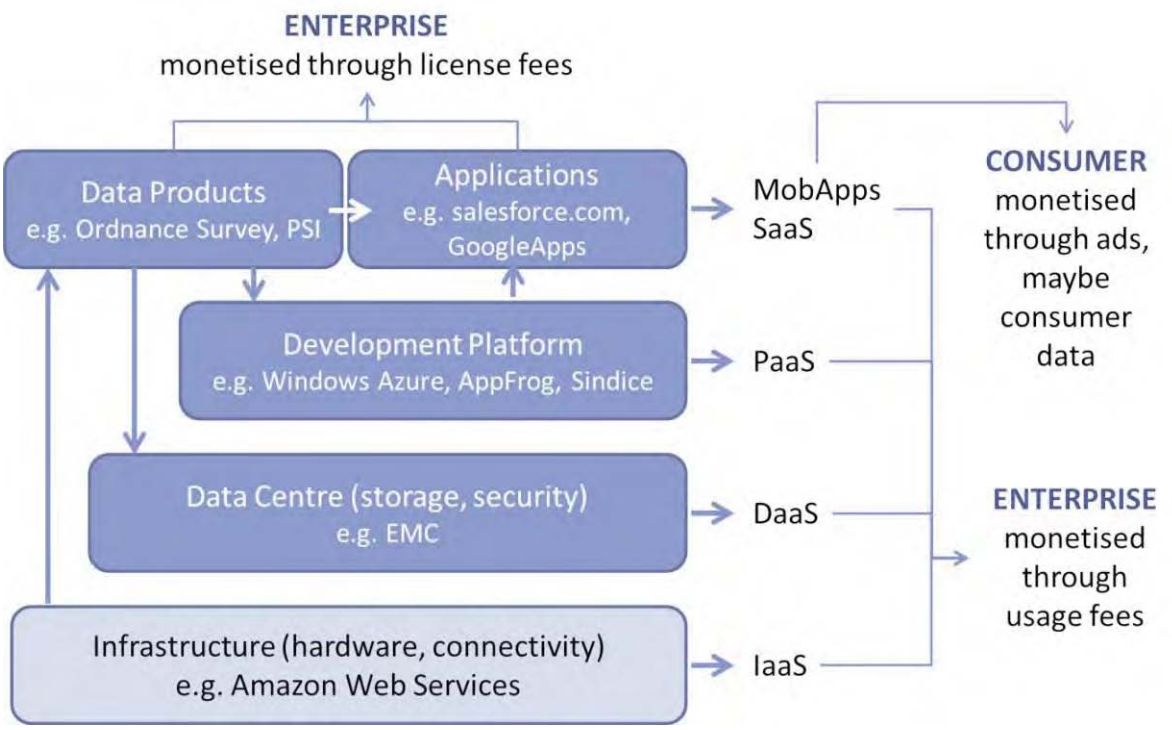


Figure 2: The LT Value Chain

New-generation PaaS offerings provide market access for small, specialist companies that would otherwise struggle for visibility to customers, analogous in some ways to the opportunities for micro-retailers on sites like Amazon or eBay. Expanding the range and scope of platforms that can carry semantics, analytics, speech and translation products is an important opportunity. And alongside these shifts in the applications market, the value of data itself has a growing impact. Platforms, and the applications built on them, are being reconfigured; data will drive innovation in much of the ecosystem, and managing the Big Data that flows from the mobile/social/global web itself is just the beginning: data (big, or not so big) is becoming crucial to innovation.

In the era of semantics – when we need to know the meaning of the data that flows around the digital universe – Language Technology is essential for innovation.

Transformational entrepreneurship will increasingly come from the intersection of engineering, design, and domain expertise. Deep knowledge about a particular field or industry, and the capacity to apply known techniques in new fields, will create value in the new ecosystem. Language Technology is key for such innovation, entwined as it is with human knowledge and the ability to understand and interpret meaning in speech and text.

2. Markets for Language Technology

2.1 The Market Context for LT

Needless to say, the reconfiguration of the ICT landscape has an impact on markets. Language Technology is an unusually broad software category that is engaged in a wide range of product and service sectors, from core technology deeply embedded in complex Enterprise systems, to sophisticated consumer gizmos, to industrial robotics (and everything in between). Although LT has been a commercial market for many years, only recently have technological conditions made it possible to exploit LT on a large scale. In the market conditions surrounding the new ICT ecosystem, take-up of Language Technology is accelerating.

Many LT-based applications are clustered in segments of Enterprise software and services. While the overall Enterprise market for software and services is currently growing at around 4.5%/year (recovering from a contraction of 3% in 2009, and an estimated loss of over a thousand billion Euro in sales during the recession), some segments are growing much faster, driven the trends outlined above. The fastest growth is in Big Data software and services, which is currently expanding at 38%/year, forecast to be worth €13B globally by 2015 according to IDC (starting from practically zero only a few years ago). The exceptional growth rate for Big Data products reflects the strategic needs of Enterprise both to manage the data deluge in all its forms and to exploit the availability of new sources of data to improve operational performance.

Software that embeds LT to manage textual and spoken information and content, and the platforms that process and communicate knowledge, are also high growth markets according to IDC forecasts.[2] For instance the Business Analytics market (including Business Intelligence, Data Warehousing platforms, and Analytics applications) is experiencing 12% growth, and is expected to be worth €36B globally by 2015. Customer Relationship Analytics (part of the Customer Experience Management market with strong growth driven by social data) is growing at 11% heading towards a value of €2.6B. Growth for Search & Discovery software and services is less strong, though still double the industry average, at 7% CAGR, growing to €2.4B in 2015; most of the activity here is in Unified Information Access platforms and applications built around them for particular business domains – the conventional "generic" Enterprise search software market is no longer growing. Content Management (including Document Capture and Imaging, Content and Records Management, Web and Digital Asset Management) is growing at 9%, aiming for €4.8B in 2015. According to TechNavio, the on-demand, cloud-based share of these markets is growing much faster than old-style packaged/licensed products, estimated by them to be 17% a year over the period 2011-2015.

The mobile market is where the convergence of LT features is most visible, e.g. combining speech with intelligent search and navigation systems, or translators. The mobile device market is the largest consumer electronics market globally, experiencing massive innovation in multiple different areas; OEMs, carriers, software vendors, and silicon providers are all scrambling to stay competitive in this rapidly changing space. Major innovation is centred around enhancing the User Interface with emphasis on interactivity using gesture, speech, and facial recognition, and adding intelligent, location-aware services; security, displays, connectivity solutions, and sensors are also areas of high innovation.

Worldwide smart-phone shipments passed a new milestone during the third quarter of 2012, with the total number surpassing 1 billion units, forecast to reach the second billion within three years (according to Strategy Analytics) and experiencing annual revenue growth of 25%/year according to TechNavio. Growth in automotive telematics, which already overlaps with the smart-phone market (e.g. sharing platforms and software) is even faster estimated at 33%/year. Tablet and eReader sales, boosted in part by the increasing adoption of eLearning technology, are growing at 21%/year.

[2] Unless noted otherwise, all Enterprise software market data is supplied by IDC.

Shooting across these shifts in ICT markets is the impact of the "globalisation" of the Internet. A medium that was initially dominated by English language users is fast becoming more representative of the world, though penetration rates still vary significantly by region. Europe (including Russia & Turkey) has a 63% penetration rate according to Internet World Stats. In the EU28 (including Croatia, which will join the Union next year), 67% of the population is now online. Although penetration rates in Asia are only slightly more than a quarter of the population, these users now constitute 45% of the population of the internet. The 275 million internet users in North America are now only slightly more than 10% of the world internet population.

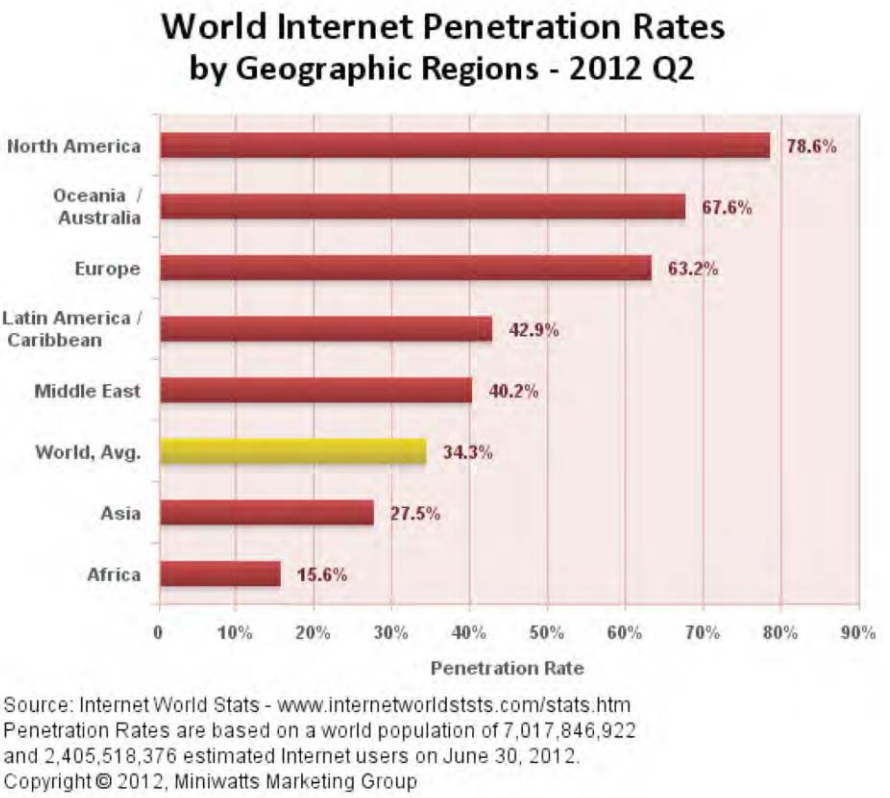


Figure 3: World Internet Penetration Rates (2012 Q2)

Speakers of Asian languages now dominate web usage, but the languages represented by web content are still heavily biased toward English. Data from LanguageConnect show that of the top 1 million websites online, 55% are published in English, only 4.6% in Chinese (many, of course, published in multiple languages); while this remains imbalanced, it is a far cry from the days when 90% of web content was in English, not so many years ago. While English is declining, the share of content published in other European languages is not increasing proportionately; the fastest growth is in non-European languages, though Spanish and Portuguese gain significance because of Latin American markets. Aside from English, Spanish and Portuguese, only five other EU languages (German, French, Italian, Polish and Dutch), out of 60 or more spoken in the Union, are published on more than 1% of the top million sites.

The digital economy is embedded in this sprawling multilingual reality, where the people using the internet, and the content in their languages, are slowly converging with demographic facts.

Linguistic diversity in Europe is a particularly compelling challenge in the age of digital commerce. Although two-thirds of Europeans are online (with much higher penetration rates in some regions), linguistic diversity fractures the unity of that market online. While the potential is for a single European digital market with

500+ million customers, the reality is a series of fragmented linguistic markets, none bigger than possibly 70/80 million customers, most *much* smaller.

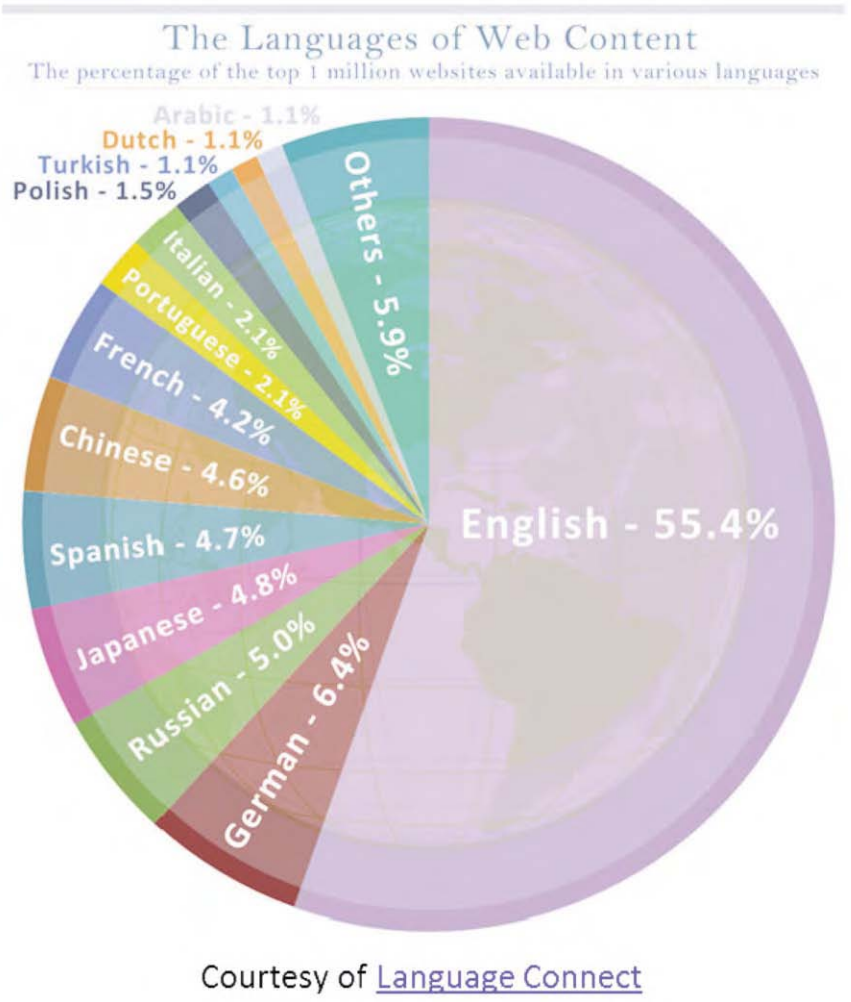


Figure 4: The Languages of the Web Content

The language landscape in which businesses operate is complex, perhaps more so in Europe than most regions. There are 60+ languages spoken in Member States of the EU, and 23 official languages (soon to be 24). Localising websites for the range of languages that might be used in digital commerce is daunting verging on impossible, and only very rich, very large companies can afford to be multilingual on a large scale (at most and very rarely, 25-30 languages globally). Very few European online shopping sites are localised; 82% are published in a single language; 11% offer two languages, and only 2% publish in five or more languages. None could be considered fully multilingual, and indeed at present no company or website could be genuinely global using the localisation techniques currently at our disposal. Only with large-scale automation will the limited multilinguality of the web be transformed into a genuinely globally accessible medium. Moreover, without the intervention of technology, the privilege of "being global" on the web will continue to be limited to the largest companies with the deepest pockets.

The EU is a culturally diverse and linguistically fragmented market, both on- and offline. According to Eurobarometer surveys, the cross-border potential of EU retail e-commerce is not being realised; 51% of EU27 retailers sell via the internet, but only 10% support cross-border transactions. While 30% of EU citizens have purchased online, only 7% have purchased from a retailer in another Member State. Businesses most likely to be involved in cross-border retailing are medium and medium-large retail Enterprises, with a limited number of outlets in other Member States and with existing language capabilities. Over half of Europeans

who have come across advertisements from other EU countries have also made a cross-border purchase. Alas, a majority of Europeans (55%) have never come across advertisements or offers from sellers/providers located in other EU countries. And two-thirds of EU consumers would not buy a product or service online unless it is offered in his or her native language.

Every aspect of the new ecosystem, and all the products and services built on it, must be delivered in the languages of their actual users, and this is the driving force behind the market for translation software and services, as well the multilingual features of both speech and content applications.

Where language is the very stuff of our digital system – customer interactions, employee conversations, technical and scientific knowledge, cultural and social objects of all kinds – the era of the Lingua Franca is over. Interacting across the many languages of the digital world is no longer optional.

There is already significant global spending for LT software & services, but while the identifiable market for LT is far from trivial, its impact extends much further, enabling the next generation of our digital systems. Every sector will rely on LT at some level to remain competitive – to manage the data tsunami, to enable cross-border trade, and to engage with employees, customers, patients and citizens flexibly and responsibly through their multiple digital devices.

2.2 The LT-Innovate Market Model

To measure the scale of the LT market involves modelling its components, and hypothesising about size, segmentation and growth rates. At present no analysts follow the LT market as a whole, though many track its components and sub-components in different ways; larger pure-play LT companies are tracked, as are the LT-related developments and products in the larger software companies. As will become evident in the discussion of trends, the borders between the technological segments (speech, translation, content) are fuzzy at best, and much of the real innovation in LT is happening at the edges, where different types of intelligent services are combined in Unified LT applications (speech and translation, intelligent content and translation, etc.).

An initial sizing of the industry is, therefore, reliant on a number of different sources whose assumptions and taxonomies inevitably vary. Is speech analytics measured in the speech, or the content analytics space? Is multilingual content management included in the content or the translation space? What about speech search? Where would you classify an intelligent, multilingual interactive virtual assistant? It is evident that the convergence of the different historically separate market segments for LT will require new market categories, but as yet these do not exist in the minds of the analyst community.

Nevertheless, an order-of-magnitude sizing of the LT industry is essential to understanding its contribution to the vitality of the ICT market. This initial sizing is based on a model that draws on a number of different sources, with input from IDC on the companies and technologies they track in their Worldwide Software Market Forecaster service. IDC's service includes reliable tracking of Enterprise software markets and companies. To measure the size of the market for Intelligent Content (which is largely, though not exclusively, delivered via the Enterprise channel) the IDC analysts identified the segments and companies within their software market taxonomy that are active in the different sub-segments of Intelligent Content (Scanning/Text Input, Content Authoring/Creation, Search & Navigation, Text Mining/Analytics, Rich Media Search/Analytics) and extracted the revenues that they consider are directly attributable to Intelligent Content Technology sales. The forecast for Intelligent Content is therefore IDC's direct contribution to the model.

IDC's forecasting model is robust for predicting the general rates of growth in software markets as a whole; their growth and exchange rates were used as a baseline for the other two segments (speech and translation), and then adjusted based on the findings from LT-Innovate research. IDC contributed additional intelligence on the companies they track who are active in these markets (i.e. large players such as Nuance,

Microsoft, IBM, Google as well as smaller market leaders such as SDL and Lionbridge), which was supplemented with data from other specialist analysts. In speech these included Opus Research (voice biometrics), Voice Information Associations (Automatic Speech Recognition –ASR-/Text To Speech –TTS-), comprehensive studies from generalist research firms (Global Industry Analysts, TechNavio), as well as the thorough coverage of the speech industry provided by SpeechTech magazine.

Only one analyst firm (Common Sense Advisory) covers the translation industry in a systematic way; while they provide good on-going intelligence about developments in the Translation Technology space, they do not publish forecasts of that market separately from translation services (the universe of so-called Language Service Providers, or LSPs, including interpreting services). The model therefore starts from CSA's service industry forecasts and adjusts based on criteria used by CSA and input from IDC and other analysts who follow translation software at more distance.

This raises the question of what constitutes "services" in the Translation Technology (TT) segment. For the LT-Innovate model, the TT services market includes all companies that provide translation in combination with technology-based development and management of linguistic resources and processes on behalf of the client. A freelancer or small agency that uses translation memory for personal productivity, but does not share and manage those resources with and for clients, is a Translation Technology user, but not a provider of TT services. An agency that, on the other hand, uses Translation Technology to develop terminology and translation management resources, and maintain and manage them, on behalf of clients, and charges for that as part of its bundle of services, is a company in the Translation Technology services space. For the largest LSPs, these types of services, increasingly supplemented with the direct use of machine translation within the process, constitute most of their business, and all of their business with large customers.

Readers should note that current business models in the translation services segment are quite different from those in speech and content, where services are directly associated with adapting, enabling, and deploying technology for the client's own use in its business processes. This is sometimes the case with large translation buyers, but even then the service provider is much more deeply embedded in the client's processes (e.g. for multilingual technical publishing) because of reliance on language data as the resource that drives the solution end-to-end. However, with the advent of Enterprise cloud services the operating models of the different LT segments may look more similar over time, as vendors enable and support processes based on language resources across the spectrum of applications that use LT.

CSA's market estimates are controversial within the translation industry, and while they are based on hard survey data from a respectable sample (around 4%) of the 26,000 companies they estimate are in the industry worldwide, the segmentation between large companies (who they say collectively earn over $5B/year, getting on for 20% of their market estimate) and the "long tail" of thousands more companies, implies average revenues for the small companies that are not supported by evidence. We therefore used the detailed study carried out by The Language Technology Centre on behalf of DGTranslation (at the European Commission), The Size of the European Language Industry, to adjust CSA's estimates. The LTC study is based on a painstaking, country-by-country review of the industry in Europe (widely believed to command half the world market) using public-record data and information from local translation associations, supplemented with a survey of 1100 companies (most operating in Europe). That survey provided data about the proportion of companies developing and using Translation Technology, which we used to further refine our estimate of the proportion of the total market that provides TT services. At the time the survey was carried out in 2008, 10% of respondents were developing Translation Technology, 10% were using machine translation, half were using terminology development and management tools, and over 70% were using CAT (Computer Aided Translation) technology.

Global Market

Based on the LT-Innovate market model the estimate of the size of the 2011 worldwide Language Technology market, including software and services, is €19.3B. Of that, around a quarter of the market is for software sales (€4.5B); the balance is for services (€14.7B).

Worldwide Language Technology Software & Services Market 2011 & 2015 (€B)

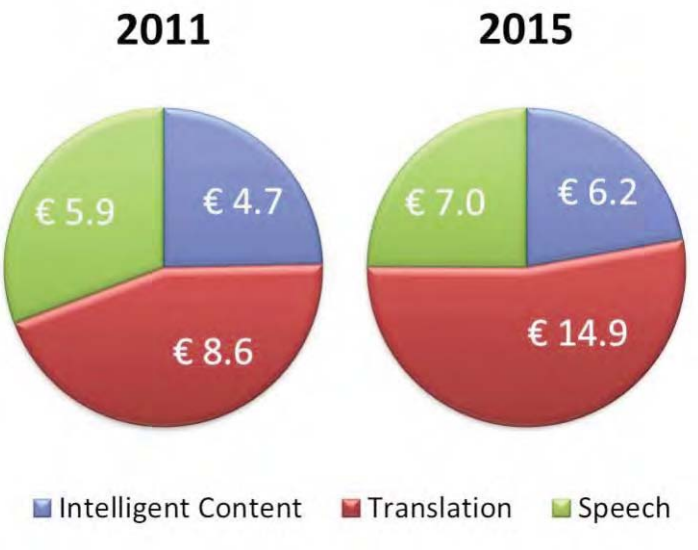


Figure 5: Worldwide Language Technology Software & Services Market

By 2015, the market should grow to nearly €30B. While the software revenue proportion remains roughly the same, the proportion spent on translation software increases significantly, and the proportion spent on content software declines.

America plays in important role in the development of LT industry. In 2011, the software revenue in America accounted 48,4% of total software revenue and Europe 25.8%. Americas' LT software market was 1.9 times bigger than European market in 2011. By 2015, the situation will remain similar. Americas' LT software market is foreseen to be 1.8 times bigger than European market and the markets will weight 40.3% and 25.2% respectively. Thus America is expected to capitalize on software revenues and continue dominating the LT software market.

Opposite to the software, the LT services market is expected to be capitalized in Europe and situation tends to benefit this continent. In 2011, the LT services market revenue accounted for 41,9% of total software revenue in America, while 40.6% in Europe. Thus, in 2011 the Americas' LT services market was 1.03 times bigger than European market. Nevertheless, by 2015, the situation will be different. Americas' LT services market will be 0.97 times bigger than European market, as European market will grow at higher pace. The markets will weight 40.3% and 41.7% respectively. The demand for LT services is expected to growth in Europe.

Currently, the European market is competitive with respect to American suppliers, with a stronger academic base, however business adoption is faster in America, lowering differences in market in a 4-5 years' time.

2.3 Trends & Growth by Segment

Speech Technology Market

Trends – Speech Technology

Speech Technology is being deployed on new platforms opening new channels and markets, as well as in analytics applications that bring the technology into new areas of the Enterprise. Speech search, audio mining, speech analytics and applications enhanced by emotion detection and "NLU" (natural language understanding, a term being used generally to describe dialogue) add spice to offerings being developed.

The market is heavily dominated by speech recognition, with a long history of commercial use. Positioned as cost-saving technology, Automatic Speech Recognition(ASR) sales were little affected by the recession; claims of 50% - 90% cost reduction in call centres are reported, and call centre penetration is extremely high. Speech transcription services (e.g. in the healthcare domain) are increasingly offered through cloud services.

Improvements in the quality of Text to Speech (TTS), combined with platforms requiring interactivity (such as mobile, gaming) are driving new opportunities for speaking applications. Notable features are naturalistic voices in many more languages, used in education and gaming environments, as well as interactive access to the web (Voice Portals).

Voice biometrics (the smallest segment of the market) is quietly gaining both market interest and investments. The performance of speaker verification products, and the ubiquity of digital access, is also boosting take-up in biometrics; the Nuance VoiceVault has 50 million voiceprints, and growing; a small bevy of niche players have emerged in this space, including the Spanish company Agnitio, and there is growing interest from the military/security sector.

Operator revenue from voice (fixed line) is in decline, and the explosion of VOIP further drives opportunity to embed speech applications. On-premise IVR solutions are also fading, cloud Interactive Voice Response (IVR) is on the rise creating scope for collecting and deploying vast stores of speech data. This, in turn, will drive quality and market acceptance. Open standards support interoperability (notably more advanced than in other LT segments), but certification regimes from major platform vendors suppress opportunities for new entrants.

Major markets for Speech:

- Call Centre is a core global market
- Medical reporting and transcription is growing (especially in the USA for compliance with new Electronic Health Records regulations)
- Large and stable government customer base (including specialised defence applications)
- Speedy growth in consumer markets on devices and social platforms.

Major gaps and opportunities for speech technology:

- Language coverage, need to develop a competitive market for languages beyond English
- Asian markets, high digital growth and low penetration outside Japan
- Speech data for developers, competitive choices on open platforms

Growth – Speech Technology

The estimate of the size of the 2011 speech technology market, including software and services, is €5.9B, of which around one-third is revenue for software.[3] The five-year CAGR for speech is 9.7%; while take-up is expanding dramatically, the market value of new products is lower than in the traditional Enterprise market.

Worldwide Speech Technology Software & Services Market 2011-2015 (€B)

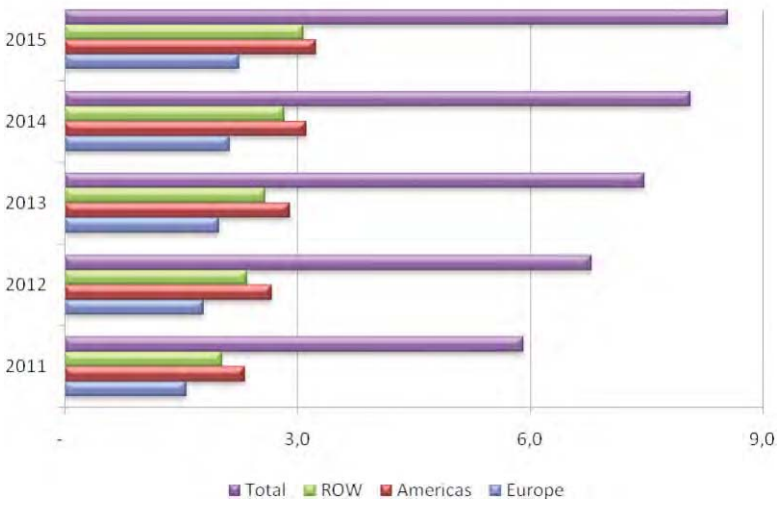


Figure 6: Worldwide Speech Technology Software & Services Market

By 2015, the value of the market is forecast to grow to €8.6B, of which services continues to be the lion's share. The fastest growth will be in "ROW" at 11%.

Translation Technology Market

Trends – Translation Technology

The rise and rise of statistical machine translation is having a transformative effect on this market. Although very good rules-based systems have been available – and used in niche market areas – for decades, no available product ever came near the ability to cover the hundreds of language pairs (and thousands of potential pairs in the future) that an engine like Google Translate can handle. Both Google and its direct competitor the Microsoft Bing Translator are available for anyone to use on the Web, and can also be licensed as an embedded service in third-party platforms and products using APIs.

Somewhat to the surprise of the translation industry, the clear line between professional translation (by a trained linguist) and what used to be called "gisting" (a machine translation just good enough to get the gist – for amateurs) gets more blurred every day, as both professionals and amateurs pile onto these translation platforms and use them. It is no secret (indeed, is widely discussed on translator forums) that trained and

[3] See description of methodology in the Global Market section.

qualified professionals have started to use online engines to create first-draft translations, which they subsequently correct. They can make the corrections in the Google interface, and choose to share that translation back to the platform.

What's significant about the vast scale of use, particularly by professionals, is that the quality of translations on these systems is directly proportional (at least in theory) to the quality of the existing translations they're based on. Anyone can translate two million characters a month at no cost on Bing (that's hundreds of text pages), and after that pay only $10 per million characters. The paid Google service costs $20 per million characters. While the services are open, and easily linked, the translation data derived from these services remains closed and proprietary.

Translation is by far the most service-intensive segment in LT; in the professional market, human translation still rules, though a quiet but pronounced shift to machine translation with post-editing (analogous to what freelance translators are doing with Google and Bing) is unquestionably underway in the agencies, some of whom are also actively developing their own engines using open-source software. The cost of human translation without automation is simply unsustainable for most applications.

Prior to the statistical revolution, machine translation was used minimally and the dominant technology in the market was translator aids, which still have a substantial share of the market. The legacy market in desktop translators, the only option for consumers prior to web-based engines, is also shifting online, though a number of packaged products are still sold. Next we will see the maturing generation of hybrid engines that combine different techniques, appropriate to different translation requirements.

Automatic translation is being embedded in Enterprise applications and services, e.g. in chat rooms for customer service. The other major trend is the emergence of consumer translator tools (including spoken translation) on mobile devices or in online applications. These systems come and go; most of them use the Google Translate API, though when Google started charging for that service, the field was cleared somewhat. Other products use their own technology, including the Jibbigo app developed at CMU and Karlsruhe, which is distinctive in delivering automatic spoken translation entirely from the device, without relying on an internet connection.

Developments in China are notable, where nearly 75% of internet users say they use online translation tools (that's nearly half a billion people) and they too are increasingly accessing translation on mobile devices. Translation is the fourth most common application online in China, after shopping, searching, and social networking. In the last year translation requests to the Chinese search engine Youdao's mobile app doubled, while web-based requests grew by 30%.

Major markets for translation technology:

- Language Service Providers
- Large, global Enterprises with substantial internal publishing requirements (particularly technical publishing)
- Embedded in Enterprise systems
- Embedded in social software
- Consumer devices

Major gaps and opportunities for translation technology:

- Coverage for languages that are either poorly handled by current technology, or for which digital resources are poor
- Custom, adapted engines for specific domains (e.g. verticals, comparable to work at the European Publications Office)
- Translation data on open platforms (data services), to continue to reduce cost and handle growing vol-

umes in professional environments

- Open translation platforms (translation services) for the not-for-profit/NGO sector.

Growth – Translation Technology

The estimate of the size of the 2011 translation technology market, including software and services, is €8.6B, the vast majority spent on technology-based services; direct software revenue is only 7% of the market.[4] The five-year CAGR for translation is 14.6%. Translation is the LT application least susceptible to full-scale com-

Worldwide Translation Technology Software & Services Market 2011-2015 (€B)

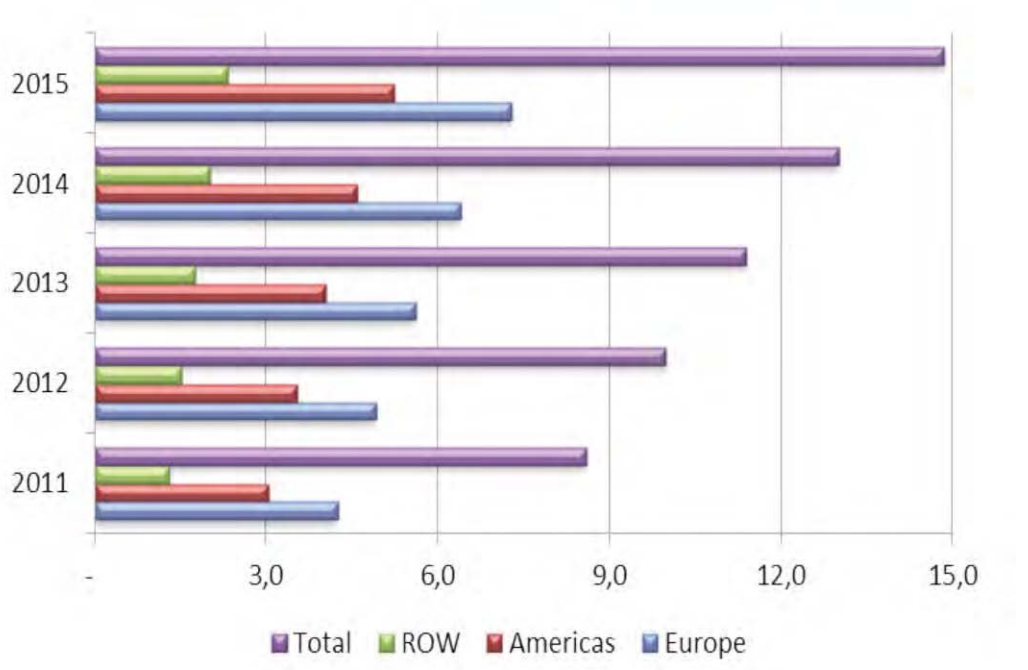


Figure 7: Worldwide Translation Technology Software & Services Market

moditisation, at least for the foreseeable future

By 2015, the value of the translation technology market is forecast to grow to €14.9B; while services continue to constitute most of the spending, the share attributable to software grows to 12%. Growth rates in the software share of the market compound as the new translation platforms mature. Services continue to grow, but at a slower pace, averaging 13%/year. The expansion of the size of the market is driven primarily by technology.

[4] See description of methodology in the Global Market section.

Intelligent Content Technology Market

Trends – Intelligent Content Technology

Intelligent applications for search, analytics, and content are generally built on (or within) existing mainstream software environments and solutions. Trends in the market for this technology, therefore, track closely with Enterprise software markets. Enterprise buyers in the search and analytics space are line of business managers, who demand compelling Unified Information (UI), task and process–based applications.

While not strictly speaking part of the LT segment of the market, new data storage platforms (NoSQL, Hadoop, SAP HANA) condition the take-up of big data and large-scale analytics applications. Business Intelligence, social analytics, decision support, business-process outsourcing, and predictive modelling dominate that landscape. Key Enterprise functions where these applications are being deployed are Marketing, Finance, and R&D. Customers are looking for faster (and therefore more actionable) insight, to handle a diverse range of data and content resources, to solve core business problems.

New search-based applications supporting a specific task or workflow (e-discovery, fraud detection, voice of the customer, sales prospecting, research, customer support) integrate domain knowledge to support the particular task, including industry taxonomies and vocabularies. Search is embedded in the process; it's search without a query box. Expertise discovery applications, such as that from Sinequa, support skills management in large Enterprises, while Textkernel uses web mining and semantics in the recruitment process. Cloud-based Enterprise search, where Exalead is a leader, has caught also the attention of Amazon and Microsoft. Open-source and mobile search are both on the agenda. Google's Knowledge Graph has made semantics mainstream, while Bing added social search to its platform. Adding OCR as a background service for search unlocks documents stored in image formats.

On the content side, we see intelligent automatic authoring of regulated documents (safety data sheets, product leaflets) in industries such as chemicals and pharmaceuticals. Checking and authoring technology from Acrolinx (recently named by eContent magazine as one of the top 100 companies in the digital content industry) deliver content optimisation, and can be applied to the creation of technical content using standards such as STE and DITA, and to "SEO at source" by guiding the use of keywords during the writing process. Intelligent creation of content overlaps with translation, as both promote the use and management of standard terminology, and use linguistic analysis of text to achieve clearer and more translatable content.

While Intelligent Content (IC) technology is mostly being deployed at Enterprise level, imaging technology has a new lease of life, as scanning and OCR applications migrate to mobile devices; ABBYY's Mobile OCR Engine SDK enables embedded scanning apps on all the major mobile platforms. Uses range from business card scanning to the address book and reading barcodes, to capturing text for translation applications. We also see the first mainstream mobile application of intelligent search, combined with Speech Technology, in Apple's Siri.

Although with its horizontal/Enterprise focus, IC Technology is used in all industries, markets leading the take-up of advanced search and analytics include:

- Banking and Financial Services
- Communications, Media and Services
- Government
- Manufacturing
- Natural Resources

Major gaps and opportunities for Intelligent Content Technology:

- Large-scale development of industry/domain/vertical specific solutions in search and analytics

- Intelligent mobile search using location and language specific data
- Cloud-based multilingual search platforms
- Multilingual authoring platforms for bloggers

Growth – Intelligent Content Technology

The estimate of the size of the 2011 worldwide IC Technology market, including software and services, is €4.8B, about 40% of which is for software.[5] The five-year CAGR for this segment is 7%, with growth of soft-

Worldwide Intelligent Content Market 2011-2015 (€B)

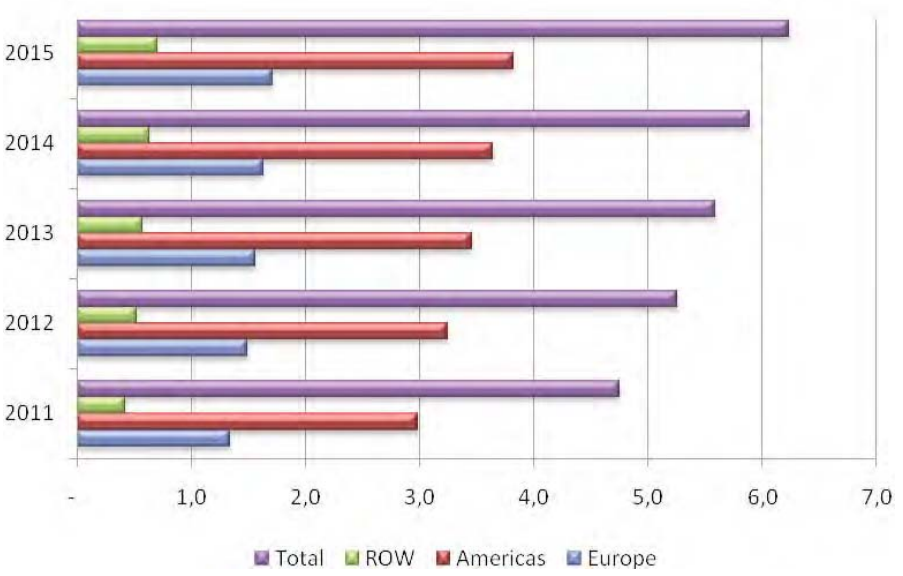


Figure 8: Worldwide Intelligent Content Technology Software& Services Market

ware slightly faster than for services.

By 2015, the value of Intelligent Content sales is forecast to grow to €6.2B. Growth rates in "ROW" markets should be significantly higher than in Europe and the Americas, where the market is more mature.

2.4 The European Market

The forecasts in our model predict that the Translation segment will continue to dominate the European LT market, and will grow to be a larger overall share (65%) by 2015. Intelligent Content remains the smallest segment in Europe, and speech is only slightly larger. The assumptions of the model are based on recent trends in the respective segments, notably the dilution of the European industry in both speech and content through acquisition by off-shore companies. By contrast, consolidation in the translation industry has historically been Euro-centric; acquiring European translation company signals, by definition, a desire to continue

[5] See description of methodology in the Global Market section. The forecast for the Intelligent Content segment was prepared by IDC.

European Language Technology Market 2011-2015 (€B)

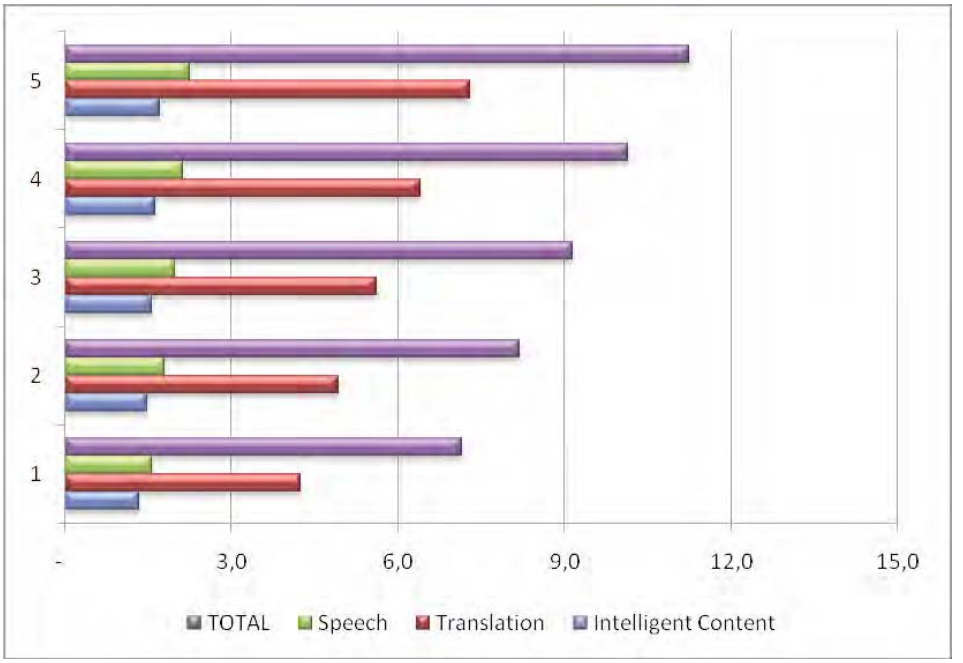


Figure 9: European Language Technology Market

to operate in the local market and develop local linguistic talent and resources. Moreover, translation technology development has historically been a European strength.

Europe's share of the worldwide market will increase slightly to 38% over the five year period, compared to 42% in the Americas. However due to the imbalance between LT segments, that share is significantly lower (24% in 2015) for the software portion of the market; as we have noted, sales of technology-supported human translation services far outweigh sales of translation software, and will continue to do so during the forecast period. The strength of the European market for translation reflects both the depth and excellence of the industry in Europe, and the need for translation into the many languages of Europe on a large scale. Large-scale multilinguality, in turn, is an inhibitor for growth in both the speech and content markets, where products and applications must be deployable in local languages.

Factors that could change the assumptions behind the market model:

- Faster and more extensive deployment of content applications in more European languages, in a coherent framework for all languages
- Development – and integration – of speech components (for recognition, generation, and identification/verification) in more European languages, affordably available for European app and solution developers
- Large-scale deployment of open source machine translation in open environments using shared resources
- Large-scale sharing of resources (paid and free) throughout the European industry
- Development of vertical and industry-specific platforms for LT development and deployment, engaging whole industries in cooperative initiatives (analogous to SWIFT in banking)

Challenges in the European market

Technological Barriers

LT is a highly complex technological domain that represents the intersection of several disciplines, including the many sub-domains of linguistics, mathematics, and information science. LT functionality remains (and may always remain) a work in progress, with few genuine technological breakthroughs. The most significant, for current technology, was the combination of NLP and computational linguistics with statistical modelling, which began more than thirty years ago, and is now a feature of many LT implementations including both speech and text.

Most improvements using today's technologies are incremental and rely particularly on the ability to access and maintain ever larger and more finely tuned linguistic data. Lack of access to that data will constrain the technological development of LT. Acquiring and using it may rely on cooperation between the LT industry and the different constituencies that own, need and use it. Collaboration between the industry and data owners will be needed. Also regulation of the use of such data should be made much more open, and core data (such as terms, concepts, ontologies) should be standardised and shared in an open environment

Copyright Issues

This news item was published on 15 November 2012 by the Dutch newspaper de Volkskrant:

> *The French president François Hollande demanded that the Dutch return all the borrowed words from the French language as of the year 2015. The Dutch people will then no longer be allowed to use many words – such as 'dossier', 'portefeuille', 'ordinair' – that have become commonly used words in the Dutch language.*

It was, of course, published as a parody - who would seriously think that anyone can claim ownership over the individual words that people use. But the intellectual property rights regarding terminology and translations imply that publishers, authors and translators can own words. Moreover, people in the language industry encounter daily conflicts between the principles of intellectual property law, corporate policies, business practices, and the pragmatic use of data, particularly (but not exclusively) in translation systems.

The principles of today's Intellectual Property (IP) legislation all stem from a last-century definition of translation whereby translation memories (stored pairs of previously translated sentences) were merely intended as a personal aid to translators and were protected by the same rules that apply to complete texts. Now focus has shifted from translation memories to massive amounts of translation data in the cloud, in the form of parallel text corpora (see the discussion of Translation Technology trends above). These translation data may be accumulated from translation memories, or from online translation service platforms or harvested ('crawled' and aligned) from localized versions of web sites and other sources. In addition, these data are often usefully annotated with attributes for domain, content type and source. In this way translation data are becoming the key to quantum leaps forward in translation efficiency. (Google has demonstrated this in the past five years by training new machine translation engines for 100,000 different language pairs using nothing but translation data.) But who owns data collected and created in this way?

In North America this use of accumulated translation data may be allowed under the exception of *fair use* and *fair dealing*; Europe tends to interpret copyright law on a much stricter basis. Yet the translation innovators are ubiquitous, and they range from very large global IT companies in the US to small start-ups in any part of the world. This creates at least the impression of unfair competition both inside and outside the translation industry.

Clarification of the copyright laws regarding such key language data, and establishing open principles for their use, would remove ambiguity and encourage more innovative use of language data, especially in Europe. Principles might include:

- A clear distinction between the way IP rights are treated for the text to be translated (the Source), the translation (the Target) and Translation Data as a new legal entity.
- Translation Data are defined as a database containing terms, phrases and segments of text, aligned between two or more languages. Translation Data in most cases contain phrases and segments from many Sources and Targets. If the database allows users to reconstruct the Source or Target, as referred to in the first principle here above, this will be considered an infringement on the IP rights assigned to the Source and Target.
- IP rights to the Source and Target may be held exclusively by the author, the translator or the company that is publishing the Source and Target.
- Translators and translation companies should be allowed to store, share and aggregate Translation Data for the purposes of developing derivative work, leveraging and reusing translations, research, and improving their services.
- The translator or the company that aggregates the Translation Data holds IP rights to the Translation Data in the form in which the data are stored and used in the database.
- Owners of Source and Target should know that they can legally protect their documents from copying when they publish on the web.

Competition

The LT market is now extremely competitive, with many small players and a few very large and very powerful participants. There are threats to small companies in all areas of LT (though not necessarily to the industry itself):

- in semantics and analytics, where the global software companies are all engaged, and most are developing cloud services that give them more reach
- in translation, where the divide between professional services (the traditional industry) and casual translation (typified by Google and Bing online translators) is slowly eroding, and competition from new entrants (in the USA and in Asia) has accelerated
- in speech, where a small number of technology suppliers service most of the market in the US and Europe, inhibiting access for small companies, or new entrants.

A healthy industry full of competitive innovators is desirable and will grow the market, particularly in Europe where attention to under-served European languages will flourish in the hands of “local” companies. However, the scale now required to compete indicates a strong need for collaborative approaches to the market. Small innovators may have more opportunities in leading-edge unified LT markets, where combining forces may give competitive advantage.

Investment Climate

In spite of persistent complaints that venture capital is scarce in Europe, this has not been observed as a particular problem for LT companies, many of which have found sources of funding. However, it is not clear whether growth paths for small European LT companies have really been tested against the realities of the climate, i.e. whether ambitions to scale have been thwarted or simply not tried. Taking advantage of mentoring and other support features of the investment community is advisable and may not be fully exploited, which may in turn hold some companies back from their true potential.

Talent Scarcity

A study by EMC reports that 65% of data science professionals believe demand for data science talent will outpace the supply over the next 5 years – with most feeling that this supply will be most effectively sourced from new graduates. But there are few degree programmes or even training regimes for data scientists (per-

haps no clarity about what a data scientist is). While linguists, including translators, may not be in short supply, there will be increasing need for linguists with technical LT skills. For both mathematics and linguistics, the shortage is, or soon will be, of technical skills that can be deployed in commercial (rather than academic) environments. Technology training in translation degrees is focused on last-century technology such as Translation Memory. The industry should consider developing a certification protocol that covers the range of technical skills needed across the range of technologies.

Valuable LT business proposition

As already proven several times, having the best product and the right technology is not enough; the proper marketing and sales strategy need to be defined and implemented. Cases such as Betamax vs. VHS demonstrated that a deficient technology could become the consumer leader. In the case of language technology, there is little awareness of how LT can be used in organisations to assist in information management and decision-making. Although lately some successful cases have already starting to appear in the market, many IT managers are still relatively unaware of the benefits that LT can provides them.

If there is no a clear business case but only technology advantages, companies from the demand side will not buy in. In this way, a classic problem in IT remains: IT suppliers need to speak the language of business. Moreover, suppliers also have difficulty reaching the right people at customer organizations, and difficulty finding a common language for communication with business partners.

Suppliers, LT vendors and IT integrators should work closer and harder to identify killer business cases, increase market awareness and deploy market strategies understanding how economic return affects clients, developing modular/incremental products, and forging cross-industry alliances for to strengthen market channels.

3. Innovation in the LT Industry

3.1 An Overview of the European Industry

LT-Innovate estimates there are around 500 companies in Europe either actively developing Language Technology, or embedding its features in their products and services in an innovative way. This excludes most of the Translation Technology service companies described in the market model, which are a separate case. It does include translation service companies that also develop and *sell* technology separate from translation services, a practice that used to be common in the industry but has receded as professional software companies that focus exclusively on technology have become more common.

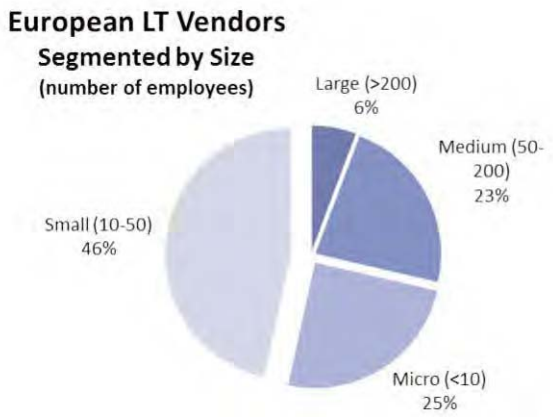


Figure 10: European LT vendors (by size)

So far we have reviewed around 400 companies, and certain features of the market are notable. The industry comprises mostly small companies, concentrated in the western and northern regions of the EU, with a mix of long-established players but also a significant number of new entrants. A quarter of companies are micro-enterprises with fewer than 10 employees, while only 6% have more than 200 employees; almost the entire industry are SMEs.

The European LT industry is a curious mix of start-ups and companies that have been active in the field for a very long time, some for 20 years or more. As a maturing technology in a benign market context, this is a healthy cocktail of technological experience and entrepreneurial enthusiasm, and no doubt reflects the fact that LT is both a very "old" technology that has been in the market in various guises for decades, but at the same time "new" as improvements in functionality, new approaches, and in particular new paths to market and business models have given LT a boost.

Over half the industry comprises companies active for more than 10 years, many that remain small. The fact that so many companies fail to scale, even after years in business, is unusual in a technology industry, and indicative of the market context for LT in Europe: local/national companies with expertise in local languages serve local markets with services based on their own languages. This state of affairs is not likely to be sustainable, as cloud-based language-enabled services are launched on a large scale. At present, few European companies are in a position to compete in an ecosystem where access to technology, rather than narrow linguistic expertise, is the driving factor.

Figure 11: European LT vendors (by length of time active on the market)

Of course many European LT companies have gained access to markets by being acquired by larger companies. This has been a particularly strong trend in language services, where huge companies that can address global clients have been built through the consolidation of many smaller, independent agencies; those same companies snapped up a range of translation tools in the process, and created the technology-based translation industry we know today, an innovation dynamic that was driven from the European base of the industry. That phenomenon owed a lot to Irish government policy to encourage software "manufacturing" (in the days of packed software that came in boxes) in Ireland, creating an entirely new industry called localisation to serve the multilingual needs of the European software market. Contrast that with the situation today, when the most visible innovations in translation technology – and business models for translation – are largely driven from outside Europe.

Some small and medium-sized European technology companies have been picked off by large US players: the Dutch semantic search company Q-go is now part of Oracle; UK intelligence analytics firm i2 is now part of IBM; Loquendo, the speech company spawned by Telecom Italia, is the most recent European acquisition by Nuance; Spain's NeoMetrics Analytics is now part of Accenture; the speech search engine of UK-based Aurix is now part of leading contact centre supplier Avaya and another call centre company Syntellect acquired Fluency Voice Technologies. And of course Europe's most successful "intelligent search" company, Autonomy, is now part of HP.

Similar acquisitions of small US LT companies by US software giants are equally common. IBM alone has invested $14B in analytics acquisitions in the last five years, not to mention its internal development of Watson (and predicts it will achieve $16B in analytics sales by 2015). Microsoft bought FAST, Oracle bought Endeca. Nuance has acquired literally dozens of companies to build its market-leading portfolio of speech and natural-language offerings.

Not all acquisitions are going to the US: Dassault Systèmes acquired the French intelligent search company Exalead; Experian, the global information services group based in Dublin, bought the UK speech-verification company 192business; OnMobile (spinoff of the Indian IT services giant Infosys) acquired the French speech company Telisma, and Wolters Kluwer acquired the US special-domain semantics company Health Language Inc. SDL in the UK acquired Language Weaver, the main statistical MT engine in the market developed for Enterprise (or government) markets, rather than as an online platform.

This consolidation in the industry is healthy, and it moves LT up the food chain into mainstream applications and markets. It does not, however, promote the evolution of a strong and self-sustaining LT industry across Europe, as evidenced by the patchy language coverage of solutions in the speech and content markets, a key constraining factor in Europe's share of those segments of the market.

Overall, companies are concentrated in the north and west of the EU, with around the same distribution among younger companies.[6]

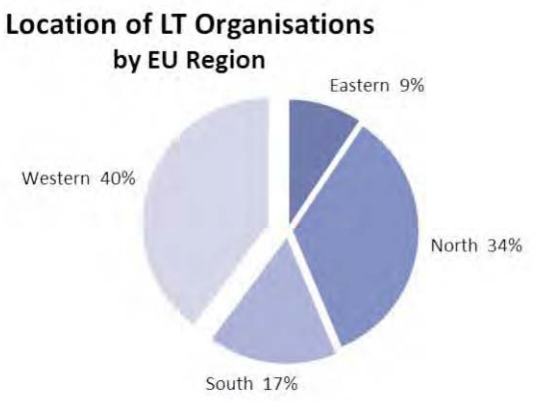


Figure 12: European LT vendors (by geographic area)

New entrants (companies less than five years old) predictably include a number of apps and products for mobile devices (smart mail, translators, voice and music search), as well as new companies in the speech segment with customised recognition applications and a batch of new speech output platforms, novel uses of LT in language learning, intelligent robotics, virtual assistants.

Speech

Global players:

- Microsoft – speech embedded in its software platforms
- Google – speech-enabled search
- Nuance – Enterprise and packaged solutions provider in the US and Europe
- iFlytek – 5,000 partners, owns 70% of the Chinese-language speech recognition market

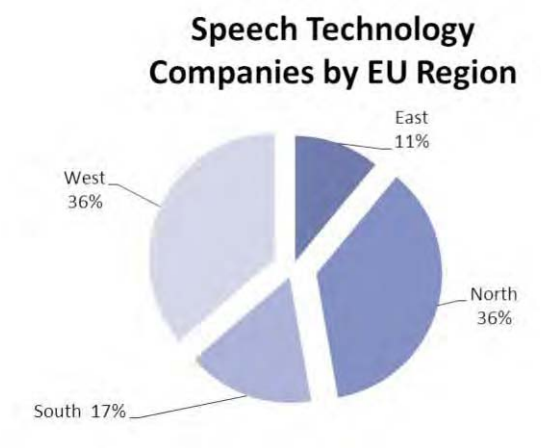


Figure 13: European Speech vendors

[6] *NORTH: Denmark, Estonia, Finland, Ireland, Latvia, Lithuania, Sweden, UK. SOUTH: Greece, Italy, Malta, Portugal, Slovenia, Spain. EAST: Bulgaria, Cyprus, Czech Rep., Hungary, Poland, Romania, Slovakia. WEST: Austria, Belgium, France, Germany, Luxembourg, Netherlands*

European Companies active in Speech (developing or embedding in innovative ways)

North			
Aculab	Connexor Oy	Lingsoft Inc	Toby Churchill
Allvoice Developments Ltd	Creative Virtual	Netcall	ValidSoft
Artificial Solutions	Eckoh	NICE Systems	Vicorp
BigHand	Fizzback/NICE Systems	Novauris Technologies Ltd	Vocalytics
Bitlips Oy	HulloMail	Phonetic Arts	Voice Technologies
Business Computer Projects	Lausumo Speech Technologies Ltd	Speech Graphics	VoxGen
Business Systems	leading software	Synthetix	
Call Trunk Ltd	Lingapps	Tilde	Verbio Technologies
CereProc Ltd	LingleOnline Ltd	TMP Media Group	VoiceInteraction

South			
Alpineon d.o.o.	Fonetic	Natural Vox	Ydilo AVS
Amebis	H-care	Pervoice	
Anboto Europe	Indisys	Synthema	
Cilenis	Lobisoft	The Corpora Robotics Company	

East			
AITIA	Phonexia s.r.o.	Speech Technology Center	
Applied Logic Laboratory	Sakrament	Speech Technology Ltd	
IVONA	Sestek	Stanusch Technologies	

West			
Acapela Group	Koemei	Semantic Edge	Vocapia
Aldebaran Robotics	Linguatec	Soundcloud	Vocapia Research
ASC telecom AG	LINGUISTIC FACTORY	Spectralmind	Voice Business GmbH
ATIP	Natlanco	SpeechConcept GmbH & Co. KG	voice INTER connect
EML European Media Laboratory GmbH	Natural Touch	Sympalog Voice Solutions	Voice-Insight
English attack	Parrot SA	Telisma	Voiceinterconnect
European Media Laboratory	ReadSpeaker	Vecsys	Yocoy Technologies GmbH
Jibbigo	SAIL LABS Technology AG	VerbaVoice GmbH	Zuuka

Translation Technology (TT)

Leading technology suppliers in the global market:

- Google – free online translation and API for developers
- Microsoft – free online translation and API for developers
- Youdau – free translation in Chinese search engine

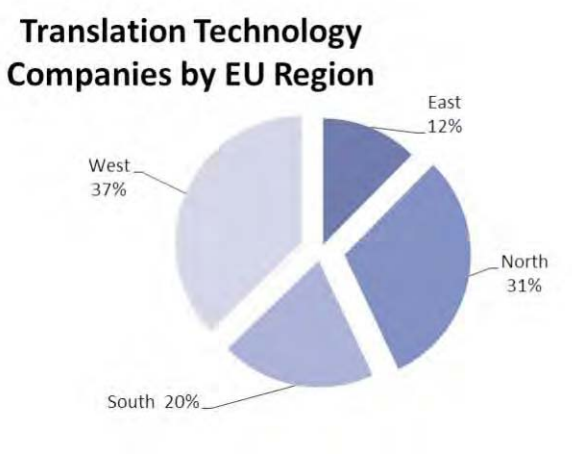


Figure 14: European Speech vendors

European Companies active in Translation (developing or embedding in innovative ways)

North			
Alchemy Software Development	Language Technology Centre (LTC)	Semantix	Translution
Applied Language Solutions	Languagelens ApS	Sunda Systems OY	Transmachina
Confirmit	Learnwell Oy – The Language Menu	SysMedia	Universally Speaking
Convertus	LINGO24	Technabling Ltd	VistaTEC
ESTeam	Lingosaur (Contatum Ltd)	Tethras	WebCertain
Existor	Lingsoft Inc	The Selfservice Company	WebWordSystem
FODINA	metatrad	The Virtual Zone	XTM International
Interverbum Tech	Multilizer (Rex Partners Oy)	Tilde	
Kielikone Oy	Rubric	Transfluent- Xiha Ltd (PremiumFan Page)	
K-Now	SDL Language Technologies	TranslateMedia	

South			
Amebis	Braser Soft	MTC Soft	Tek Translation International S.A.
ApSIC	Busuu Online S.L.	Pangeanic	Transiq Systems SL
ARANCHO DOC	Eleka	Prompsit Language Engineering	translated.net
ATRIL	IOLAR d.o.o.	RED INMIGRA	Universal projects and tools
AutomaticTrans	Linguaserve	Synthema	Vi-clone
Berca Translator	Logos Group	Tauyou (language technolog)	ZOP-CR d.o.o.

East			
Advanced International Translations, Ltd. (AIT)	Logrus Technology	One Hour Translation	Trident Software
AIT	Memsource	Skycode Ltd	XTRF
Interlecta OOD	Moravia	TiP Ltd.	Young digital planet Poland
Kilgray Translation Technologies	Morphologic Ltd	Translatica (PWN Ltd.)	

West			
ABBYY Europe Gmbh	Kaleidescope	MetaTexis Software and Services	Syn-Tactic
Across Systems GmbH	Lexcelera	MultiCorpora	Systran
Andrä AG	Lingenio GmbH	Myngle BV	Tedopres
CrossLanguage (CrossLang)	LingoKing	Plunet	Telelingua SA
Dicoland.com	Lingua et Machina	Promt	Textec Software
Digital publishing AG	LinguaSys	PROMT GmbH	Tolingo
DutchComm	Linguatec	Reverso - Softissimo	TopicZoom GmbH
Ecreation	Lingueo	Schaudin.com	Value-scope
Infovole	Lingupedia	STAR Group	Xplanation Language Services
Intertext Fremdsprachendienst e.G.	Lingvistica	Synapse Developpment	Yocoy Technologies GmbH
Jibbigo	Lucy Software and Services GmbH	Synergiums	zyLAB

Intelligent Content Technology (IC)

Leading technology suppliers in the global market:

- Endeca/Oracle
- FAST/Microsoft
- Google Search Appliance
- Lucene/SOLR
- Vivisimo/IBM
- Autonomy/HP

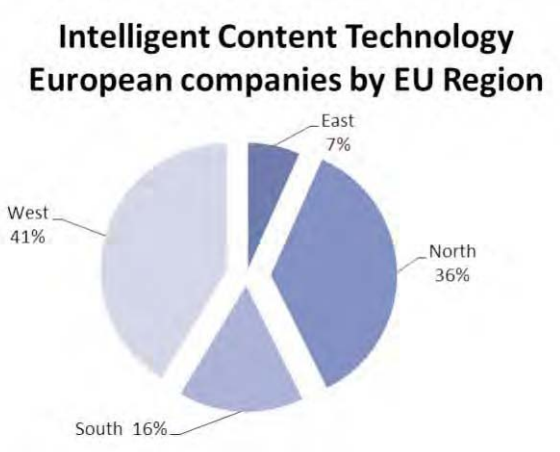


Figure 15: European Intelligent Content vendors

European Companies active in Intelligent Content (developing or embedding in innovative ways)

North			
192 business	Creative Virtual	Lexware Labs	Slicethepie
AB COMPUTING	CrowdControlHQ	Lingit	SmartLogic
Acuity Unlimited	Crystal Semantics Ltd	Lingsoft Inc	SoDash
Advance Finance Sys.	cXense	Linguamatics	Struq
Affect Labs	Digital Pebble	MetaBroadcast	Synthetix
Ankiro ApS	EDITD	NICE Systems	SysMedia
ARANCHNYS	Emediate	Nynodata AS	Talis
Artesian Solutions	Epistemics	Ontopia	Texthelp Systems Ltd
AudioBoo Ltd	Etronika	Orcawise	The Selfservice Company
Biowisdom	Etuma	Outbrain	The Virtual Zone
blinkx	Existor	PeerIndex	Tilde
Brandt Technologies	Featurespace	Profium	TMP Media Group
Brandwatch	gavagai	QIQQA	touchtype
Celebrus Technologies Limited	Greedy Intelligence	QuBit Digital Ltd	Treocht
Clinithink	GroupNos	REED ELSEVIER	True Knowledge
Cognesia Ltd	Handy Elephant	saplo	ValidSoft
Cognitive Maps	iSense/ad pepper	SDL Language Technologies	Visible Technologies
Communicative Machines	JayBee (Time Is Ltd)	Seevl	WebCertain
Concept Searching	Kognitio	SIINE	Webnodes AS
Connexor Oy	leiki	Sindice	Winterwell Associates

South			
ADmantX	Cognicor	Inbenta	priberam
Amebis	Crif/Cribis	Indisys	Strands Business Solutions
Avalon Biometrics S.L	Cybion srl	iSOCO	Synthema
Berca Translator	Cycorp Europe	lingibli	The Reuse Company
BITEXT	DAEDALUS, S.A.	Lingway Spain	Verbio Technologies
Cascaad	Experienceon Ventures	MTC Soft	Wagsoft
Celi Srl	Expert System	NOVOCAPTIS	Ximdex
Cilenis	H-care	OEG	Ximetrix
Circleme	Herta Security S.L	Okkam	Zemanta

East			
AITIA	iGlue	PetaMem	Tetracom Interactive Solutions Ltd
Applied Logic Laboratory	Kanunum.com (Karakullukcu Consulting)	Semantic Visions (Newstin)	WEBLIB
EffectiveSoft	Knowledge Hives	Software602, a.s.	Weblib LLC
Gamax	Ontotext	Stanusch Technologies	

West			
A2ia: Artificial Intelligence & Image Analysis	Eptica	Lingway	Semvox
ABBYY Europe Gmbh	ERDIL	METADAT GmbH	Senbarila GmbH
Acrolinx GmbH	Exalead	MIIA Holding Ltd.	SEOLYTICS
Actonomy	Fabasoft Mindbreeze	Mondeca	Sinequa
Aduna	FACT-Finder	Moresophy GmbH	Softlib
AI Applied	FADYART	Natlanco	Spectralmind
Aldebaran Robotics	Flimmit GmbH	NEOFONIE	Spoken Language System
Altova	fluid Operations	Netbreeze	Syllabs Sarl
Arisem	FTW	Netelligenz	Sympalog Voice Solutions
ASC telecom AG	Genkey Europe	Nomao	Syn-Tactic
Aspasia Knowledge Systems	Gnowsis	Ontoprise GmbH	Synthesio
Attensity Europe	Hanival Internet Services	Ontos	Tagmatica
Attentio	Intelartes sprl	Optelec	TEMIS
Averbis	INTRAFIND	OWI Technologies	TEXTEC Software
Berlinger System Engineering GmbH	IQSer	playence GmbH	Textkernel BV
Braintribe	Jouve SA	Proxem	Thales/Arisem
Cogia Intelligence	Kimengi	Q-Sensei	Transinsight
Collibra nv/sa	Knime	SAIL LABS Technology AG	Usoft B.V.
CrossMinder BVBA	Knowledge Concepts	SalsaDev	WCC Services
Cybertech	Kwaga	Saperion AG	Whatever
Datao.net	Language Tools	Semantic System	Whitestein
Definiens	LAYAR	Semantic Web Company	Yocoy Technologies GmbH
Dictanova	Lesson Nine GmbH	Semlab	zuuka

3.2 Collaborative Innovation for European LT

European R&D has produced a steady stream of small LT-based companies. Now opportunities are emerging to exploit LT on a wider scale that could boost the potential impact of the technology, creating both new markets and new business models to help LT companies thrive.

The pace of development could be accelerated through collaborative innovation initiatives that bring smaller companies together with their peers and with actors in the ICT value chain who are their natural partners.

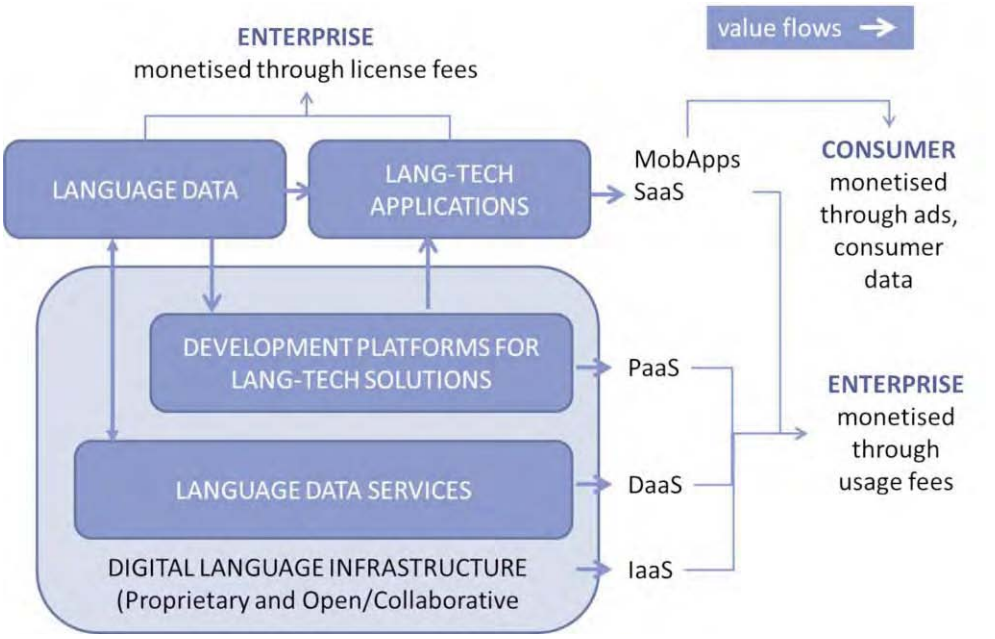


Figure 16: Collaborative Innovation

These partnerships could be oriented to one or more components in the value flow from data through development platforms to applications. No doubt there could (and should) be many different kinds of partnerships and collaborations, developing data resources in different domains, services based on those data, platforms for product development, and actual (cloud) applications.

The current European industry is concentrated in the Enterprise applications space, outside a sharable language infrastructure. Today that infrastructure (where it exists) is private and owned by major IT companies. In future there could, in principle, be an open infrastructure, available to any organisation that wanted to use its resources. The dynamics of the general software market, and the limits of what is currently possible for niche LT SMEs, strongly suggest that a Digital Language Infrastructure for Europe could both unlock potential for the industry, and help meet the need for pervasive multilinguality in Europe's digital economy.

The industry itself should define the nature and content of the infrastructure, what features are appropriately shared and open, what should remain in the commercial IP realm.

Innovation Scenarios

The review of conditions in the LT industry suggests that collaborative approaches to the market could break through the fragmentation that is evident. The following discussion explores different types of collaborative models, with scenarios that illustrate their potential use.

Styles of Collaboration

Different styles and structures of collaboration are suited to different market conditions and the incentives of partners. Styles may be open, between or among peers, or asymmetric; collaboration structures may be conventional partnership alliances or ecosystem networks. An ecosystem implies a platform intended for a large network of developers, whereas alliances partner around a particular product or business solution.

The table summarises a set of collaborative scenarios; the types of LT-industry collaborations most suitable for publicly funded projects are shaded blue.

Suitable	May be Suitable	Not Suitable
	ECOSYSTEM	**ALLIANCES**
OPEN	Core open-source technology as infrastructure for a network of organisations that participate at different points in the value chain.	Partners share intellectual capital, R&D risk, and rewards without long-term commitments.
PEER	Companies with relatively equal market strength exchange technologies across a network of overlapping application areas.	Companies with relatively equal market strength partner to share and develop complementary technologies.
ASYMMETRIC	A company with significant market strength facilitates a network of organisations that participate at different points in the value chain.	A company with significant market strength partners with a niche developer of innovative technologies.

Figure 17: Collaborative Styles

Collaborations "may be suitable" for public support in certain conditions:

- Open Alliances developing on proprietary platforms where results are likely to fill a gap (or a market failure) considered strategic, or with particular social value
- Peer Ecosystems among small market participants, where competitive imbalance is distorting the potential to innovate, or to extend LT features to desired markets or constituencies
- Peer Ecosystems among large market participants where network members represent a substantial proportion of actors in an economically significant sector
- Peer Alliances where outcomes are of particular social importance, such as public sector or security
- Peer Alliances where outcomes are likely to promote interoperability of systems that extends beyond the partnership
- Asymmetric Ecosystems where platform owners represent constituencies with particular social or economic value; associations, groups of companies but not A company
- Any of the above where valuable data is created and IPRs are contributed to an open infrastructure.

Scenario: Open Ecosystems

Core open-source technology as infrastructure for a network of organisations that participate at different points in the value chain:

- Leverages robust underlying technology
- A natural path for academic tech-transfer in collaboration with industry, good evolutionary path
- Well suited to collaboration in verticals, or clearly defined horizontal applications
- Potential for large scale

Example: KNIME - an ecosystem built on an open-source platform for a scientific domain

In early 2004 at the University of Konstanz, a team of developers from a Silicon Valley software company specialising in pharmaceutical applications started working on a new open source platform as a collaboration

and research tool. Because it was clear from day one that this product would have to process and integrate huge amounts of diverse data, the developers adhered to rigorous software engineering standards to create a robust, modular, and highly scalable platform encompassing various data loading, transformation, analysis and visual exploration models. When the first version of KNIME was released in 2006, several pharmaceutical companies began using it and, soon thereafter, software vendors started building KNIME-based tools.

Today, KNIME users can be found in large-scale Enterprises across a wide range of industries including life sciences, financial services, publishers, Retailers and E-tailers, manufacturing consulting firms, government and research – in over 50 countries.

KNIME Development Partners provide extension to KNIME for Life Sciences, Chemo- and Bioinformatics, but also high performance data analysis and other industry areas.

Channel Partners are officially entitled to resell KNIME Enterprise Products and to provide Support and Maintenance for KNIME Desktop and KNIME Report Designer.

KNIME Consulting partners provide consulting services in the area of business intelligence, CRM analysis, financial services, LIMS integration, Life Sciences and other industries.

Scenario: Open Alliances

Partners share intellectual capital, R&D risk, and rewards without long-term commitments:

- Foster more speculative development projects, potentially more early-stage (like seeding, but doesn't have to be with start-ups)
- Suitable for small scale
- Could use open technology but doesn't have to

Example: AppCampus - an open alliance between major IT companies and start-ups

AppCampus is a mobile application development program based at Aalto University in Espoo, Finland. AppCampus is a pre-seed funding program that was announced in March 2012 by Microsoft and Nokia and hosted at the Aalto centre for Entrepreneurship. The AppCampus program has been set up to foster the creation of innovative mobile applications for the Windows Phone ecosystem and other Nokia platforms, including Symbian and Series 40. The program aims to create a new generation of successful mobile start-ups by attracting and supporting thousands of students, entrepreneurs and other talented people with various backgrounds from all over the world. AppCampus has committed to invest up to €18 million over three years into new applications and services. Apart from grant funding, the program offers mobile entrepreneurs comprehensive coaching, marketing support and training in mobile technology, design and usability.

The sponsors don't take equity or commission from the investment, but successful apps are expected to be available exclusively on Windows Phone and other Nokia platforms for the first six months after launch.

Scenario: Peer Ecosystems

Companies with relatively equal market strength exchange technologies across a network of overlapping application areas:

- Industry driven, organic growth

- Low barriers to participation
- Focus on interoperability, likely to promote standardisation
- Proxy for formal consolidation with more open flavour (but not open technology)

Example: iSpeech Cloud - a Peer Ecosystem of start-up innovators with complementary technologies

The US company iSpeech was purportedly founded to develop and deliver ASR and TTS on a cloud platform, though its first product was a more modest Blackberry app. The company claims that its service is intended to compete with Nuance as a new platform for speech technology development. It claims to have extremely accurate speech recognition, created from large, unique speech datasets, and the ability to generate human-quality speech in natural sounding voices. According to the company, out of 12,000 developers who have registered on the iSpeech cloud 3,000 have actually used it, generating more than a billion transactions.

The iSpeech self-service platform allows developers to integrate text to speech and speech recognition into their own apps, products and devices, as of this year including home electronics and appliances. The service is free to all mobile developers with support for iOS, Android and BlackBerry devices. The iSpeech SDKs and APIs provide over 40 text to speech voices with support for more than 25 languages.

Developers have signed up for embedding iSpeech in applications, appliances, automobiles, websites and platforms. iSpeech Cloud handles tens of millions of connections daily, across the globe.

mylanguage, developers of the Vocre speech translation app for iOS devices that launched in 2011, is a member of this ecosystem. Vocre brands itself as a translation company for the 21st Century focused on dialogue and the user experience, and considers its competitors to be Google Translate (which it refers to as an "inferior app") and Jibbigo (which unlike Vocre offers offline translation as well as online). It is also active in the language learning space, and considers Rosetta Stone (language learning software) a competitor. It claims millions of downloads, and tens of millions of translations. Claims to have developed its own translation technology.

Scenario: Peer Alliances

Companies with relatively equal market strength partner to share and develop complementary technologies:

- More traditional partnerships where collaborators are on equal footing
- No "imbalance of power"
- Heavyweight exchanges of technology possible with large peer partners
- Low-risk access to complementary technology for small peer partners
- Potential for long-term joint ventures

Example: Nuance & IBM (large peers) – merging intelligent technologies for new healthcare solutions

Nuance and IBM collaboration on research to commercialize the Watson computing system's advanced analytics capabilities in the healthcare industry (announced in 2011). The research and technology initiative will combine IBM's Deep Question Answering (QA), Natural Language Processing, and Machine Learning capabilities with Nuance's speech recognition and Clinical Language Understanding (CLU) solutions for the diagnosis and treatment of patients that provide hospitals, physicians and payers access to critical and timely information. The two companies expect the first commercial offerings from the collaboration to be available in 18-24 months.

Additionally, Columbia University Medical Center and the University of Maryland School of Medicine are contributing their medical expertise and research to the collaborative effort. For example, physicians at

Columbia University are helping identify critical issues in the practice of medicine where the Watson technology may be able to contribute, and physicians at the University of Maryland are working to identify the best way that a technology like Watson could interact with medical practitioners to provide the maximum assistance.

Scenario: Asymmetric Ecosystems

A company with significant market strength facilitates a network of organisations that participate at different points in the value chain:

- For small companies, wider access to markets, customers, without loss of autonomy
- For large companies, build capabilities without expense and risk of acquisition
- Can provide a rich mix of technologies
- Well suited to collaboration in verticals

Example: Salesforce Marketing Cloud ecosystem

Salesforce has launched the Marketing Cloud Social Insights ecosystem in October 2012. The product allows users to track social influence and sentiment in 17 languages, as well as identify sales leads, using data produced by 20 analytics services.

The Marketing Cloud ecosystem includes partners Bitext, Calais, Caterva, Clarabridge Link, EpiAnalytics, Kanjoya, Klout, Kred, LeadSift, Lexalytics, LinguaSys, Lymbix, Metavana, OpenAmplify, PeekAnalytics, Rapleaf, Solariat, Soshio, The SelfService Company and Trendspottr.

Scenario: Asymmetric Alliances

A company with significant market strength partners with a niche developer of innovative technologies:

- Purely tactical
- Channel partnerships that give access to customers for smaller partners
- Commitment of training from larger to smaller partner
- Can have characteristics of a marketplace
- Could lead to acquisition

Example: ABBYY and Google Enterprise – OCR as a background service for search

ABBYY, a provider of document recognition, data capture and linguistic technologies and professional services, joined the Google Enterprise Partner program in October 2012. The program extends the power of Google technologies across the enterprise and helps customers get more value out of their Google enterprise deployments. The ABBYY integration provides organizations with a single solution that simplifies the management, organization and locating of image-format documents.

The Google Enterprise Partner program includes developers, consultants and independent software vendors that provide value-added services for Google enterprise products. As part of the Google Enterprise Partner program, ABBYY received training, support and deployment services from Google, allowing the company to retain a close relationship with its customers in order to provide additional services and support.

Mapping Collaboration Scenarios to Objectives

Collaboration modes may be mapped to Henry ("father of open innovation") Chesbrough's map, "Knowing How and Where to Play". LT SMEs developing collaboration strategies are advised not to play in more than one space on the map. While not strictly rigorous, the three styles map roughly to the "direction" a collaboration may take.

Figure 18: Mapping Scenarios

Asymmetric partnering for SMEs is a natural route to developing technologies in specialist areas with steep technical demands (heavy R&D), where domain expertise is key. Opportunities in verticals (high end financial systems, oil & gas exploration, pharmaceutical research) may be found through partnerships with specialist companies already active in those domains. LT developers innovate by applying new technology to known problems, or addressing business challenges in new ways.

Peer partnering takes the alternate route of creating new "breakout" categories of products or services through the collaborative combination of complementary technologies. Opportunities in the emerging social/mobile ecosystem tend toward this quadrant, though Peer Ecosystems could be developed for virtually any market area.

Dominant markets are those where technical depth meets the greatest opportunity. Leveraging the capacities of open platforms could enable niche companies to participate in markets that are otherwise dominated by much larger players.

These scenarios will be used as a basis for exploring collaborative opportunities within LT-Innovate, and outside the network with potential partners in projects.

4. Speech overview

4.1 Background

Speech technology is no longer the stuff of science fiction. Adoption of new IT platforms and hardware, coupled with unforeseen increases in the processing power, lower IT prices, and miniaturization of electronic components have supported to make speech technology nearly ubiquitous, commercially viable for speech suppliers and service providers. In fact, speech recognition is widely recognized and used not only as a cost-cutting mechanism to provide cost-effective customer care and self-service in business, but also providing to consumers an easier, more convenient, cleaner, and more fun way to control products and personal devices, such as home appliances or cell phones.

Speech technologies are human-to-computer interfaces providing multimodal and natural spoken interaction (e.g. human computer dialogue) thus human and computers or devices can interact in both directions, thus texts and speeches generated by machines could be readable, processed and understood by humans, and those generated by humans, could be readable, processed and understood by machines.

Speech interaction technology aims at enabling people to communicate with machines and devices (everything from phones to cars, e.g. PCs, cell phones, GPS, etc.) using spoken language (through natural interfaces), requiring very little training and in a cost-effective way. Thus they comprise the specification of the interaction between a user and a computer system, where the primary interface is audio input and/or output. This goal has increased importance in as mobile world grows; voice is becoming an essential interface mode for smartphones and any mobile device.

Significant milestones in speech technologies were achieved in the 1950`s and 1960`s, where research in this technology paved the way to commercially feasible designs, hardware and applications, providing not only greater understanding of speech as it related to human behavioural and conversational patterns, but also developments of artificial and synthetic sounds.

Milestones in the history of Speech Technology

1952	Creation of a small vocabulary recognizer for digits over the phone, by Bell Labs.
1962	IBM developed Shoebox - a 16-word speech 16-word speech recognizer, used to interface with a calculator.
1972	Introduction of stochastic processing with Hidden Markov models to speech recognition.
Mid-1970s	Creation of a small vocabulary recognizer for hands-free industrial applications, by NEC and Threshold Technology.
Late-1970s	Verbex launches a speech application based on a small vocabulary recognizer, which is useful for telephone toll management and financial services.
1990	Introduction of a general-purpose dictation application, by Dragon Systems.
1992	AT&T unveiled automated operator system, which is capable to understand spoken utterances that include ‘operator’ and ‘collect call’.
Late-1990s	Large-scale deployment of commercial speech recognition solutions.

Despite increase of Interactive Voice Response (IVR) technology in the 1970s to automate tasks in call centres, it was still expensive and complex. During the 90`s first large commercial applications were deployed in different areas, for example: Dragon brings the first dictation application, AT&T launches and automatic call collection operating system.

Although early 90's represented incremental and critically important progress in practical application for developing both voice-independent speech recognition and synthesis technologies, during this year's speech technologies still were facing big challenges. Speech technologies strongly depended on the user training of the system; which was a very high-consuming process. Users recording needed to be associated manually to phonetic or spelling combinations and still produced voices sounded very artificial and not human natural.

It was not until late 90`s that large scale voice recognition and unified messaging are seen, when companies started to invest in Computer Telephony Integration (CTI) with IVR systems; which became vital for call centres. Speech recognition vendors focused on small niches such as customer services with basic speech-recognition applications which understanding caller requests and respond to spoken prompts under limited acceptable options (e.g. voice activated dialling, call routing, etc.). Universal queuing and routing solutions started to be widely deployed in call centres as they acted as an agent collecting customer data and enabling intelligent routing decisions.

Currently sophisticated applications can accept widely varied and highly complex caller requests, enabling fully automated transactions or customer self-service —- for example, accepting payments and entertainment ticketing, banking transactions, or collecting personal information. In fact, nearly every industry segment (communications, financial services, government, healthcare, retail, travel and tourism, etc.) has now implemented automated speech dialog at some level, from simple call routers to fully automated self-service and even purchase/transaction applications.

4.2 Speech Input or Automatic Speech Recognition

Speech input applications convert spoken language in text or machine-readable format and they are also known as Speech Recognition, Automatic Speech Recognition (ASR), Computer speech recognition, Speech-to-Text (STT), or even sometimes known in the market as voice recognition, natural language interface, even natural language understanding. Natural Language Interface is more often used to describe environments that integrate speech and semantic information - making reference to some underlying model of "knowledge" for Natural Language Understanding (NLU) - rather than pure speech recognition, though these distinctions are increasingly blurred in the interest of marketing.

Commercially available speech recognition technology is behind applications such as voice user interfaces such as phone services (e.g. voice dialling such as saying «Call home» to our cell phone, simple data entry such as entering a credit card number), dictation and preparation of structured documents (e.g., a radiology report), speech-to-text processing (e.g., word processors or emails), and aircraft (usually termed Direct Voice Input).

Thus this technology appeals to a wide range of users going from organisations seeking for enhanced personal productivity in automated commercial phone systems to anyone who needs or wants a hands-free approach to computing tasks or communicate.

The ultimate goal of automatic speech recognition technology is to deliver speaker independent speech recognition services that not require (or requires a minimum) user training, and which can be used and performs well with multiple users and under all conditions.

The terms «speech recognition» and «speaker recognition» are sometimes confused. Speaker recognition is the task of validating a user's claimed identity using characteristics extracted from a speaker's voice. Thus we need to differentiate between speaker recognition (recognizing who is speaking) and speech recognition (recognizing what is being said).

These two terms are also frequently confused, as is voice recognition. Voice recognition sometimes is used

to refer to biometric technology that identifies a particular individual's voice, sometimes to describe a system that has been trained to recognize a particular person's voice. There is a difference between the act of authentication (commonly referred to as speaker verification or speaker authentication) and identification.

Types of Speech Recognition Technology

Although the ultimate goal of automatic speech recognition technology is to provide out of the box speaker independent speech recognition services, right now we can classify ASR in three types: speaker independent, speaker dependent and speaker adaptive. All models use mathematical and statistical formulas to yield the best work match for speech.

Speaker independent

ASR recognizes the speech patterns of a large group of people and responds to many users; they are very attractive from commercial point of view. Currently, commercially available speaker independent systems are based on *discrete or isolated word systems* and recognize only specific, well-enunciated single words. Now days there are many scenarios where it is not possible to adapt ASR to individual speakers, for example the call centres, where callers are unknown and speak only for a few seconds; thus still many routing, purchasing or transaction banking systems are based on discrete word systems.

Speaker dependent technologies

In these systems accuracy is paramount, and voice-to-text typically achieves this by having the user "train" the software during setup and by adapting more closely to the user's speech patterns over time. Speaker dependent models recognize speech patterns from only one person. This technology is seen in commercial dictation packages for medical, legal and business professional transcription. For example Dragon dictation or MacSpeech Scribe, which claim to have accuracy over 90%, recognize users` voices and speech patterns after training.

Speaker adaptive technology

This is an emerging variation of the two models above. It usually begins with a speaker independent model and adjusts these models more closely to each individual during a brief training period.

Both speaker dependent and adaptive technologies are harder to implement as each person's speech has unique spectral features. These models currently use *connected word recognition or continuous systems to respond to human normal speech*. Connected word technology recognises each word in a limited vocabulary while continuous systems provide conversational speech recognition with large vocabulary and understanding the sentences; they are difficult to implement, and few exist.

4.3 Speech Output or Text-to-Speech (TTS)

Speech output applications generate intelligible and natural sound speech over different machines or devices enabling the conversion from text or machine-readable format to verbally spoken language. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. A text-to-speech (TTS) system converts normal text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech.

The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood. An intelligible text-to-speech program allows people with visual impairments or reading disabilities to listen to written works on a home computer. Many computer operating systems have included

speech synthesizers since the early 1980s.

Vendors active in text-to-speech include Cereproc, BuddyBird, NeoSpeech, TextSpeak, Acapela, and Nuance.

Applications

Virtual assistants for business

Voice-enabled virtual assistants for the enterprise inspired by Apple's Siri application, such as Nuance Communications' Nina, Angel's Lexee, and Taptera's Sophia became available during 2012. Unlike Siri, these virtual assistants are designed for business purposes.

Emotive voices

Several European vendors are developing emotion-rich voices including Cereproc and Acapela. MARY Open-Source Emotional Text-to-Speech Synthesis is a multilingual (German, U.S. and U.K. English, Turkish, and Tibetan), multiplatform (Windows, Linux, Mac OS X, and Solaris) speech synthesis. With EmoSpeak tool, MARY can synthesize emotionally expressive speech using diaphone voices. Users can also select expressive unit voices, such as a German soccer announcer and control the intonation

Loquendo TTS voices have been enriched with expressive cues that allow for highly emotional pronunciation. These cues, which can be typed directly into the text with the appropriate punctuation or selected from a drop-down menu, contain conventional figures of speech, such as greetings and exclamations (hello, oh no, thank you), interjections (oh, well, hmm), and paralinguistic events (breathing, laughing, coughing), to convey additional layers of expressive intent, such as gratitude, doubt, or confirmation. A Pronunciation Lexicon ensures that specialized vocabulary, abbreviations, acronyms, and even regional pronunciation differences sound as the developer intended them. Loquendo supports 30 languages with a total of 72 voices.

4.4 Speech ID/Verification, Voice Biometrics

Also known as speaker recognition, speech ID / verification comprise methods designed to distinguish between real human users and computer programs that are interacting with the system, verifying that a human is interacted with the automated speech based system or application.

Speaker recognition has a history dating back some four decades and uses the acoustic features of speech that have been found to differ between individuals. These acoustic patterns reflect both anatomy (e.g., size and shape of the throat and mouth) and learned behavioural patterns (e.g., voice pitch, speaking style). Speaker verification has earned speaker recognition its classification as a «behavioural biometric».

From a security perspective, identification is different from verification. For example, presenting your passport at border control is a verification process - the agent compares your face to the picture in the document. Conversely, a police officer comparing a sketch of an assailant against a database of previously documented criminals to find the closest match (es) is an identification process.

Speaker verification is usually employed as a «gatekeeper» in order to provide access to a secure system. These systems operate with the user's knowledge and typically require their cooperation. Speaker identification systems can also be implemented covertly without the user's knowledge to identify talkers in a discussion, alert automated systems of speaker changes, check if a user is already enrolled in a system, etc.

In forensic applications, it is common to first perform a speaker identification process to create a list of «best matches» and then perform a series of verification processes to determine a conclusive match. Also, there

is a difference between speaker recognition (recognizing who is speaking) and speaker diarisation (recognizing when the same speaker is speaking).

Applications for Speaker ID

Major breakthroughs in the development of voice authentication have taken place in the last decade, and the market potential for this technology is evolving rapidly. Due to the unique characteristics of the human voice, voice authentication is being used to verify the identity of individuals for the purposes of public safety and national security or to ensure the confidentiality of sensitive data or information. The United Kingdom government's Supervision and Surveillance Programme uses voice authentication to identify and control curfew orders of youth parolees. Compared to other biometrics voice authentication is easier and more cost effective to deploy – since it works from any telephone – and can handle a range of applications from facilities access to the protection of sensitive data.

On average it takes less than two minutes to create a 'voiceprint' based on specific text such as 'Name' and 'Account Number'. This is then stored against the individual's record, so when they next call, they can simply say their name and if the voiceprint matches what they have stored, then the person is put straight through to a customer service representative. This takes less than 30 seconds and also bypasses the need for the individual to have to run through a series of tedious ID checks such as passwords, address details.

Market	Application	Drivers
Financial Services	Access to Banking, Brokerage, 401K	Reduce Financial Risk
Telecom	Calling Card	Reduce Fraud
	Cellular Roaming	Protect Personal Information
	Unified Messaging	Competitive Advantage
	Auto Attendant	
Retail	Order Entry	Reduce Fraud
	Personalized Service	Increase Revenue (1:1Marketing)
Enterprise and IT	Access to Intranet, Extranet and Corporate Applications	Increase Security
	PIN Reset	Reduce Cost
Travel	Frequent Customer Services	Convenience
		Personalization
Internet	Authenticate Users for	Reduce Financial Risk
	Internet Banking and e-Commerce	
Health: HMOs, Hospitals, Insurance	Access to Patient Information	Protect Personal Privacy
	Authorize Drug Prescription	Meet US HIPAA (Health Insurance Portability and Accountability Act of 1996)
	Authorize Insurance Payment	Reduce Fraud
Government/Military	Access to Sensitive Information	Increase Security
	Parolee Tracking	Reduce Cost

Global enterprises can use voice authentication to allow customers to conduct secure financial and consumer transactions over the telephone without needing to speak to an agent. Moreover, the identity of employees can be verified before they are granted network access. Voice authentication also adds a new level of security to pass-code controlled systems, or to entirely eliminate pass-codes (passwords and Personal Identification Numbers – PINs). Traditional pass-code oriented schemes have two major flaws: people continually forget their pass-codes and pass-codes can be guessed or stolen. Both these flaws represent

real costs to companies as they attempt to offer secure transactions. Enterprises are spending an exorbitant amount of money resetting customer passwords that are forgotten or stolen. Customers have come to expect convenient 24/7 access to their account information. In fact, companies that do not provide such access are viewed as being difficult to do business with. It is simply not cost-effective to attempt to meet this requirement with more live agents. Today, billions of dollars are spent on call centre agents. These agents must handle a wide range of functions – everything from taking orders, answering support questions, to setting up returns. It turns out that agents spend a huge percentage of their day answering simple account status questions

Financial Services

Provide effective risk management, compliance with critical industry regulations, and the ability to successfully combat increasing fraud and identity theft. Enable simple and secure remote authentication for a variety of applications including:

- Phone Banking Authentication
- Password Reset
- e-Banking Transaction Security
- Transparent Conversational Authentication
- Private Banking

Telecommunications

Support delivering services to mobile device customers and internet users securely and cost effectively. Enables simple and secure remote authentication for a variety of applications including:

- Contact Centre Security
- Secure Access to Value-Added Services
- m-Commerce
- Password Reset

Healthcare

Provides an accurate, non-invasive, cost-effective identity verification solution that seamlessly integrates with existing contact centres and VoiceXML platforms. Enables simple and secure remote authentication for a variety of applications including:

- Contact Centre Security
- Access to Patient Information
- Password Reset

Global Enterprise

Protect private information, improve efficiency, and enhance the delivery of remote services via contact centres and the Internet. Voice biometrics enables secure remote authentication for a variety of applications including:

- Password Reset
- Secured Conferencing
- Time and Attendance
- Voice Portal Access

4.5 Speech Dialogue/Interactive Speech

Speech interaction is a core feature of emerging unified interface and interactivity modes, where multiple input modalities aim at compensating for the weaknesses of one modality, offset by the strengths of another. For this purpose the Voice–user interface (VUI) is the interface of any speech application. It makes possible the human-computer/machine interaction trough a voice/speech platform as the primary interface is audio input and output and generates an automated service or process.

Applications

Voice Portal

System that uses advanced speech recognition technology and provides access to information on the Internet. Key components of most voice portals are speech recognition, text to speech, information aggregation, categorization software, telephony and Internet interfaces, and administrative interfaces. Optional components include software to support context-sensitive, personalized assistance (for example, an intelligent assistant) and support for VoiceXML.

Home appliances & Consumer electronics

The recent successes of new touch-based user interfaces (iPod, iPhone, Wii, etc.), has opened the eyes of home appliance and consumer electronics manufacturers to the benefits of giving customers new and better ways to control products and access information. Speech I/O is the perfect solution to give consumers an easier, more convenient, cleaner, and more fun way to control products in the home.

Speech synthesis output can be used for confirmations (“oven temperature set to 105 degrees”), queries (“Please say the device you would to control by voice”), or even fun and information purposes, like a refrigerator that tells a joke or a fun fact when it’s held open for more than 3 seconds.

ASR can substitute the controller that already exists in remote controls, microwaves, ovens, dishwashers and other home electronics, allowing the addition of speech I/O with very little incremental cost. Software solutions are needed for low cost speech recognition, thus enabling customers to use the best platforms for their particular applications.

Automotive

The increasing usage of cell phones, digital music players, and GPS units in automobiles distracts drivers and creates potentially hazardous situations. Thus many countries and regions have already started requiring compliance towards hands-free operation of cell phones while driving, and it is foreseen to grow and further cover personal navigation devices and other potential in car distractions; for which ASR offers the opportunity to trigger a user interface for accessing these devices safely, while on the road, with voice control and truly hands free systems, without having to locate and look at panel buttons.

Development of ASR for vehicles is based in four areas:

- hands-free use of mobile phone handsets in the car e.g. “Dial office”
- Speech instructions to navigation systems e.g. GPS-connected digital maps: “How far is it to the motorway junction?”
- In-car system interaction e.g. “Turn on the radio to the travel reports channel.”
- in-car steering systems Industrial and Medical

Industrial and medical

Ease of use and improved user experience provided by voice control and synthetic speech output is an excellent way to differentiate and add value to their products.

It is seen as a practical extended feature. The VR stamps take a Sensory RSC chip and add all the necessary external requirements to enable a simple hook up to microphones and speaker.

Banking

The smartphone is evolving at a rapid pace. The iPhone 5 and the Galaxy S3 are pushing the limits of phone-based technology with new features—near field communication (NFC), speech recognition, voice authentication—and more appealing user interfaces, such as Swype, resolution, colours, and screen size. Some capabilities require accessing the native code of the device, while others leverage hybrid solutions with parts of the application accessing cloud-based capabilities. Apple and Android, with dominant market share, take advantage of these capabilities to attract new smartphone customers.

4.6 Vendor Landscape

The Speech Recognition Industry is dynamic and is witnessing development and adoption of several new applications. There is a growing demand for speech recognition applications across a number of industries such as Healthcare, Telecommunications, Logistics, Tourism, and Financial Services. These industries are upgrading their existing manual systems such as healthcare transcriptions, biometrics applications, in-vehicle telematics, mobile applications and others with automatic speech recognition capabilities.

The majority of the growth in the Speech Technology market comes from the Americas, followed by the Europe, Middle East, and Africa (EMEA) region. Although the majority of the growth currently comes from the Healthcare and Enterprise segments, other segments such as Consumers and Mobile devices are expected to significantly increase their contributions in the next few years.

One of the main drivers in the speech market is the increased demand in the Mobile Applications segment. This segment is witnessing high demand for speech recognition applications because of the numerous regulations on the use of mobile phones while driving. The vendors in this segment are developing cost-effective and easy-to-deploy automatic speech recognition applications. The increasing demand for biometrics to address safety concerns is another major factor driving the need for automatic speech recognition applications.

The speech applications market is dominated by enterprises, which are the main end-user segment and mainly comprises contact centres. Enterprises accounted leaded the market in 2011. The Global ASR Applications area mainly used by contact centres, although embedding this technology in electronic appliances and automobiles is becoming a major trend.

Nuance Communications is the market leader. The company uses acquisitions as a major strategy to retain

its dominance in the market. It has acquired many of its competitors, including some of the global vendors such as Phillips Speech Recognition Systems and the Speech Recognition division of IBM. Recently, it acquired Loquendo Inc., which was one of its main competitors in the market.

LumenVox LLC. is in second place in the Global ASR Applications market. This vendor has a strong presence in Europe and in some Asian countries. Telisma S.A. (On Mobile Global Ltd.) is in a strong position in the market and an Automatic Speech Recognition and text-to-speech (TTS) applications provider. Telisma was acquired by India-based company On Mobile Global Ltd. in 2010. Telisma has formed strong alliances with speech recognition platform and service providers, and has a strong position in Asian countries, particularly in India.

Other major vendors in the Global ASR Applications market are AT&T and Raytheon BBN Technologies.

These vendors provide dictation software in hosted interactive voice response platforms. They provide TTS and automatic speech recognition applications to both enterprises and no business customers. Some of the other vendors in this market that pose a threat to the market shares of Nuance are Microsoft Tellme, and Voxeo Corp. Micosoft Tellme is a subsidiary of Microsoft Corp. and offers efficient ASR applications to individual end-users along with ASR applications to the enterprises. Both Google and Microsoft are active in mobile and internet markets.

Company	Description
Nuance Communications	Global speech recognition applications provider
LumenVox LLC	Global ASR and TTS applications provider
Telisma/On Mobile Global Ltd.	Global ASR applications provider
AT&T	TTS and ASR applications provider
Raytheon BBN Technologies	ASR, TTS and professional services provider
Microsoft Tellme	Cloud-based ASR applications provider
Voxeo Corp.	Directed dialog ASR provider

4.7 Opportunities and Challenges for Speech Technology

Speech Technology Opportunities

The speech applications market shows immense potential, and it is expected to grow rapidly in the next few years. The market has been witnessing a transition from the use of embedded automatic speech recognition applications to the use of networked and cloud-based automatic speech recognition applications. Vendors are gradually shifting to networked and cloud-based models to realize the full potential of the market across various end-user segments. In addition, the Mobile Devices End-user segment is witnessing numerous technological advancements, which is expected to further improve the growth of this segment in the Global Automatic Speech Recognition Applications market.

The market is mainly driven by the increased demand in the Mobile Devices segment. This segment is witnessing high demand for speech recognition applications because of the increase in the number of regulations on the use of mobile phones while driving. Moreover, the vendors in this segment are developing cost-effective and easy-to-deploy ASR applications. In addition, growing security concerns have increased the demand for biometrics, which in turn is driving the need for ASR applications.

A decoupling of Apps from phones may eventually become reality as more content and services are shifted to being hosted on the cloud rather than on native Apps. This could potentially reduce the motivation for

smartphone integration in the longer-term and encourage the use of embedded telematics.

Increased Use of Biometrics

Initially, biometrics was a terminology that was familiar only in the medical field. However, the growing concern for security has increased the adoption of biometric technologies among enterprises and governments. This is one of the key drivers in the Global ASR Applications market because ASR applications are used along with voice authentication applications. These applications are used to verify the authenticity of users using their voice patterns. A combination of these two applications helps to identify the authenticity of the voice source along with interpreting and replying to the person. This ensures a high amount of accuracy and security. In addition, the low cost of biometric devices has increased the demand in small and medium businesses (SMBs). The use of ASR applications for attendance purposes is another major criterion which is increasing the demand for ASR applications across various enterprises.

Improved Accuracy Levels

Currently, vendors have the capability to provide ASR applications with more than 90% accuracy. In addition to this, these applications have the ability to recognize various languages and generate speech-based automated replies. Because of these features, enterprises are increasingly adopting ASR applications to improve their productivity and turnaround time. In addition, high accuracy levels of the ASR applications have substantially reduced the cost and time that is spent on customer care. Thus, the accuracy of the ASR applications in recognizing various languages and vocabularies has increased demand among small and large enterprises that are seeking such applications to optimize their processes.

Customer Care in many languages

In a competitive environment, it has become imperative for enterprises to differentiate their marketing activities with better customer care services. Thus, to achieve product differentiation, vendors have increased the capability of the ASR applications to support various languages. This language-specific customization has helped enterprises improve customer satisfaction levels.

Consumer-focused applications can support Speech awareness: Speech interfaces on products such as mobile phones (e.g. Siri) and within vehicles will lead end users to the potential capabilities of Speech. Due to Consumerisation of the workplace, we will also see employees expecting enterprises to offer similar capabilities to help them undertake their work tasks

Growing markets in speech recognition can anchor growth for LT and benefit from LT enhancement: Speech leads to IC (recognizing contexts and concepts), applications such as call routing also include analytics, natural language processing and semantics.

Speech Technology Challenges

Inability to Suppress Ambient Noise

One of the main challenges is the inability of speech recognition applications to suppress ambient noise. Speech recognition applications are highly sound sensitive and hence any ambient noise reduces their accuracy levels. Although the Speech Recognition industry has witnessed several technological advancements, the issue of high sensitivity to noise still remains a barrier to the adoption of automatic speech recognition applications.

Despite the technological advancements in the Speech Recognition industry, ASR applications are highly sen-

sitive to ambient noise. This is one of main barriers to achieving accuracy. As a result, any disturbance nearby prevents the application from accurately recognizing the voice source. This also disrupts the automated reply for the spoken command. The inability of the application to suppress the ambient noise is the only factor that is keeping vendors from achieving 100% accuracy.

Limitations in Automated Speech Services

ASR application vendors offer automated speech recognition applications which can replace humans by providing automated speech services. These services include travel reservations, hotel bookings and in contact centres. Although vendors are offering human-like experience in automated travel reservations, the limited grammar and language input into this automated system is a challenge in the Global ASR Applications market. In addition, the consumers must have some knowledge about the system and the limited number of words that the system understands. Since the system cannot understand all the human needs it is still limited in the market. If given a choice human operators will generally be given a preference over automated voice generated machines. This is the major reason behind the lower adoption rate of ASR applications in most developing countries.

Out of human tendency, most users will prefer to talk to a human operator. It is a general perception that a human can understand the needs of another human in a better way and so users calling a customer care contact centre of any enterprise prefer to talk to a human operator who can provide an efficient solution to their queries and enquiries. Therefore, preference for a human operator is a challenge not only in developing countries but also in developed countries.

Lack of Awareness about ASR Applications

One of the major challenges in the Global ASR Applications market is lack of awareness about the applications and their benefits. Most enterprises are still under the misapprehension that these applications are difficult to adopt and implement. Hence, the adoption rate of these applications in the Enterprises and Government sectors has been low.

Speech Technology Drivers and Inhibitors

The following table presents macroeconomics, global megatrends, specific market trends and labour supply-factors that might affect the speech market.

Market Force	Assumption	Impact	Time Frame	Accelerator /Inhibitor/ Neutral	Certainty of Assumption
Macroeconomics					
Economic situation	There is a strong correlation between the economy and IT expenditures. The global economy situation is impacting IT budgets, business and consumer confidence, the availability of credit and private investment, and internal funding.	High	Short-term	Inhibitor	★★★ ★★
Global megatrends					
Mobile	Spending on mobile devices will grow 23%, driving 43% of IT growth; a mobile strategy is priority number one for all industry players in 2012. Consumers are interested in personalised services which could be enabled through LT technologies.	High	Short-term	Accelerator	★★★ ★★
Cloud Computing	Cloud as a new paradigm of computing that is reshaping IT will help to evolve speech technologies. The key advantage to cloud services should be the ability of IT organizations to shift IT resources from maintenance to new initiatives. IDC estimates that cloud services (public cloud) increased 34% in 2010 to nearly $22 billion, or about 1.6% of IT spending, and that percentage should increase to 3% by 2014.	High	Medium-term	Accelerator	★★★ ★★
Big Data	The data growth is pushing the need for storage, analysis and big data technologies. The latter has intersection with LT technologies and its growth will push new developments and revenues for Lt vendors.	High	Medium-term	Accelerator	★★★ ★☆
Smart Verticals	Cloud, mobile, social and big data / analytics are the base for the vertical industry evolution. Many organizations have started the path to become a smart vertical: industries generating new competitive advantages based on the third platform pillars.	Moderate	Long-term	Accelerator	★★★ ☆☆
Open source	Nowadays, there are few speech open source projects. However, open source has the potential to transform the speech services industry. For example, siri-based virtual assistant for Android, a mobile open platform, is helping to educate new potential customers.	Low	Long-term	Accelerator	★★★ ☆☆
Privacy and security	As speech analytics, new services and data insights increase, new concerns about the privacy and security of such data appears. Organizations must not only protect against new threats but to respect the limits of the private lives of their employees, partners and customers.	Moderate	Medium-term	Inhibitor	★★★ ☆☆
Emerging markets		High	Medium-term	Accelerator	★★☆ ☆☆
Specific market trends					
Multilingual speech recognition costs	High quality voices requires high quality sources. The least common is a language the more expensive and time consuming will be the requires audio resources. This barrier increases the multilingual speech recognition costs.	High	Short-term	Inhibitor	★★★ ★☆

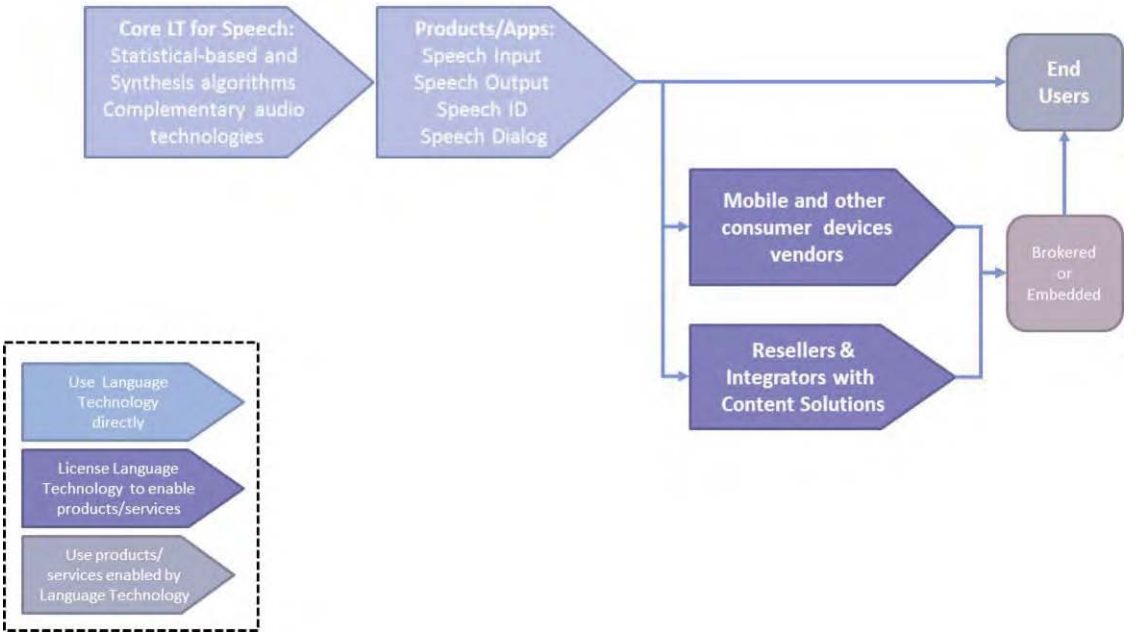
Lack of standardization and shared resources	Linguistic and audio data resources are usually expensive or no shared by companies. LT companies must recreate the required resources inhibiting a faster time-to-market and the proper market development.	Moderate	Short-term	Inhibitor	★★★☆☆
LT Consolidation	LT are mature and are living a progressively convergence not only between them but also with other technologies such as content analytics, In-memory processing, in-database analytics, search-based applications, non-row databases, and non-SQL analytics.	High	Long-term	Accelerator	★★★★☆
Accuracy	Speech technologies can be considered mature as they have 90% accuracy. This performance is helping the market to growth whenever a solid business case is found.	Moderate	Short-term	Accelerator	★★★☆☆
Business solutions awareness	Although business applications for call centres are well-know, there still exists ignorance regarding the potential uses for speech in other industries. Speech vendors should increase the awareness of new business cases.	Moderate	Medium-term	Inhibitor	★★☆☆☆
Biometrics massification	The performance of speaker verification products, and the ubiquity of digital access, is also boosting take-up in biometrics.	High	Medium-term	Accelerator	★★☆☆☆
Speaker adaptive technology maturity	The highly costs of speaker adaptive technology are an important factor to deliver new business value propositions. Whenever the cost will be reduced (using for example Cloud), the market will be able to boost its growth.	High	Medium-term	Accelerator	★★★☆☆
Labour supply					
Talent scarcity	The availability and the skill level of talent have a direct impact on LT markets such as speech, translation or intelligent content. Whilst in the previous decades, the LT talent scarcity has been covered by IT experts, the current availability may inhibit adoption rates and market development and growth.	Moderate	Medium-term	Inhibitor	★★★★☆

Legend: ★☆☆☆☆ very low, ★★☆☆☆ low, ★★★☆☆ moderate, ★★★★☆ high, ★★★★★ very high

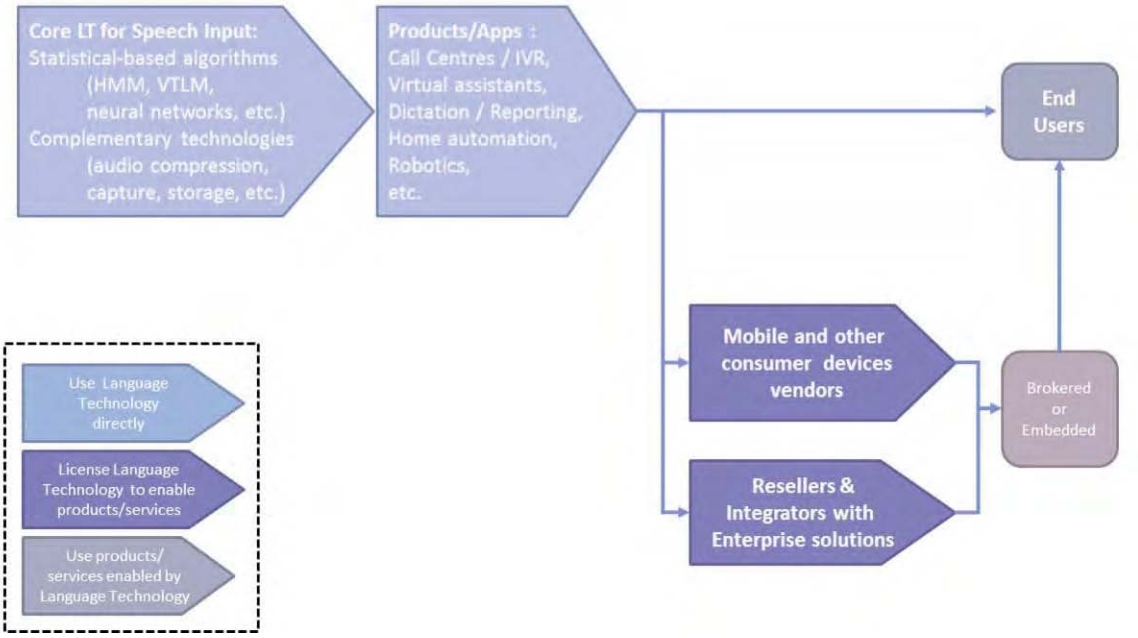
Speech Technology Value Chain

The following graphs present the value chain for Speech Technologies as well as for the different speech technological components and tools. This value chain also represents the value of speech technologies when often implemented for high throughput in operations.

Value Chain: Speech Technologies
Core Tools & Techniques

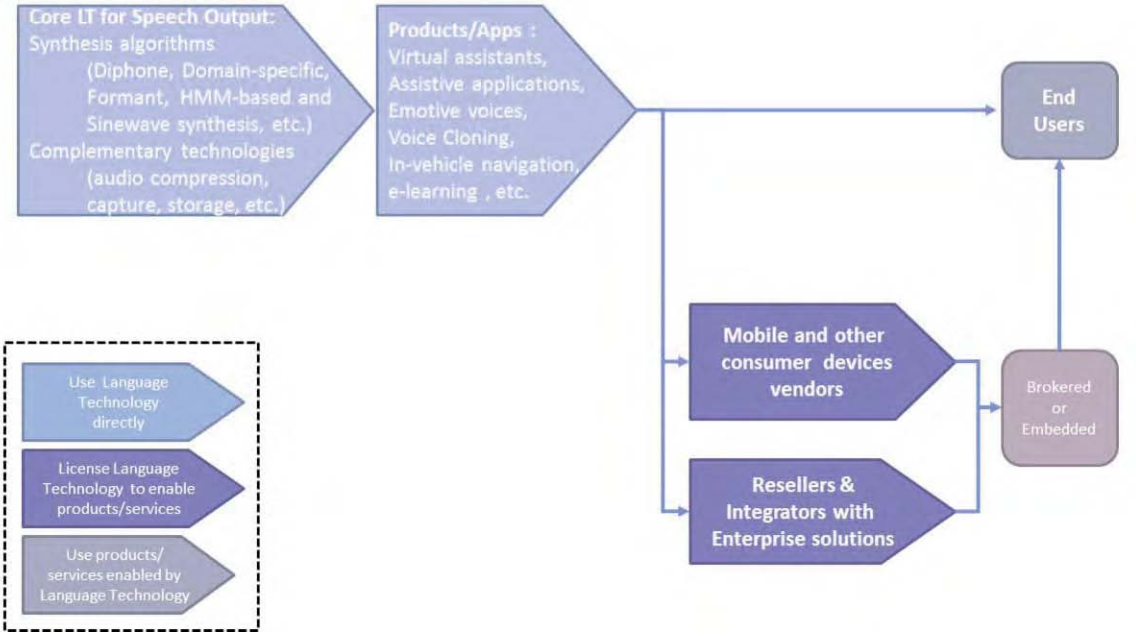


Value Chain: Speech Input / Automatic Speech Recognition
Core Tools & Techniques



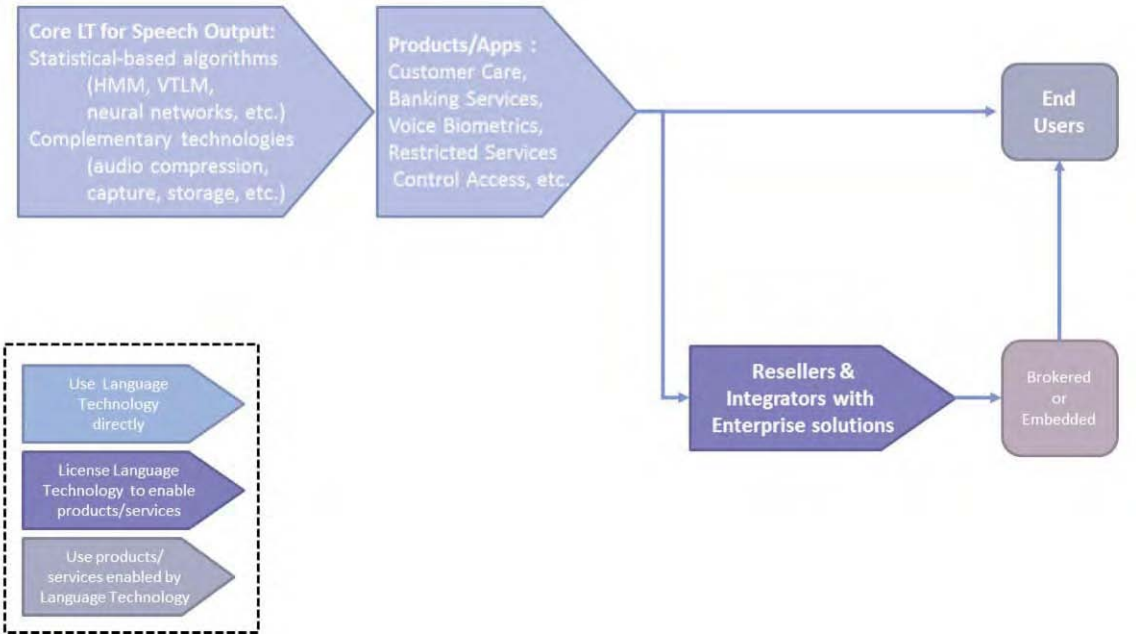
Value Chain: Speech Output / TTS

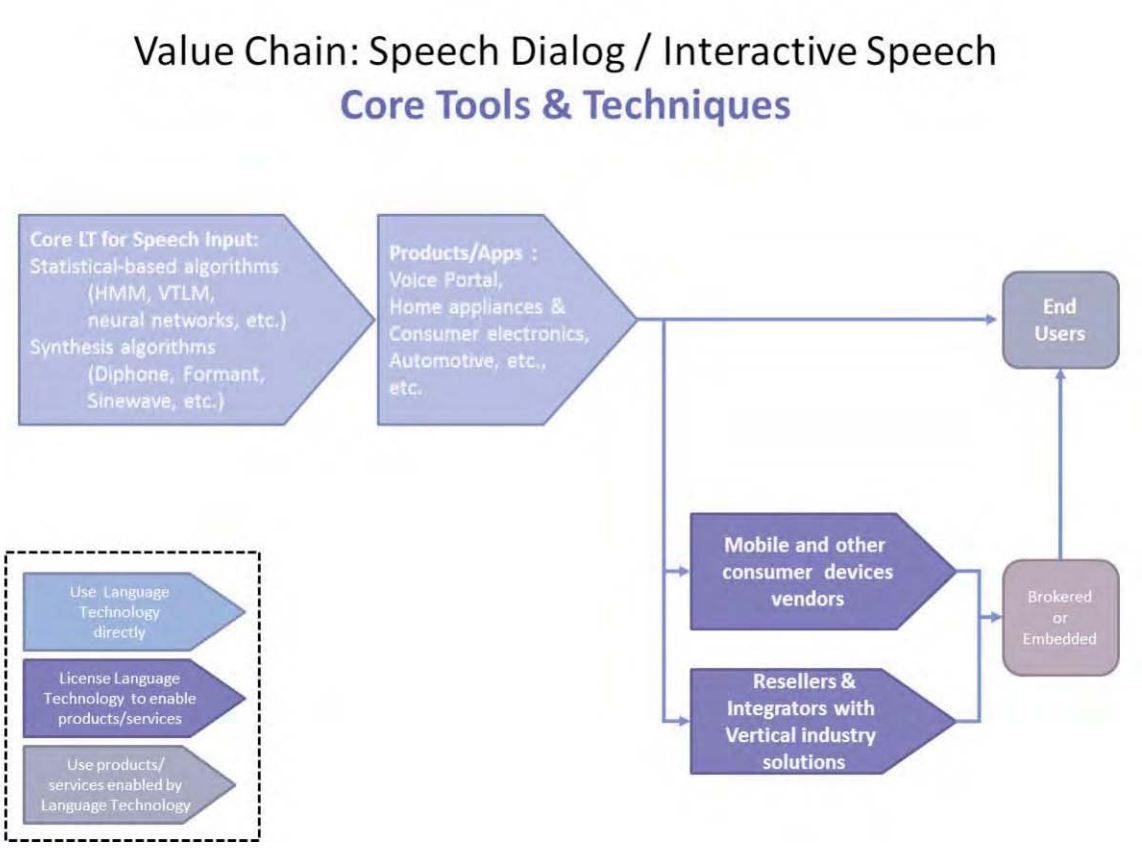
Core Tools & Techniques



Value Chain: Speech ID / Verification

Core Tools & Techniques




Value Chain: Speech Dialog / Interactive Speech
Core Tools & Techniques
Core LT for Speech Input:
Statistical-based algorithms
(HMM, VTLM,
neural networks, etc.)
Synthesis algorithms
(Diphone, Formant,
Sinewave, etc.)
Products/Apps :
Voice Portal,
Home appliances &
Consumer electronics,
Automotive, etc.,
etc.
End Users
Mobile and other consumer devices vendors
Resellers & Integrators with Vertical industry solutions
Brokered or Embedded
Use Language Technology directly
License Language Technology to enable products/services
Use products/ services enabled by Language Technology

Translation Technology Overview

4.8 Background

We can debate whether the translation industry's response to rapid globalisation and growth in content has been the right one. Has the industry made best use of technology to raise its capacity and stay profitable? Or has the content explosion marginalized an industry of artisans?

To understand the current landscape we need to take a ride into translation technology history. The onset of the cold war highlighted the need for machines that could translate with both the US and Russia wanting to keep tabs on each other. However, despite significant investment, by the late 1960s it became clear that machines could not translate anywhere near well enough. Machines and translation were rarely sighted together for the next fifteen years. A notable exception was the European Commission's use of Systran's systems for registering purposes beginning in 1976.

By the mid-1980s a few service providers in a burgeoning localisation industry began to provide translation memory tools, which enable recycling of translations, alongside glossary tools to their translators. This they felt at least left translators in control.

While these translation technology tools became more scalable and the feature set matured, the core technology for improving capacity in commercial translation services was left largely untouched for the next twenty years.

The localisation industry continued to grow rapidly, if not innovate, and many in the traditional translation industry were largely left in its wake. Many translators demonstrated reluctance to adopt translation technology for a range of reasons - some well-founded, others based on ill-informed fears of replacement.

The lack of innovation also meant low barriers to entry to the translation technology segment. Many translation services companies developed their own computer-aided translation tools, creating panoply of mediocre systems with no differentiation. While the translation services industry flourished, the technology segment remained minuscule.

In that same period the cold war ended, economic growth began shifting from English speaking to non-English speaking nations, digital communications largely replaced print, consumers and citizens became publishers *en masse*, and people all over the world got *connected* via mobile phones.

It is only in the last 5 years that we begin to see wholesale changes in the technology and business models being used for translation; for the large part change leaders have entered from outside the industry.

4.9 Technology for Professional Translation

Translation Processes and System Features

Translation technology can be broken down into several major components. Translation tools, also known as CAT (computer aided translation), enhance the productivity and consistency of human translators, and include several component technologies, including: translation memory (which enables translators to re-use and learn from previous work), translation management systems (which automate project management and publication), terminology management systems (which encourage the use of standard terms, names, and translations) and quality assurance tools.

The workflow diagram below provides an overview of the many stages that may be needed for translation and localisation. The translation technologies outlined in this section seek to optimize and where possible

automate the process.

Translation Processes & Features

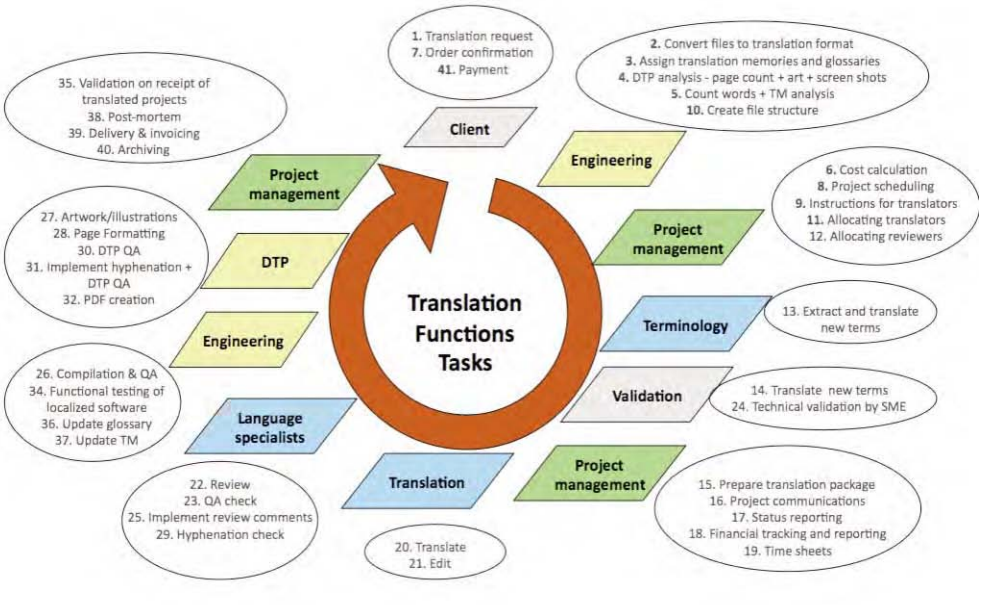


Figure 19: Translation Process

The features highlighted in this section can be found throughout the translation technology tool chain. These are best thought of, not as products, but as technologies or features that are embedded within larger systems. These features are generally utilized to increase translator productivity, improve efficiency/cost, reduce errors, and encourage the use of consistent style and terminology, all of which are important to delivering high quality output.

Translation Memory

Translation memory is a common feature in most computer aided translation (CAT) tools. It comes in two distinct forms, one of which is used to eliminate redundant work, and one which is used to assist translators in recalling similar translations from previous jobs.

Exact match translation memory simply looks for translations whose source text exactly matches a previously completed translation. This type of search/lookup is easy to do, and is not computationally intensive, so an exact match translation memory can search a corpora of billions of texts nearly instantly. This is typically used to auto-fill translations for source texts that have already been translated. For applications where there are a lot of repeating texts, this can yield significant cost savings.

Fuzzy match translation memory looks up not just exact matches, but also looks for similar source texts, and scores them by their degree of similarity to the index text. While this is conceptually simple, calculating the "distance" between texts is not trivial. At best these systems can generate an approximate "distance" between candidate texts. Because of this, fuzzy match translation memory is best used as a style and memory aid for human translators, who can pull up a list of similar texts and their translations, and then use the

relevant pieces as needed. This form of translation memory generally does not yield significant cost savings (the translator has to manually decide which texts to use and how), but is best used to reduce errors, and to encourage the use of a consistent writing style.

Both forms of translation memory are found in most translation and localisation management systems (although localisation tools have less need for fuzzy match translation memory, but do rely heavily on exact match to prevent the redundant translation of strings).

Translation memory was first found to be useful when the user documentation was updated for new versions of software. There is typically a high reuse rate in these documents.

Advanced Leveraging

Advanced leveraging, otherwise known as sub-segment analysis, is used to construct a translation using sentence fragments. This type of CAT tool tries to stitch together translations from different segments, for example by identifying translations for phrases which, in turn, appear in longer texts elsewhere. This technology is useful as a translator productivity enhancer, but like fuzzy match translation memory, cannot be fully automated. It is an intermediate approach between translation memory and machine translation.

Advanced leveraging has been offered as a feature in a small group of tools, including MultiCorpora's MultiTrans and Atril's DeJaVu for a number of years now. The main adopters have been public bodies and the finance sector. These sectors do not benefit as greatly from translation memory as their focus is not product updates. Instead advanced leveraging tools help them to efficiently reuse translations from large translation datasets, which are called corpora. By using such technology they aim to enhance productivity and terminological consistency.

There is a general trend away from large sets of documentation to smaller volume, sometimes continuous streams of publication. Given this trend there is scope for growing adoption of advanced leveraging technology for years to come.

Translation Process Management

Translation process management (TPM) refers to administrative and workflow control tools that enable a system owner or administrator to control when and how translations are done, the quality assurance/review process, etc. The details of how this is implemented vary widely from system to system. Some systems provide the administrator with the ability to set broad controls on how things are done, while others allow the administrator to control the translation process at a fine-grained (per job or task) level.

Generally speaking, TPM tools allow the system administrator to control the following:

- Decide which assets or asset classes are routed for translation
- Decide which translation resource or service to use for an asset class and target language (e.g. machine/human translation, select LSP, SLA, etc.)
- Decide whether returned translations should be auto-approved, or queued for review and post-editing
- Decide who is allowed to review/post-edit returned translations or mark translations as complete/published
- Decide who is allowed to edit or translate assets, on a per site, project or asset level (e.g. invite professional or crowd translators to a project)

Terminology Management

Terminology management systems enable users to create, translate and manage dictionaries or term glossaries on a customer, project or asset level. Term glossaries are useful for defining how a set of words, phrases or proper names should be translated (or not translated). This is used not so much to reduce translation cost, but to encourage consistent vocabulary and style, and to prevent dissonant translations of phrases that recur frequently.

While glossary terms can be auto-translated, the best practice is to recommend their use to translators, who then click on an accept button or link, to paste the recommended phrase into their translation editing environment. Translators should be able to override the recommendation because there are often situations where the translated text needs to be edited for grammatical correctness.

This is basically a required feature for any serious CAT tool, and is available in some form on almost every translation and localisation platform or tool in use today. There are also a few vendors that specialize in terminology management tools, such as Interverbum Tech.

Controlled Authoring

Controlled authoring tools, such as Acrolinx, are used before translation even begins. These tools are designed to maximize the source content's quality and consistency, and provide the following key features:

- Spellcheck and grammar correction: to catch basic mistakes during authoring
- Terminology management: to insure that technical terms are used consistently
- Brand protection: to insure that brand names and proper names are used correctly
- Edit for MT: to enable users to prepare source text for machine translation

The basic goal is to prevent authors from using inconsistent terminology, catch common errors, and generally produce standardized output that is search engine friendly, and translation ready.

Quality Assurance

Translation quality assurance relies on a combination of technology and processes to prevent errors from creeping into translation projects. The QA process starts before a project is sent out for translation, for example be sanitizing text to protect non-translatable elements, disambiguate the source text, provide comments and context, etc.

Once translation is in progress, QA is implemented in several ways at different stages of the process:

- [Prior to translator assignment]: decide which translator(s) are best match to the task, factoring in skill level, prior QA scores, availability and domain of expertise.
- [during translation] : computer aided translation (CAT) tools draw upon resources such as spellcheckers, term glossaries, and translation memory to increase productivity, catch common errors, and encourage the use of consistent style and terminology.
- [Post translation]: completed translations are generally sent to an editor or trusted reviewer to be spot checked, and post-edited as needed (or sent back to the translator(s) for additional work.
- [Post translation]: highly automated systems may send translations to one or more randomly chosen translators for a blind peer review and score. Editors only intervene to evaluate translations with an ambiguous score.
- [Post-delivery]: highly automated translation services may also provide a widget or API through which the customer can comment on, score or request re-translation of texts. This allows high volume translation applications to continually improve translations without subjecting every text to embargo, and also enables the client's customers to contribute feedback about translations (e.g. in moderated crowd

translation scenarios).

Examples of a specialized QA tools are QA Distiller and Error Spy. TAUS has developed the TAUS Dynamic Quality Framework, which documents best practices for quality evaluation, provides a method to select fit-for-purpose evaluation approaches and a set of tools to establish quality evaluation metrics and benchmarks.

Computer-Assisted Translation Systems

Computer-Aided Translation Tools

Computer aided translation (CAT) tools have been in use since the introduction of personal computers. Their primary purpose is to improve translator productivity and accuracy by providing tools such as document editors, glossaries, and translation memory, in a single integrated environment or workbench. These tools have evolved along with the computing and networking industries, first as stand-alone tools to be used on a single computer, to client-server tools to be used on a company network, to web based tools where the service and tools are delivered via the Internet.

Client/Server based CAT tools

Client/server computer aided translation tools (CAT) have been in use for approaching twenty years, and are well entrenched within enterprise translation and localisation departments. As a general rule, these tools were developed before the web and cloud based services caught on. These products represent a mature technology and include nearly every feature users expect from a translation platform.

The typical implementation consists of a central server, usually hosted on the customer premises, but not always, which in turn interacts with client software, typically Windows based that is installed on employee and translator machines. This was the standard configuration for most types of enterprise software until about ten years ago, when vendors began migrating to web based environments. Newer systems that have been built from the ground up since then largely bypass this model.

These systems, because of their age, offer a complete set of component technologies including translation memory, glossaries, document editing and spellchecking tools, and in some cases, project management tools. While they offer a complete feature set, many of these tools have significant disadvantages relative to new tools, among them:

- Lacks of cross-platform support, most of them are heavily focused on Windows, a problem for organisations that use other operating systems.
- Lack of native mobile support. Translators and reviewers increasingly demand mobile applications so they can work where and when they want to.
- Steep learning curve. These systems were designed primarily for translation and localisation professionals, and can be intimidating to newcomers. They also have user interfaces that are pretty dated by today's standards.

High IT costs, due to the need to maintain the servers, update software and install client software on a potentially large number of machines.

In general, companies that are not already vested in this class of tool are moving to web oriented systems. Established vendors are re-tooling their client server solutions as cloud based products as well. On the other hand, companies and government agencies that have stringent security requirements will probably want to continue hosting these services in their own data centres, or at least on computers they directly control.

Web based CAT tools

Web based CAT tools are an important new class of translation tool, and are likely to replace traditional client/server tools (e.g. Windows products) in many use scenarios for several reasons:

Cross-platform access via Windows, Mac, Linux and other operating systems.

Ability to support users on mobile devices either via native mobile apps or HTML5/JavaScript, Java is also an attractive language for building highly capable, cross platform apps.

Support for agile development processes, where server side software is continually improved without forcing labour intensive client software upgrades

Cloud based asset management, translation memory, and other features to centralize project and asset management. Users can work on a job from different devices at different times without losing work

SaaS business models, with per user or volume based licenses enable customers to scale up and down as needed

Web based translation agencies typically develop their own translation workbench and CAT tools, with varying levels of sophistication depending on the company and its customer base. Examples of companies in this category include: Gengo, Straker Translations, Fox Translate, Elanex (aka ExpressIT), and One Hour Translations. These companies provide clients with web based and API based ordering and project management interfaces (front end), and provide translators with an entirely web based editing and CAT environment (back end).

The Google Translator Toolkit (GTT), which is offered as a standalone CAT tool for translators, is an interesting example of how powerful web based tools can be. It is fully integrated with Google Translate, so translators can pre-translate and post-edit texts using machine translation followed by human review. GTT was also recently integrated with YouTube to support user-generated captions.

Some client/server CAT tools, such as those provided by Across Systems, have web accessible interfaces. However in these systems, web access tends to be a second-class citizen compared to the native client app, which is typically written for Microsoft Windows. This is often explained by the product's history and the timing of its development. Most of the client/server tools trace their origins to the late 90s and early 00s, before SaaS had caught on, and before non-Windows operating systems such as Mac OS X, Android and iOS became commonplace.

Mobile translation tools

Mobile translation tools are currently a novelty in the computer aided translation space, but will become a material requirement in most translation management systems because translators, especially in developing economies, use mobile devices as their primary means of accessing the Internet. Tools that do not support mobile access, either via HTML5/JavaScript, or via native iOS/Android applications, will find it increasingly difficult to bring these translators into their workforce.

Hong Kong based OneSky (www.oneskyapp.com) offers an example of how translation can be done via mobile devices. Their service is specifically geared to mobile app localisation, so it fits well with a mobile translation-editing tool. The primary challenge in making CAT environments accessible via mobile devices is to deal with the restrictive display and user input interfaces common to these devices. Typically this means constraining the type of work mobile translators can do, for example, by having them work on shorter texts and documents. This limitation is less of an issue for tablet type devices since their displays are similar in size and resolution to a PC or laptop display.

Standalone utilities

Stand-alone CAT tools are designed for use by independent translators who do a lot of their work offline. Their primary advantage is the ability to work independently of Internet connectivity, so a translator can download a project, copy it into their CAT tool, complete the task, and then check it back into the translation management system when they return.

These tools are well entrenched and heavily used by professional translators, and will enjoy continued success in the marketplace, especially if they are upgraded to inter-operate with cloud based translation management systems. These tools combine a number of functions including: document editing, translation memory, glossary, and spellchecking.

Examples of these tools include MemoQ and SDL/Trados.

Translation Management Systems

Translation Management Systems

Translation management systems enable users to centralize and control their translation workflow, as well as to manage the assets being translated (documents, videos, and other content). These systems range from simple, purpose-built TMS solutions, such as Word Press translation tools, to complete supply chain management systems that, in addition to managing translation workflow, also automate the process of interacting with external translation service providers.

Translation management systems provide an additional layer of process management and automation. They are used to manage how translation work is assigned to different participants, and to import/export documents and their translations to and from other systems, such as content management systems.

Document TMS Systems

Translation management systems (TMS) provide a similar function as content management systems. These systems enable operators to:

- Centrally manage resources to be translated (documents, video captions, localisation files, etc.)
- Control which languages each project or resource is to be translated to
- Invite and assign translators, editors and reviews to each project/language
- Define workflows for translation, for example whether translations are auto-approved on receipt (from trusted sources) or must be independently reviewed
- Receive quality feedback and defect reports, and automatically route these to the appropriate translators and project managers
- Use machine translation (for pre-translation), translation memory, term glossary and advanced leveraging to re-use previously completed translations, to boost efficiency, and to increase quality and consistency
- Export completed translations to external systems (e.g. content management system, e-commerce platform, etc.)

These systems are fairly mature products in terms of functionality. The major shift underway today is a migration from customer premise based client/server systems (where the customer deploys and manages their TMS system) to cloud based SaaS (Software-as-a-Service) offerings. Virtually all translation vendors starting up in the last few years offer cloud-based solutions. These include companies like MemSource and XTM International. Major translation technology vendors such as SDL have been retooling their product offerings across the board as SaaS services.

Localisation TMS

Localisation management systems are a specialized form of translation management system that are used primarily or exclusively for software localisation. Localisation has a unique set of requirements compared to document translation, so this differentiation makes sense. For example, when localizing software, it is very important to pay attention to word length, as this affects the layout of a user interface. For example, German, with its longer average word length and abundance of compound words, can easily break the layout of a webpage or application. Localisation management tools are designed with these issues in mind, whereas document oriented tools are less concerned with issues like this.

There has been a proliferation of vendors that offer localisation as a turnkey SaaS offering over the past couple of years. Examples include companies like GetLocalisation, OneSky, Smartling, Tethras, and Transifex. These companies provide turnkey, cloud based services that enable customers to upload their prompt files and other assets, control how they will be translated and to what languages, and if needed, order bulk translations from a professional language service provider that has integrated with the service. The customer can generally combine machine, crowd (bring your own translators) and professional (outsourced) translation.

Translation Memory & Terminology Management

Translation memory and terminology glossaries are generally implemented as a feature within larger systems. These services are used to improve translator efficiency and accuracy. Term glossaries are translation dictionaries that are built from frequently occurring words or phrases, for example technical terms, brand names, etc. These dictionaries are used to pre-translate recurring words and phrases, and to assist translators in using consistent translations, and also to avoid translating items that should be left as is, such as brand names.

Translation memory is a record of previously created human translations. Typically, this is used to display similar source texts and their translations, both as a memory aid for translators, and as a style guide. Translation memory can also be used to boost translator efficiency, for example by enabling them to make small changes to previously created texts (see section 3.1.3.1 for information on translation memory).

Both services are typically integrated into a translation management system, but there are examples of stand-alone translation memory services, including TAUS Data repository (estimated to be the largest), or MemSource, a commercial translation memory application that is available both as a self-hosted product and as a SaaS offering.

QA Tools & Processes

Translation management systems typically employ a number of tools and processes to maximize quality at different stages of the project.

Technology Providers

A number of translation process management platforms are commonly integrated into the translation supply chain, including translation management systems, localisation management systems, captioning and subtitling platforms, stand-alone translation memory services and live interpretation systems. Each of these directly manages or interacts with the translation workflow, each with its own use cases and special requirements.

Translation management systems

Purpose built TMS systems are often embedded within other products, such as content management systems. While these do not provide full supply chain management, they provide most of the functions needed to manage translation workflow for typical users. Examples include the Translation Management Tool, by MD Systems, for the Drupal content management system, and Word Press ML (wpml.org), a multilingual translation management tool for the popular Word Press platform.

Stand-alone TMS systems, on the other hand, are used to manage translation workflow for many different types of assets, from documents, to websites. Since different types of content require different workflows, and often different service providers, an enterprise TMS enables operators to manage not just the translation process, but also the vendor supply chain. Examples of vendors in this category include SDL, Across Systems, and Lingotek.

Localisation management systems

Localisation management systems are a special type of translation management system, and focus on the tasks and challenges that are unique to software localisation. There has been a proliferation of SaaS based services in this area, especially for mobile app localisation for iOS and Android platforms. These services enable users to upload their application prompt catalogues using a variety of localisation file formats, and manage them and their translations via a centralized repository. Newer services borrow from collaborative development platforms like Github to support agile localisation, where localisations are continually refined and deployed with incremental upgrades.

Examples of these services include GetLocalisation, Onesky, Tethras, Transfluent, and Transifex, all of which are accessible to both small and large companies. On the high end of the market, companies like Moravia Worldwide and Welocalize provide software localisation and testing to large software companies such as Microsoft and Oracle.

Translation memory (standalone)

Translation memory, while it is generally integrated into translation management systems, is also available as a stand-alone, cloud based service. TAUS provides a translation memory that pools submissions from a large number of sources; the translation memory is accessible via a web API, and currently contains over 50 billion words spanning over 2200 language combinations.

Most translation management systems and CAT tools provide some form of translation memory. Nearly all provide exact match translation memory, to avoid re-translation of repeating texts, as well as terminology glossaries or dictionaries, to promote consistent translation. Most also provide fuzzy match translation memory, as a memory aid and style guide for translators. If a TMS or CAT tool does not provide at least basic translation memory, this is a serious deficiency that should rule out the use of that product.

Audio/video captioning systems

Audio/video captioning and subtitling systems have a unique set of requirements that differ from text based content. Because of this, vendors tend to specialize in this area. Captioning systems must deal with a number of technical issues, including:

- Support for a wide variety of video file/stream formats
- Tools to transcribe audio tracks to create source language captions
- Tools to translate captions into one or more languages
- Ability to time code captions so they appear at the right time during playback

- Tools to review and post-edit translations

Several vendors, including dotSub, Amara, and Viki specialize in video captioning and subtitling. dotSub and Amara provide tools that enable video content producers to generate captions using a combination of crowdsourced translations, and optional professional translators. Viki, meanwhile, is a purely crowd based system, and has created a vibrant translation community (several million active users) around captioning video programs from around the world.

Interpretation systems

Interpretation systems and services enable users to have telephone calls translated in real time, using either sequential or simultaneous interpretation (simultaneous interpretation enables both parties to converse naturally without pausing for an interpreter to repeat what they said). Like video captioning, this is also a specialist market, with a different set of dominant vendors compared to other sectors of the translation industry.

These services are typically accessed via a telephone or voice over IP (VoIP) call to the interpretation service which, in turn, bridges on or more interpreters onto the call.

The leading providers in this category include: Language Services Associates, Language Line.

Business models

Licensed

Until recently, most translation tools were sold as licensed software, typically priced per user/seat. This is how most enterprise software was sold prior to the transition to cloud/SaaS based services. Even now, many enterprise software vendors sell their software this way. Cloud based offerings will, however, force many vendors to rethink their pricing model because these offerings enable customers to scale their tools budget up and down in response to usage, and to avoid committing to expensive upfront purchases.

Cloud/SaaS

Cloud based (SaaS) offerings are becoming more common in the translation industry, and are the overwhelming favourite among new companies. These services are typically priced as monthly subscriptions, with the fee based on any number of factors including:

- Number of registered or active users
- Number of target languages supported
- Number of projects or assets stored on system
- Number of words hosted on the system
- Translation volume per month

In all cases, customers are largely able to avoid up front commitments, and can also evaluate a product at relatively low cost and risk prior to scaling up to production use. Vendors that fail to offer a viable SaaS option will risk losing market share to emerging companies and services that do.

Translation services

Another business model employed by some tools and platforms is a bundling strategy, where the software or platform is offered for free, while the vendor charges for professional translations brokered through their system. Cloudwords, for example, operates a hosted marketplace and translation/project management serv-

ice that enables customers to select from many translation agencies, and to centrally manage their projects. While the Cloudwords service is not free, it is quite inexpensive. Cloudwords, in turn, charges a commission for the projects brokered through its platform.

Many localisation service providers follow a similar model, and offer their hosted platform for an inexpensive monthly fee, then make the bulk of their money by selling professional translations on a per word basis. Translation agencies, in turn, make most or all of their money by selling translation.

Channels and platforms

These translation services are offered through both direct and channel partner systems, depending on the amount of automation and system integration required. For example, a translation agency that offers a self-service web translation tool for ordering translations for Word documents will typically sell direct to end users. On the other hand, systems that require a lot of automation will often have translation built in.

Examples of integrated solutions include:

- Content management systems that have translation management built in (e.g. Drupal)
- Captioning and subtitling services that support translations as part of the captioning process (e.g. 3PlayMedia, Amara)
- Multilingual e-commerce systems that automatically translate source language content as new products are added
- Custom applications built around a language service provider's system or API

The advantage of integrated solutions is that they greatly reduce the amount of work the customer needs to do to utilize translation (in some cases, they automatically request translations from LSPs behind the scenes so no administrative work is required of the users). Their main disadvantage is that these integrations are difficult and expensive to do, so LSPs and third party solution providers are slow to create integrated tools for new markets and applications.

Demand for Professional Translation Technology

The demand for translation technology comes from several sources: individual translators, language service providers (translation agencies), and publishers/content producers. Individual translators and agencies are typically looking for computer aided translation tools, so they can work more efficiently. Larger agencies will also invest in translation management tools (or use their customer's translation management tools, depending on the situation). Publishers and content producers, on the other hand, are generally looking for process automation, and are less concerned about the details of how translators do their work (this is often done by an outsourced agency).

Individual translators

Individual translators participate in the supply chain in three main ways: via translation agencies, translation marketplaces and direct-to-customer relationships.

Traditionally individual translators would interact with customers via language service providers (translation agencies). LSPs provide several services to translators: customer acquisition (sales), project management, and administrative support (billing, collections, etc.). They are still a dominant channel translators go through, but new technologies are enabling customers to automate more of the process, and in some cases build direct relationships with translators. Google's G-Community is an example of such disintermediation.

Translation marketplaces, such as Cloudwords, enable translation buyers to request competitive quotes,

place orders, and manage their projects via a SaaS offering. This is attractive for companies that have complex translation needs that require the use of multiple agencies. Simpler marketplaces, such as ProZ and Translator's Cafe, enable users to request competitive quotes, but do not provide integrated project management tools.

Translators will often decide to work directly with their favourite clients. New tools make it easier for people who are not translation industry professionals to assume many project management tasks, and therefore to work directly with a hand-picked crew of translators.

Language service providers (translation agencies)

Language service providers typically interact with the supply chain via more direct means, since they serve in an intermediary role for individual translators. They typically focus on direct sales, and to a lesser extent reach customers through translation marketplaces.

One of the services they provide is outbound sales and account management to larger corporate clients; something individual translators are not necessarily skilled at or financially equipped to do.

There are a significant number of translation agencies that work primarily as subcontractors for larger service providers. These typically focus on a small number of languages or specialize in specific sectors. They undertake the production, while the larger firms maintain the relationship with customers.

To a lesser extent, translation agencies will access customers through translation marketplaces like ProZ and Cloudwords. The larger agencies generally don't feel they need to participate in these communities, except to recruit translators, since they have well developed outbound sounds and account management capabilities. Small and mid-sized agencies, which tend to specialize by language, services or domain of expertise, do actively participate in these.

Content producers and publishers

Publisher organisations, or content producers, are generally concerned with delivering translations once completed, and generate requests for translation for new content, whether it is a website article, travel listing or other item. They will typically use a content management system or e-commerce platform that has been integrated with the translation supply chain to implement an automated or semi-automated workflow.

There are many different types of publisher organisations, since anyone who produces content in print or digital form can be considered a content producer. Different types of content producers will typically use different types of content management and e-commerce systems. For example, the online version of a newsmagazine might use a content management system like Drupal to host its website, while a flight booking service would use a completely different system.

These customers are typically looking to integrate their existing publishing and e-commerce platforms with translation resources that automate the process of detecting new content, queuing it for translation, and then storing/displaying the translated content when needed. Each type of content producer has different requirement where translation is concerned. For example, an e-commerce site may have tens of thousands of product listings that need to be translated and kept in sync, and may be concerned more with search engine visibility than the highest possible translation quality. A magazine publisher, on the other hand, will have fewer but longer texts needing translation, and will be much more concerned with output quality.

In most of these cases, the customer's primary concern is system integration between their content delivery platform and their translation management system.

Trends for Professional Translation Technology

There are several trends affecting this sector: the transition from desktop to client/server and then to cloud based (SaaS) services, the translation from licensed software to subscription (SaaS) business models, and the trend toward integration solutions (for example, translation management systems that are integrated with web publishing or CMS platforms).

Computer aided translation tools have gone through three distinct phases of development since their inception. First generation tools were largely designed to be used as stand-alone applications, largely due to the fact that network connectivity was limited at the time of their development, and due to the "low tech" history of the translation industry, as it predates the computing industry by far. Some of these tools, such as SDL/Trados, continue to thrive, and can be upgraded to interoperate with new cloud based translation platforms.

As companies deployed local and wide area networks, and then later connected to the Internet, client/server CAT tools and systems became commonplace. These tools were developed during the 1990s, and mirror the technologies available at the time. These systems enabled their operators to centralize the storage and management of translation assets and project management, which enabled significant productivity gains in translation workflow. These tools were largely Windows based, and most utilized proprietary communication protocols for client/server communication.

The latest generation of CAT tools are largely cloud based SaaS (Software-as-a-Service) offerings that eliminate the need for the operator to own and manage on premise equipment. These services, many of them developed by emerging companies such as XTM International and Smartling, are also built from the ground up around web based technologies, and are accessible from virtually any operating system or device. This is a significant advantage over second generation tools, as they are largely tied to the Windows operating system, and are not easily accessed via the web or mobile devices.

From licensing to professional services and SaaS

Translation platforms have gone through a similar evolution from per-seat perpetual licenses to professional services and most recently, to a SaaS (Software-as-a-Service) pricing model.

Early products were sold as shrink-wrapped software, essentially a perpetual license based on the number of installed seats. Most software was packaged this way at that time, so this approach made sense in its day. Since then, small and large translation technology companies, from start-ups to industry leaders like SDL, are shifting to SaaS pricing models.

These contracts are typically priced as monthly or annual subscriptions, with rates linked to one of several variables, including: the number of words or segments stored in the system, translation volume, number of target languages, or number of active users. Transifex, for example, prices its localisation management service based on the number of words stored in its central repository. The pricing formula varies from company to company, but is generally linked to usage, storage or translation volume, so each customer pays based on their usage of the service.

Integration with content management systems

Leading translation management systems, such as SDL World Server, are integrated with popular content management systems, including Drupal, Sharepoint, and others. This has become a requirement as companies use CMS platforms to manage their source content, editorial and publishing workflows. Integrating with these systems enables translation companies from re-inventing tools that have already been done well by other companies.

Middleware companies, such as Clay Tablet Systems, offer services that connect a variety of translation management systems with the leading content management systems. They have integrated with most of the popular corporate CMS platforms. This is an attractive option for translation technology companies, who can integrate once with Clay Tablet's system, and then automatically inherit support for every CMS platform they have integrated with.

4.10 Technology for Automatic Machine Translation

Approaches to Machine Translation

Rules Based

Rules based machine translation, developed several decades ago, was the first practical approach to automatic translation. This type of translation engine works by parsing a source sentence, analysing its structure (for example, determining which words are used as verbs or nouns), and then converting this into an intermediate, machine-readable code. This is, in turn, transformed into the target language.

The advantage of rules based translation is that a sufficiently sophisticated translation engine can translate a wide range of texts without having been trained with a large number of examples, as in statistical machine translation. The disadvantage is that it is necessary to build custom parsing software and dictionaries for each language pair, and that it is quite "brittle". Rules based translation engines don't deal very well with slang or metaphorical texts, for example. For this reason, rules based translation has largely been replaced by statistical machine translation or hybrid systems, though it is useful for less common language pairs (where there are often not enough parallel texts to train a statistical machine translation engine).

Products and Practitioners

The two primary providers in this category are Systran (commercial software) and Apertium (open source). Systran has been in operation for decades, and was a pioneer in web translation (their translation engine powered the Babelfish web translation service back in the 1990s). They cover most major language pairs, and most recently Systran has released a hybrid rules/statistical translation engine to upgrade their product line.

Apertium is an open source project sponsored by Universitat d'Alacant in Spain. They have developed an open source rules based translation engine that enables users to create custom translation engines for any language pair. This solves an important problem for rules based translation engines, as commercial vendors do not invest in development for less common language pairs, such as Spanish ← → Catalan. Developing a custom engine is a large task, as it requires the development of dictionaries, parsing rules, etc., and the involvement of linguists who are experts in the source and target languages.

Example Based

Example based machine translation is similar to statistical machine translation, as it uses a large volume of parallel texts (source segments and their translations) to train the system. The logic behind example based translation is that it treats sentences as a collection of often repeated phrases that can be translated independently and then combined to form a translated sentence.

The problem with this approach is that you need a very large corpus of phrases and their translations. This requires a lot of data, and also requires the phrases and translations to be perfectly aligned, which typically requires manual effort, whereas statistical machine translation systems can be trained in a fully automated process.

Example based machine translation has not been widely deployed as a commercial service. However, there

is an open source platform, Cunei, which enables developers to build their own example based MT engine (similar to the Apertium platform for rules based translation). Most translation engines in development and commercial use today are statistical or hybrid systems.

Products and Practitioners

Example based machine translation is not available as a stand-alone commercial product or service, but you can find two open source projects: Cunei and Marclator. These are open source projects, and are only suitable for expert software developers and system administrators, as they are not turnkey solutions designed to be touched by end users. They are great for experimental use, but if you are looking for user-ready platform, look for statistical machine translation platforms.

Statistical

Statistical machine translation is currently the gold standard for machine translation, and is used by the most popular translation engines in use today. The process works by training the translation engine with a very large volume of parallel texts (source texts and their translations), as well as monolingual corpora. The system looks for statistical correlations between source texts and translations, both for an entire segment, but also smaller phrases, or N-grams, within each segment. It then generates confidence scores for how likely it is that a given source text will map to a translation. The translation engine itself has no notion of rules or grammar.

The key advantage of statistical machine translation is that it eliminates the need to handcraft a translation engine for each language pair, as is the case with rules based translation. Provided you have a large enough collection of texts, you can train a generic translation engine for any language pair.

The main disadvantage of statistical machine translation is that it fails when it is presented texts that are not similar to material in the training corpora. For example, a translation engine that was trained using technical texts will have a difficult time translating texts written in casual style. Therefore, it is important to train the engine with texts that are similar to the material you will be translating on an on-going basis. Even with large and suitable training corpora, statistical machine translation does not generally produce publication quality text. It frequently translates items out of context or uses the wrong word order. However, it generally translates well enough that it is suitable for comprehension. If you need publication quality translation, you'll want to have some sort of human review and post-edit process, which many commercial MT engines provide as an option.

Products and Practitioners

Many companies offer statistical machine translation, among the key providers are:

- AsiaOnline (customized engine): AsiaOnline provides statistical machine translation as a hosted service for corporate clients. They distinguish themselves by providing custom MT engine that have been trained with the customer's texts, in addition to standard sources of training corpora.

- BeGlobal (SDL): BeGlobal is SDL's machine translation offering. Derived from its acquisition of Language Weaver several years ago, BeGlobal enables users to combine machine translation with professional translation and post-editing. A common workflow is to machine translate a text on the first pass, and then have human translators and editors review the output and correct it. These corrections can, in turn, be fed back into the translation memory to further train it.

- Google Translate (free): Google offers a free web translation service that is based on its own translation engine and research. The service can translate to and from over 50 languages, and is regarded as a benchmark for translation quality for non-specialized translation engines.

- Microsoft Bing Translator (free): Microsoft also offers a free web translation service that is similar to Google Translate, but also includes many options for people to score and post-edit translations using an interactive (WYSIWG) editing tool. This is an especially interesting option for companies that have a large community of readers who can be tapped to edit and improve translations to benefit other users.
- Moses (Open Source): Moses is an open source statistical machine translation engine. It is widely used within the industry to build customized MT engines. Because it is open source, people wishing to develop a custom engine can focus on obtaining the training corpora rather than writing their own statistical machine translation engine (a difficult task that is beyond the abilities of most developers).

- Tilde (customized engine): Tilde is a Latvia based translation technology company that focuses on building custom translation engine for poorly resourced languages, and for specific industries.

Hybrid

Hybrid translation engines combine elements from rules based and statistical machine translation to leverage the strengths of each approach. This is an area of ongoing development, so we expect many systems to evolve into hybrid platforms.

There are two main categories of hybrid systems: rules based engines that use statistical translation for post processing and clean-up, and statistical systems that are guided by rules based engines.

In the first case, the text is translated first by a rules based translation engine. This translation is then processed by a statistical machine translation engine which corrects errors made by the rules based engine, or replaces the text entirely if needed. In the second case, the rules based translation engine does not translate the text but assists the statistical translation engine by inserting metadata (e.g. noun/verb/adjective, present/past tense, etc.).

Products and Practitioners

Several companies offer hybrid translation platforms, mostly focused on the enterprise market, among them:

- LinguaSys: they have developed Carabao, a hybrid translation engine that targets the enterprise market.

- Systran: Systran has been developing machine translation software for 40 years, and has upgraded its tools to combine statistical and rules based translation.

Technology Providers

Types of Providers

There are a wide variety of service providers in the machine translation space. These include:

- Consumer/web translation services, such as Babelfish, Google Translate and Microsoft Translator. These services are trained with general purpose corpora and offer decent translation for comprehension, but not publication quality output. One exception is Babelfish, which provides an option to request professional translation.

- Custom/adapted machine translation for specific language pairs or subject matter (domain of expertise). These providers typically help clients build domain specific translation engines using their own training corpora (enterprise customers often have very large translation memories which can be used for this purpose). XXXXX are examples of this.

- Hybrid translation engines. These providers combine rules based and statistical translation technology to develop adapted systems. Systran is a good example of this type of provider.

- Human/machine translation engines. These typically use machine translation for a first pass, and then have human translators review the proposed translations and edit or replace them as needed. SDL BeGlobal is a good example of this type of engine. Microsoft Bing Translator also enables this type of workflow, although they use a more ad hoc approach that's designed for crowd translation.

Business Models

Machine translation is typically offered as Software-as-a-Service, although a few vendors like Systran also offer licensed products designed for standalone use. Machine translation is a memory and CPU intensive application, and generally requires fairly high-end hardware to perform well. Some systems, like Moses, are also difficult to administer. SaaS offerings enable customers to offload upfront capital costs and system administration to their service provider.

These services are generally priced based on the volume of material being translated. Pricing models may be per word or character, or split into tiers that correlate fairly directly to word count. For custom or hybrid systems, there is typically a baseline monthly or annual fee that accounts for the cost of hardware dedicated to each customer.

Channels and Platforms

Machine translation services for consumers (e.g. web translation) are generally marketed direct to end users via sites such as Google Translate and Babelfish. While Google Translate has a dominant position in this market, there is plenty of room for companies that specialize in other language pairs to enter this space. This is particularly true of countries that have a different ecosystem of service providers, for example Yandex in Russia or Baidu in China both of whom are dominant search and services providers in those countries.

Corporate translation services are typically offered to customers via a combination of direct, stand-alone offerings, and integrated solutions that are part of a larger translation toolset. Direct offerings include services like Systran or PangeaMT, which can be used as an independent service, with or without a translation management system. SDL, on the other hand, promotes its BeGlobal product as part of a translation supply chain management system, and have done a lot of work to integrate it into other tools such as their translation management system. This allows them to sell machine translation as an option in a larger suite of services. Both approaches require the vendor to have an outbound sales capability since these clients take a long time to cultivate, and also often need time to transition off of legacy systems.

Demand for Machine Translation

Language Service Providers

Language service providers are increasingly using machine translation as part of their process. Though initially resistant to using it at all, they need it to serve customers who need to translate a lot of material very cheaply and very quickly. Machine translation can be offered as part of a tiered service where the customer can decide on a per request basis, what level of quality they need for a particular item.

Tokyo based Gengo, for example, offers a highly automated translation service for software developers. Their service is accessed via a web services API, so the process of requesting and retrieving translations can be completely automated, even if the translations are done behind the scenes by humans. They offer several different levels of translation, including a free machine translation option (powered by Microsoft Bing Translator). We expect to see offerings like this become fairly standard, as they allow the client to decide on a case-by-case basis how much they are willing to spend on a given task.

The type of machine translation they need varies depending on the type of clients they have. Gengo typically focuses on fairly generic texts, so consumer/web translation engines are suitable for them. LSPs that work with less common languages or domains of expertise will probably want to use a custom MT engine such as LetsMT or AsiaOnline.

Consumer/Individuals Direct

Google Translate and Microsoft Translator (Bing Translator) are the two dominant machine translation services for consumers in the US and Europe. In addition to machine translation, there is also a well-developed market for web based professional translation, where the customer uploads a Word document, PDF file, etc. to have it translated by professional translators. Several companies including Gengo, One Hour Translation, expressIT (Elanex), Straker Translations and others offer some version of this type of service.

We expect hybrid translation service to become a standard offering, as it enables customers to use free machine translation when they need to comprehend the contents of a document, but don't need to publish it or share it with customers. By offering the ability to easily switch over to paid, professional translation, as Babelfish does, these providers can offer a convenient all-in-one solution for day to day translation requests.

SME and Enterprise Direct

Some companies, such as SDL BeGlobal, AsiaOnline, and others offer products that can be sold direct to customers, enterprise clients in particular. They frequently have a very large volume of material to be translated, and cannot afford or wait for professional translation. The companies best suited to sell direct have a good outbound sales capability, as the direct to enterprise sales channel has a slow sales cycle (months or 1-2 years).

In this scenario, the machine translation platform can be used either as a standalone service, or can be integrated into other systems where automation is a requirement. If the customer has a fairly standard workflow (for example, to upload documents for translation, and then post-edit as needed), they can often use the built in toolset provided with the translation engine. If the customer needs to integrate machine translation into a highly automated system, for example an e-commerce server with tens of thousands of product SKUs, they'll probably need to do some system integration work. Nearly all of the translation engines we've seen provide some sort of web services API, so they can be integrated into external systems in a relatively straightforward way.

Government/Institutions

Government is an important market for machine translation technology, especially investigative and intelligence agencies, which use it to sift through vast amounts of source material before it is reviewed by analysts. Machine translation is a vital tool in making information visible to analysts. This is an example of a market that will be well served by hybrid translation platforms, where machine translation is used for a first pass. The machine translated documents are then fed into automated and human assisted search tools to flag potentially interesting documents, which are then queued for professional translation, and then for review by analysts and specialists.

The downside of this market is that the procurement process is quite arduous. Some governments also impose considerable security requirements that commercial clients are less concerned with. The decision to pursue this market is a strategic one, as it requires considerable investment in ancillary services, as well as a sales force that is experienced at securing public sector accounts, and military/intelligence agencies in particular.

Trends for Machine Translation

Custom and Adapted MT Engines

Statistical machine translation software is generic, meaning the same translation engine can be used for any number of language pairs or specialized applications. There is no need to handcraft the translation software as there often is with rules based translation. There is a market need for customized, or adapted, translation engines that are trained with corpora for a specific language pair, or for a particular industry or domain of expertise. Doing so results in higher quality and more consistent translation for that application.

A number of companies are focusing on language or domain specific translation engines as their speciality, among them: SDL/BeGlobal. We expect this to be an area of ongoing development, along with hybrid machine/human translation platforms.

Open Source MT Technology

Open source has already had a major impact on machine translation. There are mature and actively used open source platforms for all types of machine translation. Apertium enables people to build their own rules based translation engines. Moses is widely used by people building custom statistical machine translation systems.

Open source is important not so much to end users, but to language service providers who want to develop customized or adapted translation engines. Because they don't have to build or support the underlying translation engine, they are freed to focus primarily on compiling training data needed to build a client's system. Thus its primary role, in the context of MT, is to reduce the R&D costs that these companies would otherwise incur in bringing their products to market.

Data Sharing for MT

Public translation memories will play an important role in the improvement of machine translation because they can be used to generate high quality training corpora. This will also reduce development costs for companies because they can re-use an ever growing baseline corpora that many parties feed into, and can focus on collecting the high value information that is specific to their client's project. The ability to share translation memories is also important for poorly resourced languages where there are often smaller batches of translations stored in many different locations, typically language service providers. If these can be pooled and shared, this will make it much easier for companies to create high quality translation engines for secondary language pairs.

Human/Machine Translation

Another important trend is the growing use of machine translation for a first pass, with human translators (users, professional translators or both) providing feedback, suggested translations, or direct post-edits. This approach is becoming popular because it is often unknown if a particular document will be read by enough people to justify the cost of professional translation. For example, a document might be translated by machine, but then when traffic reaches a defined threshold, would be sent to human translators and editors for further review and post-editing. We expect this type of cost optimisation workflow to become popular, especially with web and mobile content producers who have to generate fast and low-cost translations. Also, if customers know that content is being machine translated, and only later cleaned up by people, their expectation of quality is markedly different.

SDL BeGlobal is a good example of this type of integration. Their machine translation engine, which can also be trained with custom corpora, is fully integrated into SDL's Translation Management System, and can be used to generate first draft texts that are, in turn, processed and post-edited by staff translators or outsourced workers as needed.

From Licensing to Professional Services

Across the board, vendors are migrating to a Software-as-a-Service business model. Few companies want to

pay up front for site licenses anymore, and would prefer to pay based on usage. Machine translation lends itself nicely to this business model, as translations can be billed on a per-word or per-character basis. Google, for example, charges $20 per million characters for the use of their Translate API (version 2.0). Pricing models vary. Some vendors charge on a per word basis, others per language, but in general expect some combination of the two for most offerings.

MT Interoperability and Standards

Interoperability between machine translation systems is desirable, but the systems are so similar in the way they present themselves to outside users that transitioning from one system to another, from a system integration standpoint, isn't that difficult. The APIs they expose to external systems all perform similar functions (request a translation, post-edit a translation, etc.), and are not terribly complicated. Moreover, many vendors have committed to implementing the TAUS web services reference API, as a way of providing a standard API that developers can use to interact with a variety of services. The reference API defines how to make a variety of requests that are common to all human and machine translation systems.

Translation memory and corpora, if they are stored separately from the machine translation engine, in a standard localisation file format (e.g. TMX), can easily be ported from system to system. Changing from one translation engine to another shouldn't prevent customers from using the corpora they've built up over the years.

Measuring and Benchmarking MT Quality

Measuring and benchmarking quality using an automated process remains a difficult challenge. While there are quality scales, such as the BLEU scale, they only provide a comparative measure of quality. This is important because what's really needed is an automated way to identify problem texts so they can be routed for human review and post-edit. At present, the standard practice is to have human reviews look at a certain percentage of texts, or spend an assigned about of time reviewing a subset of a project. As the volume of material being translated grows, it becomes easier for reviewers to miss defective translations. An automated process that could identify problem texts without generating a large number of false positives would be highly useful. Several companies are working on this problem and have fielded beta products.

Cost of MT Customisation

The cost of building custom or adapted translation engines remains quite high, mostly due to the cost of obtaining and pre-processing high quality parallel texts with which to train the translation engine. Globally shared translation memories will encourage translation vendors to pool translations so they can be combined to create large, high quality training corpora. While the technical challenge to building a shared translation memory has been solved, the primary challenge going forward is to encourage translation vendors to share their translations by default. LSPs often resist doing this, so it will take time to make this a standard practice.

4.11 Opportunities and Challenges for Translation Technology

Cloud Computing

As with other technology segments, the client/server model for Translation Technology is being quickly replaced with a cloud (SaaS) based service model. The advantages of hosted services compared to customer premise software are extensive, and include:

- Continual upgrades, with a continuous software release cycle (agile development)
- Greatly reduced system administration and IT costs
- High availability with highly redundant storage
- Subscription based cost model, versus perpetual licenses, reduces up front capital cost

- Accessible across broad range of devices and operating systems, including mobile
- Generally improved user interface and usability compared to enterprise client/server software

While there are a few situations where customer premise software makes sense, these use cases are generally limited to corporations and government agencies that have stringent security and data protection requirements (e.g. intelligence agencies), although in most situations, the case can be made that a cloud based solution can be every bit as secure as on site software. Therefore, we expect this migration away from client/server software (thick client) solutions to continue, and expect most new products to be designed around cloud computing.

Crowdsourcing

Crowd-sourcing has been a by-product of web 2.0. Crowd translation is becoming an increasingly important mode of translation, and has been widely used by popular web services to localize all or part of their offering. While pure crowd translation is relatively rare, most web companies that are expanding internationally are using crowd translation to some extent, typically by recruiting their users to contribute to their translation and localisation efforts.

The ideal solution combines crowd and professional translation in a way that enables large scale user participation while using professional translators in the background to score and vet new users (to prevent bad actors and incompetent translators from getting into the labour pool). This approach enables large scale, low cost translation while guaranteeing quality levels.

Translation management systems are beginning to incorporate support for crowd translation, so that crowd translators can work alongside professional translators, albeit with different access rights on the system. Transifex, a localisation management system, is a good example of this, as they allow any combination of machine, crowd/user and professional translation. We expect that support for crowd translation, and tools to manage crowd translators; will become a standard feature in translation and localisation management systems. Lingotek is one on the most established providers in this space.

Big Data

Tools to manage and query large data sets, such as map/reduce, enable developers to build applications that would have been prohibitively expensive to create otherwise. Statistical machine translation is one example, as these translation engines require millions to billions of aligned texts for training purposes.

Mobile

With most people shifting to laptop, tablet and mobile devices, translation technology vendors have to develop a mobile strategy to remain competitive. Translation management systems, for example, should enable translators to work on projects from tablets and smartphones. Small form factor devices, of course, have limited real estate, so this is not a trivial problem. The companies that develop an intuitive mobile interface will have a distinct advantage in the marketplace compared to platforms that only work with conventional computers. This is particularly true for systems that enable crowd translation, since crowd translation often involves large numbers of people doing individually small amounts of work (a perfect use case for casual translation via a mobile device).

Mobile also represents an opportunity for language service providers. Localisation to multiple languages is becoming a de facto requirement for mobile app and service developers. Companies that provide simple, high quality localisation solution for popular mobile environments and frameworks will put themselves in a position to capture other translation business as these companies expand internationally.

Social Platforms

Social media has become an important marketing and distribution channel for web services and publications, and will continue to grow in importance. Social media translation services will enable content producers and service providers to expand their reach in multiple languages, and to drive usage from these regions. This is new territory, with only a few providers, such as Helsinki based Transfluent, specializing in this area.

Social translation is another interesting opportunity. Amara and Viki, for example, both operate crowd translation and captioning services for web video. Viki in particular has attracted several million users via word of mouth, and has built up an impressive catalo of translated videos from major content providers. We expect to see more services like this for translating other types of content.

Interoperability and standards

The translation industry is notorious for its lack of standards. While there are standard file formats, like TMX and XLIFF, and a long list of secondary file formats, these standards tend to be over-engineered, and because of this difficult and expensive to integrate into products. XLIFF and TMX in particular are file formats that are specific to translations and aligned texts. This is the opposite of the situation with web services, where simple REST APIs and data interchange formats like JSON are dominant.

The basic issue facing the translation industry today is not the lack of a standard file format in which to store texts and their translations (both TMX and XLIFF work well for this), but the lack of standard procedures to initiate common translation tasks (for example, when a content management system needs to call a third party translation service to request a translation for a new document, there is no standard protocol). The TAUS web services reference API standardizes the way common tasks are done in the context of a publicly accessible web service. This enables translation technology and service providers to provide a common implementation that is the same for participating vendors. This, in turn, will enable developers to leverage re-usable code, libraries, and extensions rather than build custom integrations for each translation service or technology they choose to work with. The API was made public in September 2012.

Measuring and benchmarking quality

Measuring and benchmarking translation quality in a consistent way is another challenge for the industry. For many years now the industry has applied one size fits all approach to quality. It is an area with an inherent level of subjectivity. There has been no industry-level consensus on the different expectations for quality for different type's content and purpose. The TAUS Dynamic Quality Framework is an industry initiative that has begun to address this common issue. In 2012 a knowledgebase on quality assurance best practices, content profiling methodology and set of tools for quality evaluation was launched; these tools introduce industry benchmarking using the business metrics have been is missing in the translation industry.

A number of automated metrics have been developed to measure quality. These can provide an understanding of relative levels of quality, for example to indicate if one system has produced a better translation than another. However, they do not provide absolute measure of quality. They don't provide a measure of how quality is measured, nor do they measure factors such as style, grammar, etc. Unfortunately due to the variance of human language, as well as the fact that there are often many ways to translate the same thing, there may not be an algorithmic way to measure quality beyond counting obvious defects.

The likely solution will to combine automated and human mediated quality assurance processes to provide a more precise measure of quality. For example, statistical methods can be used to measure the likelihood

that a given text is a decent translation (using corpora of high quality translations as a baseline for comparison). Humans, on the other hand, are much better at catching more subtle problems with word order, grammar and word choice.

Sharing Language Data

Now that virtually all systems are connected to the Internet, it is straightforward to create shared translation memories that are continually updated as new translations are created. This is particularly important for machine translation and hybrid translation systems that require a large corpus of aligned texts to train the translation engines.

Language service providers are an ideal source of texts, since they continually produce high quality, aligned texts that, with client permission, can be fed into these systems. This process can be completely automated so that the language service provider incurs little or no cost to participate in the shared translation memory.

The main cost associated with participating in shared translation memories is the cost of modifying the translation management systems used by language service providers to mirror, or copy, completed translations to the translation memory's web API. Since most LSPs use one of a few TMS platforms, such as Globalsight or SDL's TMS, if these vendors do that system integration work, all LSPs using those platforms will have the option of participating in projects that share language data.

Companies that are using a home built translation management system, on the other hand, will need to do the integration work themselves. Fortunately, the process of posting completed translations to a public web service is about as complicated as submitting an HTML form. If an LSP is sophisticated enough to build their own TMS, they can add support for a public translation memory without much effort (probably a few days' worth of dedicated developer time).

The primary opportunity created by public translation memory will be for statistical and hybrid machine translation services, as these will serve as a high quality source of aligned texts to train these systems. This is particularly true for MT engines that focus on secondary languages for which aligned texts are difficult to source, and also for translation engines that target a particular domain of expertise, such as the automobile industry.

The risks of participating in public translation memories are pretty minimal. Their primary benefactors will be statistical and hybrid machine translation projects, especially those dedicated to poorly resourced languages, such as Tilde, which focuses on the Baltic languages. The main issue LSPs will need to deal with is getting permission from clients to share their translations with these projects.

Translation Technology Drivers and Inhibitors

The following table presents macroeconomics, global megatrends, specific market trends and labour supply factors that might affect the translation market.

Market Force	Assumption	Impact	Time Frame	Accelerator/ Inhibitor/ Neutral	Certainty of Assumption
Macroeconomics					
Economic situation	There is a strong correlation between the economy and IT expenditures. The global economy situation is impacting IT budgets, business and consumer confidence, the availability of credit and private investment, and internal funding.	High	Short-term	Inhibitor	★★★ ★★
Global megatrends					
Mobile	Spending on mobile devices will grow 23%, driving 43% of IT growth; a mobile strategy is priority number one for all industry players in 2012. Consumers are interested in personalised services which could be enabled through LT technologies.	High	Medium-term	Accelerator	★★★★ ☆
Cloud Computing	Cloud as a new paradigm of computing that is reshaping IT will help to evolve translation technologies. The key advantage to cloud services should be the ability of IT organizations to shift IT resources from maintenance to new initiatives. IDC estimates that cloud services (public cloud) increased 34% in 2010 to nearly $22 billion, or about 1.6% of IT spending, and that percentage should increase to 3% by 2014.	High	Medium-term	Accelerator	★★★ ★★
E-commerce	According to Eurobarometer surveys, the cross-border potential of EU retail e-commerce is not being realised; 51% of EU27 retailers sell via the internet, but only 10% support cross-border transactions. Translation technologies are the key enable technology to enable cross-border transactions.	High	Short-term	Accelerator	★★★ ★★
Social Media	Social media networks (such as Facebook or Twitter) and social business initiatives yields social data, where customer, partners and employees conversations create an insight-rich goldmine for businesses. New unstructured data (in the form of text, audio and video) analysis are required to generate profound insights.	High	Medium-term	Accelerator	★★★ ★☆
Big Data	The data growth is pushing the need for storage, analysis and big data technologies. The latter has intersection with LT technologies and its growth will push new developments and revenues for Lt vendors.	High	Medium-term	Accelerator	★★★ ★☆
Open Source	Open source has already had a major impact on machine translation. There are mature and actively used open source platforms for all types of machine translation such as Apertium or Moses.	High	Medium-term	Accelerator	★★★ ☆☆
Internationalization and globalization		High	Short-term	Accelerator	★★★ ★★

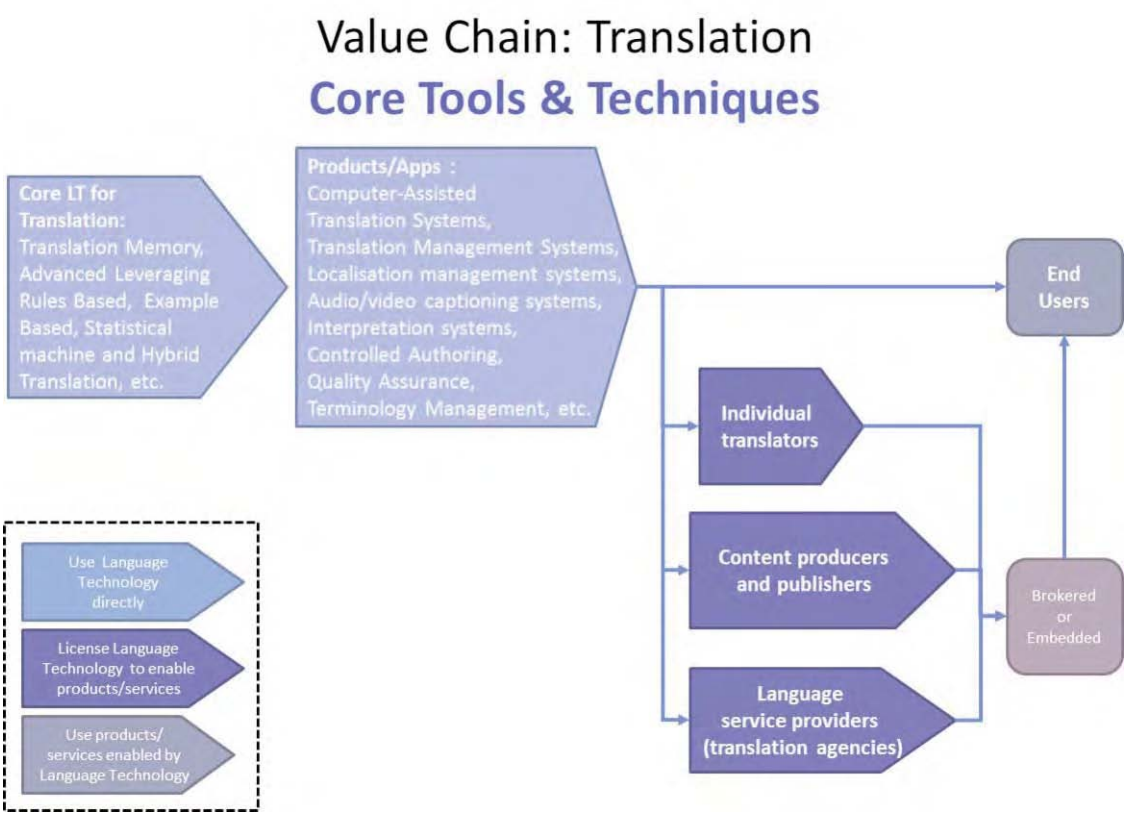
Specific market trends					
Translation Technologies convergence	Taking into consideration the accuracy limits of translation memory, companies are combining memory translation technologies to increase accuracy. This is producing a progressive convergence to a single translation engine able to use the most accurate technology depending on the situation.	High	Medium-term	Accelerator	★★★★☆
LT consolidation	LT are mature and are living a progressively convergence not only between them but also with other technologies such as content analytics, In-memory processing, in-database analytics, search-based applications, non-row databases, and non-SQL analytics.	High	Medium-term	Accelerator	★★★★☆
Lack of standardization and shared resources	Linguistic data resources are usually expensive or no shared by companies. LT companies must recreate the required resources inhibiting a faster time-to-market and the proper market development.	Moderate	Short-term	Inhibitor	★★★☆☆
Labour supply					
Talent scarcity	The availability and the skill level of talent have a direct impact on LT markets such as speech, translation or intelligent content. Whilst in the previous decades, the LT talent scarcity has been covered by IT experts, the current availability may inhibit adoption rates and market development and growth.	Moderate	Medium-term	Inhibitor	★★★★☆

Legend: ★☆☆☆☆ very low, ★★☆☆☆ low, ★★★☆☆ moderate, ★★★★☆ high, ★★★★★ very high

Translation Technology Value Chain

The following graphs present the value chain for Translation Technologies as well as for the different translation components and tools.

In the case of translation technologies, translation components and tools need to interact together for creating a translating project; in addition these components do not interact on their own for other purposes different from translation and localization processes. Thus all get represented by a unique value chain.



Intelligent Content Technology Overview

4.12 Background

The starting point of the modern and computational approach to the Intelligent Content industry could be found in the late 1940s when the first computer-based application related to natural language was developed to break enemy codes during World War II, and in the early 50s with the publication of Alan Turing's famous article «Computing Machinery and Intelligence».

Intelligent Content technologies aim to solve problems related to human and digital content transformation, search, analysis, delivery, etc. to produce intelligent content.

«Intelligent content» refers to content that is structurally rich and semantically aware, and is therefore discoverable, reusable, reconfigurable and adaptable and which is not limited to one purpose, technology or output.

These technologies rely on underlying techniques and tools such as natural language processing (NLP), categorization and clustering engines, and statistical approaches for processing the outputs of human language, such as written or spoken texts.

Many Intelligent Content applications require (or generate) linguistic resources such as word lists, terminologies, dictionaries, thesauri, taxonomies, or ontologies to enhance their semantic accuracy.

During the following decades, the relevance of Intelligent Content technologies has continued to grow with the explosion in digital information and the needs of businesses, consumers, governments, and other stakeholders to find, organize, navigate, publish, and make sense of it. These technologies continued to spread tentacles of functionality into any applications that required language understanding: in enterprise applications, in consumer Web businesses, and in online social environments.

The Intelligent Content Technologies are specifically designed for unstructured data and usually enable, both unstructured and structured content, from across the enterprise to be cleaned, restructured and enriched in a consistent manner.

As a matter of fact, Intelligent Content industry is playing a key role in many growing technology markets. For example, semantics is helping to improve enterprise search systems accuracy or social media analysis based on text mining is helping to understand the voice of the customer.

4.13 Core Tools and Techniques for Intelligent Content Technology

Natural Language Processing (NLP)

Natural Language Processing is a field of computer science, artificial intelligence (also called machine learning), and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, the process of a computer extracting meaningful information from natural language input and/or producing natural language output.

The market globalization puts pressure on the need to address language barriers to trade between countries. Those technologies that reduce these barriers automatically become more important in a 24 hours open market. The customers' use of online platforms, personal assistants, Interactive Voice Response systems (IVR) and other automated systems has increased in the last decade. So, companies depend more and more on the right interaction between their systems and customers. Natural Language Processing provides mechanisms to deal with interactions between humans and machines.

The foundations of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc.

In the field of Natural Language Processing, there exist two distinct focuses: language processing and language generation. The first refers to the analysis of language for the purpose of producing a meaningful representation, while the latter refers to the production of language from a representation.

Applications of NLP include a number of fields of studies, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems, SEO and so on.

Semantic Search Engines

Semantic Search Engines are systems to provide rapid access to text or unstructured data inside databases on desktops or across the entire Internet that formally encoding meaning (i.e. semantics) separately from data, content or applications, generally ontologies and the tools to build them.

Timely access to critical information in companies and institutions is vital in today's information economy. The knowledge worker's productivity relies on his capability to access to the right information in the proper time. In fact, information is useless if it can't be found and retrieved within its digital data life span.

So, as digital information is more and more abundant within the organizations, the search engines become more and more important to work effectively. However, unstructured data and language barriers have limited the success of search engines.

For that reason, search itself has evolved into a larger collection of technologies than just a basic search engine. The move from technology to product requires not only additional technologies, like text analytics and categorization, but also tools to adjust relevance, create filtering rules, protect information security, deliver appropriate results to different devices, and analyse usage.

The application of semantics to search engines seeks to boost accuracy search by aiming ambiguity via an understanding of context. Indeed, search and semantics have similar goals and rely on similar technologies: to make information more findable and usable.

Semantic search companies used several methodologies to integrate semantics into their engines: RDF Path Traversal, meaning-text functions, keyword to Concept Mapping, graph Patterns, Logics, fuzzy concepts, relations and logics.

Natural Language Query Engines

Natural Language Query Engines are systems to analyse and classify a question, and use NLP techniques to understand an answer, usually from a text source.

Under overload information, the key considerations for improving access include accessibility, high precision, ease of use, transparent retrieval across heterogeneous data sources and accommodation of rapid language change in the domain.

Another approach to provide innovative search solutions is to consider natural language query engines. The idea behind is that people can use human language to ask for information rather than Boolean operators, keywords or complex query structures. However, people's keywords usage habits undermine the value of this type of engines that needs a complete sentence to understand what has been asked.

Natural language queries are used to access information in a database. The database may be a base of structured data or a repository of digital texts in which certain parts have been marked as potential answers.

As the competition in the search engines is getting harder, more and more companies are combining different search approaches to refine and enhance the results and response time.

Categorizers and Clustering Engines

Categorizers and Clustering Engines are tools that provide semantic tagging for content elements using mathematical analysis or resources such as taxonomies or ontologies and Mathematical techniques to identify related categories of content. So, clustering engines organize search results by topic, thus offering a complementary view to the flat-ranked list returned by conventional search engines.

In a context of growing data, the efficient use of information can't be achieved unless the right mechanism is applied to organize information. This is the main objective of categorizers and clustering engines that automatically cluster results into categories that have been intelligently selected from words and phrases contained in search results. Categories help to navigate and access information providing human-level accuracy.

These engines can help companies in several areas: for example, to capitalize expert and domain knowledge or to develop legal discovery practices.

Statistical Algorithms

Statistical algorithms, such as statistical inference or data mining, provide applications based on automated algorithms for machine translation, machine aided human translation tools, localization tools, information retrieval and extraction, question answering, speech recognition, spelling and grammar checking, generation, etc.

Among the specific task that the statistical algorithms can provide, we must highlight: automatic summarization, co-reference resolution, discourse analysis, machine translation, morphological segmentation, named entity recognition, natural language generation and understanding, topic and word segmentation, etc.

Semantic technology

Semantic technologies provide an abstraction layer from data /content files and application code. It includes tools for auto recognition of topics and concepts, information and meaning extraction and categorization.

The applications based on semantic technologies can be found in several industries such as healthcare (electronic medical records), finance (regulatory reporting) or emergence response (disease monitoring).

Data Resources

Many intelligent content applications, and the tools they are built with, rely on underlying linguistic data resources such as annotating textual data (morphological categories, constituency and dependency syntactic trees, anaphora, discourse structure, word-sense disambiguation, parallel-text alignment etc.), lexical data (wordnets, translation dictionaries, valence lexicons etc.) and ontologies.

Although there exist many open repositories for specific linguistic data resources (most of them in English), intelligent content providers must create their own specific resources to develop applications. Some of them obtain them through shared projects with universities departments, but many times as an internal develop-

ment.

This lack of open linguistic data resources have been seen as an open opportunity for many companies that have created linguistic data resources markets for their platforms.

Technology Trends

Intelligent Content core language technologies are in a constant process of evolution. The current trends are mainly driven by market needs. Vendors intend to provide better solutions based on new developments.

- **Natural Language Programming:** is an ontology-assisted way of programming in terms of natural language sentences that can be used to produce formal representations of procedural knowledge or program the reasoning cycle and operational logic of intelligent agents. Companies like Wolfram or Sysbrain are approaching the market with this technology.
- **Towards to a semantic web:** companies like Google, Facebook or Twitter among others are progressively including semantic technologies into their systems with initiatives such as Google's Knowledge graph, Facebook's Open Graph protocol and Twitter Cards. Beyond the use of publishing standards (RDF, OWL, etc.) and linked data principles, these companies are also considering artificial intelligence technologies.
- **Convergence with other technologies:** Companies are interested in having solutions that provide a unified approach to information management (from access to analysis). Progressively the main vendors are integrating more and more technologies such as intelligent content, search, business intelligence, database, business analytics and big data. This integration is not closed to software.
- **Many vendors provide appliances: specific hardware and software configurations that solve a problem. The return of artificial intelligence:** Artificial intelligence (AI) has come in and out of many times in the past 20 years due to promises that underestimate the challenge of meeting the needs of business users. The current interest of artificial intelligence is related to language, learning and cognition. As language goes beyond understanding speech, parsing sentences or disambiguating multiple senses of words and includes understanding, artificial intelligence could help to identify patterns to analyse better language.
- **Resurgence of symbolic approaches:** IC companies are looking for new differentiation sources. Some vendors aim having a unique service proposition, others focus on technology differentiation. As a result, the market is living a resurgence of symbolic approaches.
- **Patents wars affect the R&D budget:** patent wars have become more and more usual in the IT market and they are producing significant constrains in the R&D budget and in the company's survival. LT vendors such as Vlingo fought a patent war against Nuance. Although Vlingo survived the battle, due to financial problems it was finally acquired by Nuance.

4.14 Scanning/Text Input

Scanning/Text Input are systems that "read" existing text or support the physical input of text through keyboards, keypads, etc. Optical Character Recognition (OCR) was one of the earliest commercial applications of pattern-matching LT and is still an important market. Keyboards (including Virtual Keyboards) can be enhanced with LT-enabled features such as predictive typing, continuous touch/swipe and even gesture input.

Scanning and text input software comprises two different types of applications (1) capture and image management software and (2) text recognition software.

Capture and image management software helps automate document-centric workflows that continue to be

paper based. It provides:

- Capture capabilities that convert images of scanned documents to text (via an embedded optical character recognition engine). Currently there are three approaches: Optical Character Recognition (OCR), Intelligent Character Recognition (ICR) and Optical Mark Recognition (OMR).

  - OCR is a specialized type of software that leverages artificial intelligence, pattern recognition, and machine vision to convert scanned documents into machine-readable text.

  - ICR is advanced OCR that is neural network based, and is typically used for handwriting recognition.

  - OMR is another capture technology that has been used for forms recognition and is in use today for barcode recognition.

- Indexing capabilities that automatically generate metadata for the scanned document for search purposes, provide an application user interface that facilitates manual indexing, or provide a combination • of the two.
- The ability to store and retrieve imaged documents, whether via the vendor's own repository and search services or via integrations with other content management solutions

Scanning solutions typically include one or more of the following functions:

- Classifiers, which automate the categorization of documents (both scanned and electronic) so that they can be automatically routed for further processing to appropriate subprocesses or applications. Classifiers use neural networks or other types of artificial intelligence so that users are spared the effort of creating rules, which would otherwise make deployment complex and time consuming, and require significantly more on-going maintenance to extend the system to new document types.

- Intelligent extraction facilities, which extract «fielded» information from the content of scanned and electronic documents so that the information can be mapped to enterprise applications and leveraged by business processes.

- Full text search: often an embedded capability within the solutions, especially those that encompass repository and records management (RM) services and capabilities.

- Business process automation facilities to implement workflow tasks such as notification, approvals, and exception handling.

Text recognition software helps to identify and predict text patterns. In fact, Text-based input technologies seek address user experiences issues around desktop, web, mobile, in-car navigation, seat-back entertainment systems and televisions text input. These technologies can be found in word processors, browsers, mobiles and other devices.

There are several types of text input technologies: input technologies and optimization techniques.

- There are two paradigms for input technologies: pen-based input and (virtual) keyboard-based input. The latter is becoming more and more the standard the facto for all type of devices.

  - Pen-based input: produces digital ink that should be recognize and it a similar problem as scanning.

- Keyboard-based input: produces machine-readable text that is suitable for indexing, searching, and handling by contemporary character-based technology.

- Regarding optimization techniques, they are a combination of movement minimization and language prediction (that exploits the statistical nature of a language to predict the user's intended letters or words).

As a result, text input solutions includes:

- Capture capabilities (via a virtual or physical keyboard)
- Predicting capabilities: based on linguistic (semantics, ontologies, thesaurus, etc.) and artificial intelligence to predict the word that is being typed.
- Error Correction: provides meaningful correction on real-time.
- Personalized text input experience: the system progressively is being adapted to the customer.
- Multilingual: provide writing capabilities in multiple languages.
- Industry specific dictionaries: provide access to specific domains such as finance.

Applications

Scanning and text input software provides several business applications:

- **Accounting:** Minimizing human touch points (and human error) helps companies reduce errors and exceptions and recognize errors earlier in the process (e.g., incorrect shipments), reduce cycle time (process more orders, more quickly), and improve customer relationships. This can be applied to: invoice processing, purchase order processing, sales order processing, new account opening and contract management.
- **Customer correspondence / support:** is a high-volume document processing opportunity for capture and image management vendors.
- **Human resources documents:** HR must cope with large volumes of documents of very diverse types, including recruitment, hiring, and on boarding documents; benefits enrolments and electronic forms; and employee agreements and annual reviews.
- **Case management:** increasing use of case management in government, healthcare, and legal.
- **Claims processing:** must deal with many different document types, including paper forms collected by branch offices, telephone documents, mail-in documents, faxed-in documents, externally generated documents in electronic formats, internally generated electronic documents, email documents, and Web documents. This can be also applied to insurance and warranty processing.
- **Product life-cycle management:** Barcodes and other technologies are used combined to manage the life-cycle of products in the retail industry among others.
- **Marketing:** QR Codes and the required scanning technologies are starring becoming mainstream in

marketing but also in new business models like virtual supermarkets.

- **Smart devices interaction:** text input technologies are the pillar for many applications for smart devices such as SMS, one-to-many text communication, email, etc.

Trends

Scanning and text input technologies are under different trend scenarios.

From one side, with the progressive digitalization of business processes the adoption drivers for scanning are aligned with the cost reduction culture. Companies seek for reducing overall costs, shortening cycle times, reducing errors, improving information sharing and reducing paper filing/storage. On top of that, these initiatives may be linked to other intelligent content projects such as ensuring compliance and auditability, improving visibility into business processes or improving litigation preparedness. As a result, cost reduction initiatives that include scanning technologies are helping to create a paperless culture and reduce the carbon footprint.

From the other side, the main drivers for text input technologies are related to providing better user experiences on any device that requires text input (from a mobile to an ATM) based on focus of attention, text creation versus text copy tasks, novice versus expert performance, quantitative versus qualitative measures and the speed–accuracy trade-off.

Market leaders

Scanning and text input technologies markets are currently influenced by several IT industries such as mobile industry (smartphones but also ruggedized devices) or the document imaging industry. Ruggedized mobile devices include devices of a handheld form factor that are of industrial design, have mobile computing capabilities, are designed for data capture, and have wireless connectivity. These devices must have a high-level operating system as well as application processors. Such devices are designed for data capture and have 1D, and often 2D, barcode scanning capabilities, representing the evolution of barcode scanning handheld devices to include on-board intelligence and wireless connectivity to back-end enterprise IT systems.

Ruggedized vendors such as Honeywell, Motorola, Casio, Datalogic, Fujitsu or Denso Wave relay on internal developments and third-parties scanning / text input software vendors to add functionalities to their products. This happens as well for the mobile industry, were Apple, Samsung or Nokia relays on third parties.

Scanning and text input vendors usually have OEM partnerships in several industries which translate into small revenue margins. One of the main actors in this market is Nuance due to the progressive acquisition of several companies. However, this market still remains strongly fragmented.

Challenges / opportunities

Opportunities

- **Mobile:** With the exponential growth of smartphones, tablets and mobile information workers the need for better mobile services is increasing. Use of mobile devices to capture document will introduce/educate users on the benefit of document capture and drive demand for scanners as users deploy an infrastructure to support digital document.

- **Scan behaviour should rise in the cloud.** The need to take paper content to the cloud is evident.

Although most survey respondents, similar to the results for print, expect no change in their personal or their company's scanning behaviour, for those respondents indicating a change, a strong percentage expect an increase. Hardcopy vendors would be wise to heed this expected transition and think about offerings (in the form of products or services) that address this need.

- **Digital data growth:** As IT continues to modify its annual spending away from technology toward information (i.e., content), the demand by employees, business partners, and clients to access information in digital format will be a major driver for scanner growth during the following years.

- **Vertical industry applications:** Environments such as warehousing and retail continue to represent the key verticals for scanning and text input applications. There exists a progressive expansion onto the retail shopping floor for both workers and shoppers alike. For workers, these new form factors are being used for inventory tracking and messaging. For shoppers, these devices are used to scan barcodes while shopping, ultimately expediting the checkout process.

- **Marketing:** As augmented reality is becoming mainstream, more and more applications for digital imaging will appear.

- **Combination / Convergence with NUI:** as Natural User Interfaces like touchscreens are becoming the standard the facto, new text input methods and techniques are required and should be integrated with these technologies.

Challenges

- **Capture more revenue in the mobile chain.** Text input has become a commodity and standard in the mobile industry.

- **Innovation:** due to the maturity of the technologies, innovation has to come from the natural user interface side rather than from the intelligent content technologies.

- **Competition**: With the rise of app markets for mobiles, pcs and companies, scanning and text input applications are rising. However, this competition is eroding price and margins.

- **Mobile scanners.** Mobile scanners are personal workgroup scanners that are small enough to fit in a personal bag for travel. Mobile scanners tend to require human interaction in feeding multiple pages (no ADF), but if an ADF is included, the document feeder is limited to fewer than 10 pages.

- **Scan behaviour should rise in the cloud.** The need to take paper content to the cloud is evident. Although most survey respondents, similar to the results for print, expect no change in their personal or their company's scanning behaviour, for those respondents indicating a change, a strong percentage expect an increase. Hardcopy vendors would be wise to heed this expected transition and think about offerings (in the form of products or services) that address this need.

Authoring/Creation

Authoring/Creation are systems to create, or re-create text and documents. Authoring tools provide support when creating textual content, particularly in formal and technical communication authoring. Authoring/Creation tools include checking tools (for spelling or grammar), controlled authoring environments, author memory systems, summarization engines, and new applications that use text mining to enhance the writing or publishing process.

Authoring and creation applications include products for authoring any type of content. Authoring and creation is only concerned with those that have specific features for facilitating the creation of customized content. Such customization features include the use of a content ontology that enables the description of content components for subsequent personalized selection, the direct assembly of customized documents from previously defined components, and the embedding of customized content selection formulas within a document template. The range of technologies combined for authoring and creation applications go from NPL to statistical algorithms.

Several markets have contributed to the evolution of authoring and creation:

- **Content management:** These solutions cover a great deal of territory - everything from document and Web content storage to XML content management, digital asset management, and related workflow.

- **Output management.** This technology focuses on improving the efficiency of generating and distributing paginated documents through print job queuing, multichannel distribution, and delivery workflow.

- **Enterprise information portals.** These products can already customize content based on user profile or other criteria. Because they currently focus almost exclusively on HTML output, they lack a paginated presentation model.

- **Content access tools.** These products provide content search, document categorization, and question answering capabilities so they already perform parameterized content processing.

Applications

The motivation to adopt authoring and creation technologies stems from a desire to improve process efficiency and customer relationships by communicating more effectively.

Efficient customization and personalization of documents is one means to improve this communication. A customized document can incorporate the timeliest data, eliminate extraneous information, and address the specific needs of each target audience. These advantages deliver different benefits depending on the documents business purpose:

- **Offers.** Offers inform another party about the availability and conditions of a potential transaction. Offers include documents such as direct marketing letters, quotes, and proposals. Customizing these documents can make them more appealing, generate higher response rates, and accelerate the closing of business, especially in time-based business (e.g., quote-to-cash).

- **Contracts.** Contracts document the terms of a potential or past transaction or other business agreement between two or more parties. Contracts include documents such as subscriber agreements, software licenses, and insurance policies. Customizing these documents can reduce conflicts by making them more comprehensible and can also allow enterprises to be more selective about when they offer certain terms.

- **Products and publications.** Documents can themselves be products. Externally-focused products

include newspapers, books, and guides. Internal-use products include training materials, compliance audits, and management reports. Customizing these documents can improve demand for existing products by offering more relevant information. Authoring and content can also expand the realm of possible products through finer market segmentation.

- **Documentation.** Documentation describes the properties of a product, process, or service. Documentation includes documents such as product specifications, user guides, and repair manuals. Customizing these documents can make them shorter by reducing irrelevant content and improve their effectiveness by eliminating confusion over whether certain content is applicable to a particular item.

- **Notification.** Notifications further or update an on-going relationship with a party. Notifications include documents such as statements, reminders, and alerts. Customizing these documents can engender an immediate psychological feeling of satisfaction in the party as well as enable them to more effectively use a product or service, further enhancing satisfaction.

Each of these document purposes corresponds to a generic type of business process, and the scale of a process for a given enterprise will affect the scale of document customization. An automobile insurance carrier may offer hundreds of quotes a day, customized with applicant information and special offers for which the applicant qualifies. An aircraft manufacturer may have to send out thousands of maintenance updates a month, customized to reflect the precise configuration of every aircraft. A pharmaceutical company may have to generate dozens of regulatory filings a month, based on pieces of content created by hundreds of different staff members. Segmentation by document purpose and segmentation by vertical market focus provide additional ways to cluster and examine vendor positioning and the market.

Specific solutions based on authoring and creation applications are corporate communication, marketing campaigns, print services, professional publishing (like newspapers), requirements quality validation, standardize document creation (for example corporate curriculum), translation cost reduction, automated documentation for manufacturing, compliance, customer satisfaction or e-learning.

Trends

Quality pursue is one of the major trends that pushes authoring and creation market evolution.

- **Technology integration:** The focus of authoring tools in maintaining a consistent text quality is pushing the inclusion of linguistic intelligence in these tools. Vendors are including technologies such as semantic technologies or statistical algorithms to enhance the common authoring tools (spell, grammar and style checker). Moreover, other technologies are being integrated such a as translation memory systems.

- **Machine data:** as internet of things extends in business processes, authoring and creation applications can be used to automated use of data into technical documentation.

Market leaders

After a consolidation phase where companies such XyEnterprise, Astoria, Arbortext were acquired, the authoring and creation market currently consists a wide fragmented set of niche companies providing services for specific industries.

Among them, the top worldwide vendors are Acrolinx, Congree, Across Systems, Oracle, SDL Global Authoring management or Tedopres. Small companies such as The Reuse Company have entered the last two years providing services for the IT sector.

Challenges / opportunities

Opportunities

- **Digital content growth:** as digital content is exponentially growing, the more important become authoring and creation tools to assure quality.

- **Focus on new vertical solutions:** with exception of media or manufacturing, the majority of markets are still early adopters.

- **Collaboration:** increasing collaboration among employees is one of the pillars to boost employee's productivity. Authoring and creation vendors should focus on enable collaboration capabilities for their platforms. In that sense, cloud computing is the key technology component to consider.

Challenges

- **Non-mature Market:** although authoring and creation technologies have been widely adopted by markets like media, manufacturing or high technology companies, other markets that could benefit from these technologies are not mature enough. For example, requirement quality analysers could increase the quality of software code and create a consistent corporate framework.

- **Verticalization requires new partnerships**: expanding authoring and creation technologies beyond media industry requires deeper and specific industry knowledge. As authoring / creation vendors are mainly focused on technology, they should find the proper VARs.

4.15 Search & Navigation

Search and Navigation are systems to provide rapid access to text or unstructured data inside databases, on desktops or across the entire Internet. Search & Navigation systems include search engines, platforms, and applications with browsing and navigation, applied to enterprise or Web content. LT in search appears in unified information access platforms.

Search and navigation applications create access to unstructured and structured information. This range of software applications and technologies analyses, tags, and searches text — often in multiple languages — and rich media such as audio, video, and image files. It combines several technologies such as information access platforms, extended search platforms, search engines, multilingual modules, question-answering applications, categorization/metadata tagging tools, categorizers and clustering engines, visualization tools for information navigation and analysis, filtering and alerting tools, and text analytics applications. This progressive Integration of different techniques seeks to boost search accuracy.

Traditionally, search has been classified as:

- Enterprise search: index data and documents from a variety of sources such as: file systems, intranets, document management systems, e-mail, and databases.
- Web search: index documents on the open web.
- Mobile search: index the content on a single computer.
- Desktop search: index the content on a single mobile.

Nowadays, with the increased demand of unified access platforms, this distinction is becoming less clear as

enterprise search should include the whole set of tools.

Applications

Search and navigation has become a central concern for companies as they need to deal with information overload. Due to that, these technologies provide a wide range of business purposes for which organizations are using them:

- Knowledge management
- Business intelligence applications/access to database data
- Social media monitoring/reputation management/listening platforms

Search engine optimization/search marketing

- Customer service/call centre application
- Mobile search from handheld devices
- Compliance
- Voice of the customer
- Online store/retail Web site
- Competitive intelligence
- Tag, classify, and categorize documents automatically
- R&D/innovation support
- eDiscovery
- Fraud detection (insurance claims, warranties, financial, etc.)
- Social tagging, profiling, network development (Enterprise 2.0)
- Expert location
- Automate master data management (MDM)
- Multilingual and translation software for content management, cross-language access to information
- Warranty claims analysis

As the buyer shifts from IT to business departments, the main drivers to invest in these technologies are becoming more business-related:

- Lower cost of managing and analysing information
- Unified access to different information sources — both structured and unstructured (providing a comprehensive view of the information across email, file servers, applications)
- Provide faster and easier access to structured data, to more users, in legacy systems
- Improve customer support
- Document classification and categorization for alerts, improved search, or archiving and records management
- Decision support and management
- Normalize information across repositories (ETL for content) to gather all the related information from different sources

Identify trends in our market quickly

- Support R&D (i.e., identify new products and develop new ideas quickly)

- Better customer understanding by email mining , call centre interactions, and so forth
- Find new sales prospects
- Comply with regulations (risk management)
- Monitor what people are saying about the company and their products on the Web or in the traditional media
- Publishing uses such as automatically creating specialized topical newsletters and alerts or displaying live, structured news/social media newsfeeds
- Mine for cause and effect (e.g., determine causes of product or campaign failures — the «why» behind the data from the BI systems)
- Reduce call centre costs
- Litigation support
- Near-duplicate document detection
- Precision support of foreign languages

Trends

The enterprise and navigation search market has seen successive waves of change. These upheavals have been caused by a confluence of factors that change the demands of the market for new features and functions. These factors, in order of their appearance, include:

- Search as a self-service environment for business users, not designed exclusively for information professionals.
- Cross-organization intranet search as part of a knowledge management portal.
- An expanded definition of full-function enterprise search, including both internal uses and outward-facing Web site search.
- Search deployed as a unification point for multiple content sources, either through integrating content in a central index or through federation.
- Emergence of search-based applications to address specific business process requirements (e.g., eDiscovery, online commerce catalogue search, and expertise location).
- BI and search converge in unified information access offerings, which take search beyond text and BI beyond transactional data analysis. This demand for unified search of data will push more mergers and acquisitions coming, for example, from traditional BI vendors that seeks to add search functionality.

Search-based applications means a disruption point for this submarket as it allows creating workspaces for information workers that:

- Are tailored to fit a specific task or workflow
- Combine multiple technologies and tools, particularly search, collaboration, authoring tools, content management, and analytics
- Integrate information from multiple sources
- Incorporate domain- and organization-specific term lists, taxonomies, and knowledge bases
- Hide technical complexity below an easy, compelling UI that may have dashboard-like qualities

Currently the top five drivers for search and navigation acquisition will be: eDiscovery; search engine optimization/search marketing; mobile search from handheld devices; tag, classify, and categorize documents automatically; and fraud detection (insurance claims, warranties, financial, etc.). However, new applications will appear like competitive intelligence or cybercrime detection.

Market leaders

The top 3 vendors in 2011 based on worldwide revenue were HP, Microsoft, and Google Inc., accounting for

34.3% of the market total. In Europe, Exalead acquired by Dassault Systémes is experiencing a significant growth that is helping to consolidate its position within the top ten vendors.

There is continued progress on the part of enterprise software providers like Microsoft, IBM, Oracle, and SAP to provide embedded search at advanced levels within their proprietary application platform offerings.

The main innovation source is to gather unified search and navigation platforms. To date, smaller search vendors have been at the forefront of creating unified information access platforms. Companies like Endeca, FAST (now part of Microsoft), and Exalead (now part of Dassault Systèmes) were quick to see the need to find all information, regardless of its source or format, and were early to add search across structured data to their offerings. They were joined by newer vendors such as Attivio, BA-Insight, MarkLogic, Sinequa, Palantir, Inbenta and Perfect Search.

Therefore, innovation comes from the small players and big players are acquiring these companies in order to complete their portfolio and have access to these innovations. So innovation in big players takes the form of being able to combine several technologies at the same time. For example, Autonomy has added content management, NLP, analytics, visualization, workflow, rich media analysis, and data analysis to its IDOL platform.

Challenges / Opportunities

Opportunities

- **Unified search platforms:** Newer approaches and technologies are breaking down the barriers between the BI and the search worlds.
- **Data is going to grow faster and faster:** Big Data, Social Media Data, Internet of Things are just a few examples of the shape of things to come. Companies needs to be ready to deal with huge volumes of data, combine information of a variety of sources and provide results on real-time. So, companies will need to upgrade their current search engines.
- **Search-based applications:** in the current context, ROI is king. Search-based applications provide a faster ROI for companies.

Challenges

- **Lose search- focus:** More focus on Big Data than in search. The first driver to invest in Big Data is to solve data volume issues.
- **The run for search accuracy:** accuracy will increase more from complementary technologies rather than incremental innovation.
- **Garbage-in garbage-out**: the first step to have a good data search is data quality. Search and navigation vendors should start including data management solutions to their portfolio.
- **Platform vs. application deployment:** creating a broad search platform will produce better accuracy but slower-results (search based applications) in the short term.
- **Closed internet sources are growing:** Facebook, Twitter and other social platforms are progressively changing their access policies. Vendors will face the challenge of providing an integrated search experience for their customers.

4.16 Content Analytics

Content Analytics are systems to provide access to or to derive specific high-quality information about entities, locations, actions and evaluations from text, unstructured data and rich media. Content analytics tasks includes categorizing text, clustering related text elements, extracting concepts or entities, modelling the relationships between entities or concepts, producing taxonomies, mining graphics, audio, and video, etc.

Content analytics provides insights of unstructured data combining several types of intelligent content technologies such as natural language processing, machine learning, clustering and categorization engines or semantics.

Traditionally, content analytics can be classified as:

- Text analytics: gather insights from text sources.
- Speech analytics: gather insights from audio sources.
- Rich media analytics: gather insights from video sources.

As companies are interested in vertical business solutions rather than generic platforms, this classification is less significant. Specialized solutions includes a combination of text, audio, still images, animation, video and interactivity content forms analytics.

Applications

Analytics is one of the top IT priorities worldwide. In particular, in 2011, 20.1% of companies consider that smart technologies (technologies that capture, process and present real-time information from partners, suppliers, customers, etc. via devices such as phones, sensors, chips in order to enable better data analysis and real-time / smarter business decisions) is one of the top IT priorities, as well as 18.4% of companies believe that for Business Intelligence / Analytics technologies[7].

As a result, content analytics applications are emerging in many sectors and disciplines at the same time. Among them, it is necessary to highlight:

- Security applications: to analyse plain text sources such as internet news, to discover threat patterns in the new undefined corporate perimeter, to track and monitoring terrorist activities, etc.
- Biomedical applications: to access provide based-data diagnostics, automated cancer recognition, etc.
- Search: to provide new ways to improve search results, etc. This is vital for example in those sectors that needs information index and retrieval (for example, academics).
- Legal discovery: to detect possible legal threats, etc.
- Media: to curate content, RSS subscription optimization, topic trending, screen recording, etc.

Marketing: churn analysis, customer micro segmentation, digital marketing strategy optimization, RSS subscription optimization, targeting, etc.

- Customer Experience Management. Including: sentiment analysis (to detect emotions and negative or positive opinions about a brand or product), loyalty analysis, influence score analysis, user feedback, user surveys, etc.
- Social media analytics.
- Testing: to improve quality, automated coding generation, etc.

Moreover, content analytics could help, from a generic point of view, to support the whole decision making process when combined with business analytics solutions:

- **Operational decision-making process:** Simulation, Risk management, Profitability KPIs, Collaboration,

Planning

- **Tactic decision-making process:** Ad-hoc, exploratory analysis, Root cause analysis, Operational metrics, Granular data from multiple sources and of multiple types, Rapid scenario evaluation, Collaboration
- **Strategic decision-making process:** Rules based automation, Pattern and exception identification, Guided navigation, Real-time monitoring

Trends

Content analytics is growing strongly under the influence of the Business Analytics and Big Data markets, as it is becoming more and more important to be able to analyse unstructured data. Therefore:

- **Technology convergence:** content analytics needs to be automated, on real-time and integrated into corporate information flow. So it's expected that in the context of the business analytics platforms, several technologies will converged: content analytics, In-memory processing, in-database analytics, search-based applications, non-row databases, and non-SQL analytics. As a result it is expected that content analytics functionality will migrate into the database.

- **Market consolidation:** This market is living a consolidation. New rounds of M&A will include industry and business process specific content analytic vendors. The main IT vendors are completing their information management portfolio acquiring companies that provide content analytics services and solutions. For example, HP acquired Autonomy, IBM acquired Coremetrics, Salesforce acquired Radian6, IBM acquired SSPS, SAP acquired Inxight, Oracle acquired Endeca and Accenture acquired Neometrics. So, a growing portion of buyers of content analytics technology will be services providers and application vendors that incorporate content analytics in their broader offerings.

- **Customer behaviour requires an integrated approach:** Although rich media analytics hold an important piece of the puzzle of the customer behaviour, it's not the single source of relevant information. Customers interact with companies through several channels and the social based interactions create social data that reflects the relationship of people to people, topics, ideas, or locations. So it is required to combine and integrate easily online (web, mobile, video and off-site) and offline content (CRM, POS, call centre).

Market leaders

In the content analytics industry we observe that there exist a rich open source ecosystem with solutions like R, Weka or Rapid miner and commercial solutions with important vendors like SAS, IBM, SAP, Microsoft, etc.

Big vendors not only include vertical content analytics solutions but also a complete portfolio that covers the main deploying needs (from infrastructure to visualization).

The top 3 vendors in 2011 based on worldwide analytics revenue were SAS, IBM and Microsoft. In Europe, hundreds of small companies operate in the content analytics market such as Temis, Attensity, Linguamatics or Picturesafe. Some provide a single tool or application, while others offer software that spans multiple market segments. Some of these vendors are highly focused on specific business processes and/or industries, while others offer horizontal technology applicable across the market. There is also a range of business models (e.g., commercial software and open source) and go-to-market strategies (e.g., on premises and software as a service [SaaS]) among these vendors revenue.

Challenges / Opportunities

Opportunities

- Data has become the more valuable asset and companies are looking forward to exploiting vertically in order to create competitive advantages.
- Companies lack specific resources to exploit data, so content analytics services providers have a strong opportunity rather than pure product vendors.
- Content analytic vendors should approach with packaged business solutions rather than with an open technology / platform approach to focus on ROI benefits.

Challenges

- Content analytics is just a small part of the analytics market and companies still lack having the proper analytics infrastructure.
- Companies lack of sufficient computing resources. Vendors should include cloud-based solutions to overcome this issue.
- Big vendors already have a wide portfolio of content analytics vertical solutions and they can focus on services. Small vendors should focus on innovation and niche solutions to survive.
- Data integration complexity and poor data quality are the two main corporate problems that vendors should help companies to solve.
- IC Works well into the intranet, still standardization needed for interoperability in the internet and cloud platforms.

Intelligent Content Technology Drivers and Inhibitors

The following table presents macroeconomics, global megatrends, specific market trends and labour supply factors that might affect the Intelligent Content Technology market.

Market Force	Assumption	Impact	Time Frame	Accelerator/ Inhibitor/ Neutral	Certainty of Assumption
Macroeconomics					
Economic situation	There is a strong correlation between the economy and IT expenditures. The global economy situation is impacting IT budgets, business and consumer confidence, the availability of credit and private investment, and internal funding.	High	Short-term	Inhibitor	★★★ ★★
Regulations	New regulation of data and information compliance will require LT investments	High	Medium-term	Accelerator	★★★ ☆☆

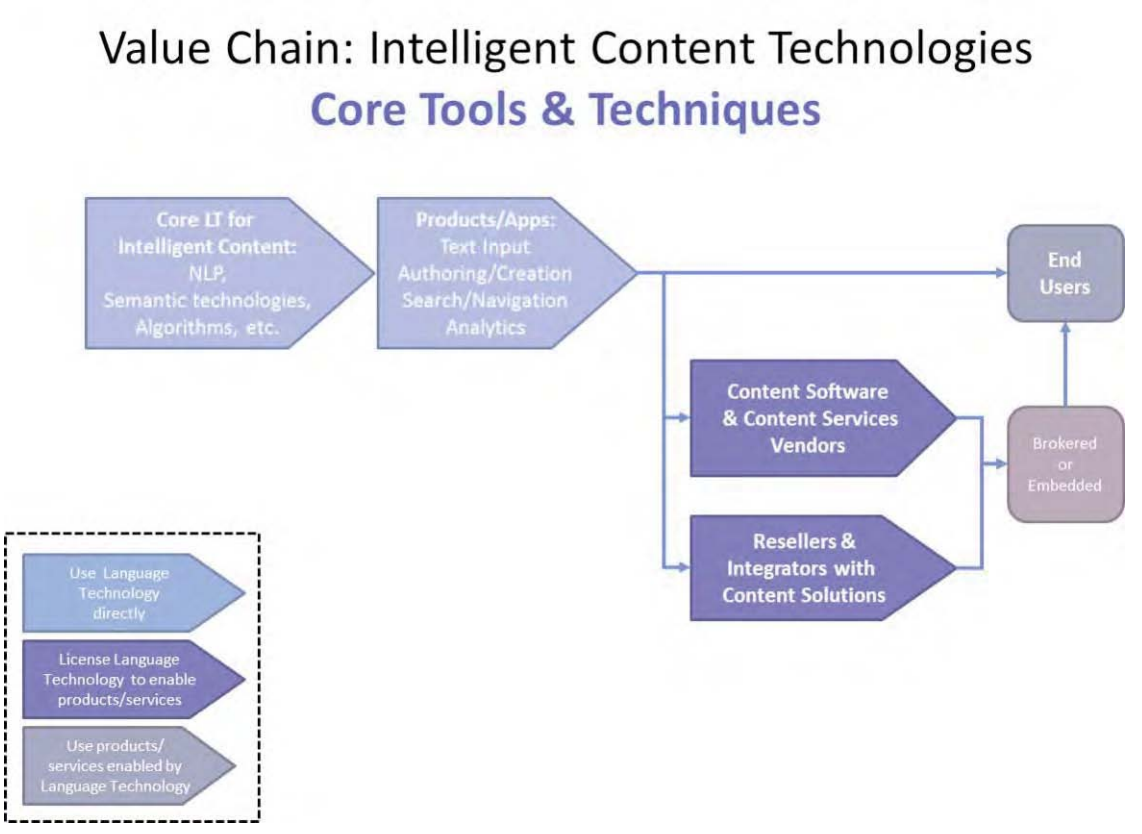
Global megatrends					
Social Media	Social media networks (such as Facebook or Twitter) and social business initiatives yields social data, where customer, partners and employees conversations create an insight-rich goldmine for businesses. New unstructured data (in the form of text, audio and video) analysis are required to generate profound insights.	High	Short-term	Accelerator	★★★ ★★
E-commerce	The progressive increment of e-commerce and m-commerce transactions generate more data to be analysed. This information is mostly unstructured. On top of that, new customers ask for more personalized services.	High	Short-term	Accelerator	★★★ ★★
Analytics	Frequently linked to Big Data, analytics will play a fundamental role transforming traditional industries to smart industries. In this context, text mining technologies will be fundamental to understand customer behaviour and sentiment.	High	Short-term	Accelerator	★★★ ★☆
Big Data	The data growth is pushing the need for storage, analysis and big data technologies. The latter has intersection with LT technologies and its growth will push new developments and revenues for Lt vendors.	High	Short-term	Accelerator	★★★ ★☆
Mobile	Spending on mobile devices will grow 23%, driving 43% of IT growth; a mobile strategy is priority number one for all industry players in 2012. Consumers are interested in personalised services which could be enabled through LT technologies.	High	Short-term	Accelerator	★★★ ★☆
Open Source	Open source has already had a major impact on intelligent content. There are mature and actively used open source platforms such as R or K-nime.	Low	Short-term	Accelerator	★★★ ☆☆
Internet of Things	Although the IoT deployment is getting slower than expected, the machine to machine and machine to person data is continuously increasing and pushing the boundaries of IT systems. In the following years, industries such as energy, automotive and healthcare will deploy data-based services that will require an accurate data management	Moderate	Medium-term	Accelerator	★★★ ☆☆
Privacy and security	As text analytics, new services and data insights increase, new concerns about the privacy and security of such data appears. Organizations must not only protect against new threats but to respect the limits of the private lives of their employees, partners and customers.	High	Medium-term	Inhibitor	★★★ ☆☆
Smart Verticals	Cloud, mobile, social and big data / analytics are the base for the vertical industry evolution. Many organizations have started the path to become a smart vertical: industries generating new competitive advantages based on the third platform pillars.	Moderate	Long-term	Accelerator	★★★ ☆☆
Cloud Computing	Cloud as a new paradigm of computing that is reshaping IT will help to evolve speech technologies. The key advantage to cloud services should be the ability of IT organizations to shift IT resources from maintenance to new initiatives. IDC estimates that cloud services (public cloud) increased 34% in 2010 to nearly $22 billion, or about 1.6% of IT spending, and that percentage should increase to 3% by 2014.	Moderate	Short-term	Accelerator	★★★ ☆☆

Specific market trends					
LT Consolidation	LT are mature and are living a progressively convergence not only between them but also with other technologies such as content analytics, In-memory processing, in-database analytics, search-based applications, non-row databases, and non-SQL analytics.	High	Medium-term	Accelerator	★★★☆☆
NUI technologies	NUI technologies such as multitouch are defining the new computing era. Intelligent content technologies are a key enabler to deliver a personalized experience in this context.	High	Medium-term	Accelerator	★★★☆☆
Market maturity	The maturity has brought a market consolidation. New rounds of M&A are including intelligent content vendors	Moderate	Short-term	Inhibitor	★★★☆☆
Labour supply					
Talent scarcity	The availability and the skill level of talent have a direct impact on LT markets such as speech, translation or intelligent content. Whilst in the previous decades, the LT talent scarcity has been covered by IT experts, the current availability may inhibit adoption rates and market development and growth.	Moderate	Medium-term	Inhibitor	★★★☆☆

Legend: ★☆☆☆☆ very low, ★★☆☆☆ low, ★★★☆☆ moderate, ★★★★☆ high, ★★★★★ very high

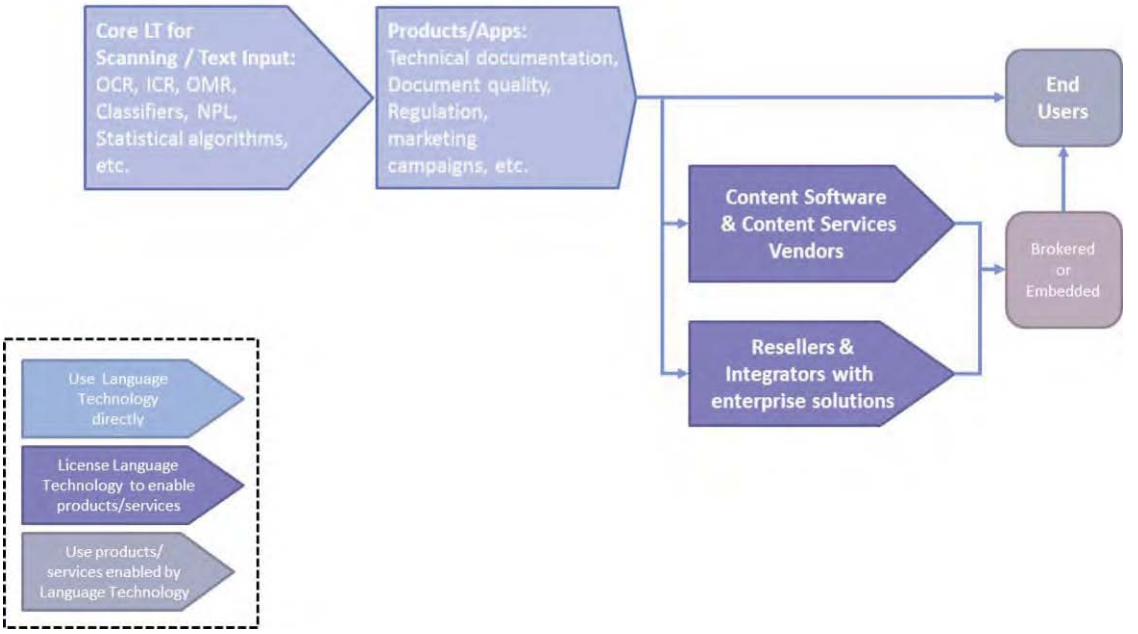
Intelligent Content Technology Value Chain

The following graphs present the value chain for Intelligent Technologies as well as for the different translation components and tools.



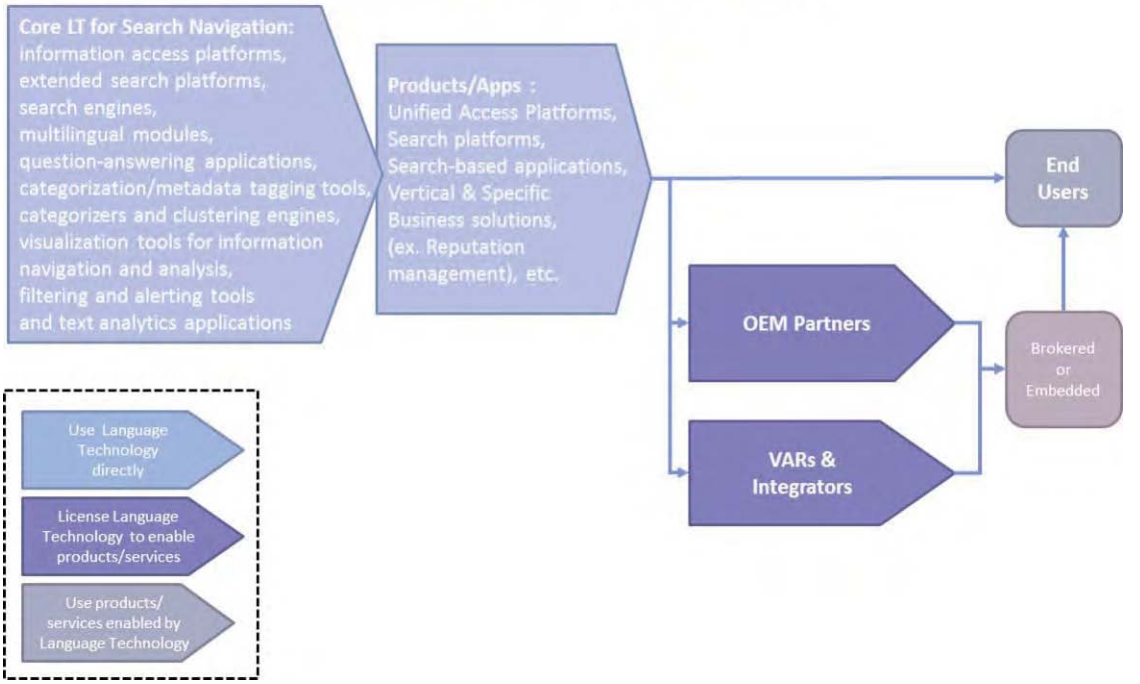
Value Chain: Scanning / Text Input Technologies

Core Tools & Techniques



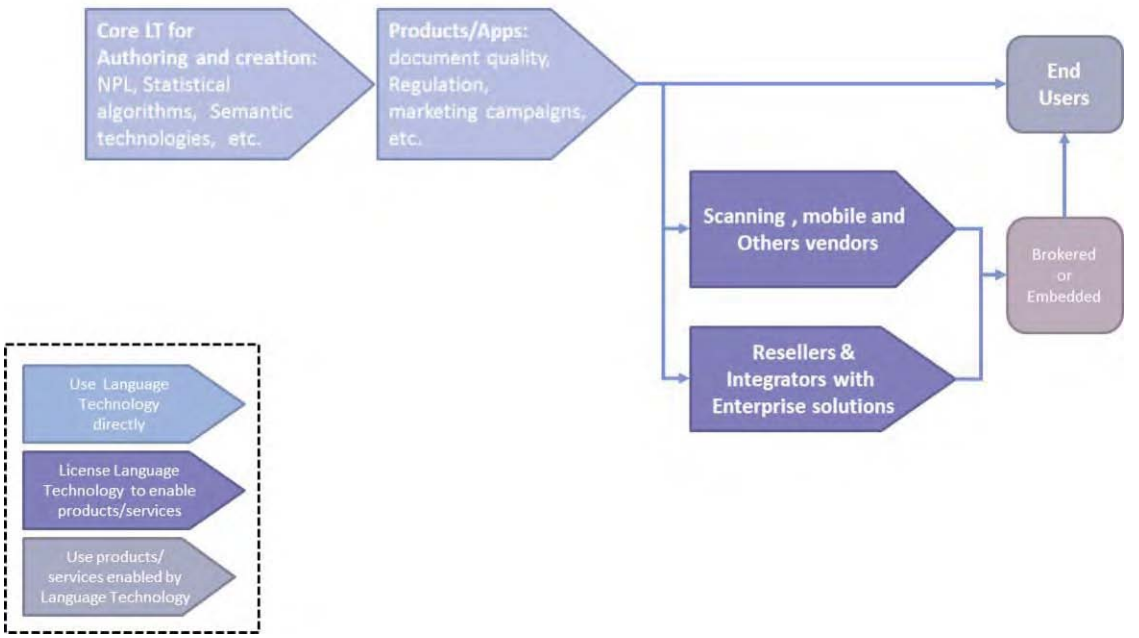
Value Chain: Search & Navigation Technologies

Core Tools & Techniques



Value Chain: Authoring and creation

Core Tools & Techniques



Annex: Business Analytics and Big Data definitions

Business Analytics Software Market Definition

IDC defines the business analytics software market as an aggregation of several software tools and application markets, with the functionality to aggregate, manage, organize, analyse, access, and deliver structured and unstructured data. As shown in Figure 1, the business analytics software market has three primary segments: performance management and analytic applications, business intelligence and analytic tools, and data warehouse platform software. Each of the primary segments has sub segments as follows:

- **Business intelligence and analytic tools.** Business intelligence and analytic tools include query, reporting, and analysis tools; advanced analytics tools; spatial information analytics tools; and content analysis tools. The business intelligence (BI) tools segment includes both standalone and embedded software:

  - Standalone BI software refers to tools that are packaged and marketed under separate products and are accounted for within the IDC functional markets of end-user query, reporting, and analysis and advanced analytics.

  - Embedded BI software refers to tools that are components of other software, specifically relational database management software or application server software. These products are not priced separate of the software into which they are embedded. Examples include Microsoft SQL Server Analysis Services embedded within Microsoft SQL Server, Oracle Reports embedded within Oracle Application Server, and IBM InfoSphere Warehouse Data Mining embedded within IBM InfoSphere Warehouse.

- **Data warehousing platform software.** IDC defines data warehousing as a process that organizes time-based data coming from multiple sources according to subjects meaningful to the business and driven by the need to inform decision makers. The data warehouse platform software competitive market includes two market segments:

  - **Data warehouse generation.** These software tools are software used in the design, cleansing, transformation, loading, and administration of the data warehouse.

  - **Data warehouse management.** These software tools are database management software used to manage data in the data warehouse.

- **Performance management and analytic applications.** IDC defines these applications as software that must meet each of the following three conditions:

  - **Business process support.** Commercial application software that structures and automates a group of tasks pertaining to the review and optimization of business operations (i.e., control) or the discovery and development of new business (i.e., opportunity)

  - **Separation of function.** Can function independently of an organization's core transactional applications, yet can be dependent on such applications for data and may send results back to these applications

  - **Time-oriented, integrated data from multiple sources.** Extracts, transforms, and integrates data from multiple sources (internal or external to the business) — supporting a time-based dimension for analysis of past and future trends — or accesses such a database

- IDC tracks the following performance management and analytic applications submarkets:
  - Customer relationship analytic applications
  - Supply chain analytic applications
  - Services operations analytic applications
  - Workforce analytic applications

- In addition, IDC classifies two functional markets as being part of the performance management and analytic applications market. These markets are:
  - Financial performance, strategy management, and GRC applications
  - Production planning analytic applications

Big Data Definition

The intelligent economy produces a constant stream of data that is being monitored and analysed. Social interactions, mobile devices, facilities, equipment, R&D, simulations, and physical infrastructure all contribute to the flow. In aggregate, this is what is called Big Data. However, this document sizes and forecasts the technology and services for managing, analysing, and accessing Big Data, not the data itself.

IDC's definition of Big Data technologies describes a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis.

Infrastructure

- External storage systems purchases by enterprises and cloud service providers and direct purchases of HDDs by select large cloud service providers (It also includes supporting storage software for device, data replication, and data protection of Big Data storage assets. Internal storage installed directly on servers is included in the server segment, not the storage segment of the market sizing.)

- Server revenue (including internal storage, memory, network cards) and supporting system software as well as spending for self-built servers by large cloud service providers

- Datacentre networking infrastructure used in support of Big Data server and storage infrastructure (Specifically, this forecast models spending based on IDC's research into the following markets: Ethernet switches, Fibre Channel switches, InfiniBand switches, and application delivery. Datacentres owned by enterprises and cloud service providers are counted.)

- Cloud infrastructure services that combine server, storage, and networking services, which are delivered through public cloud offerings

Software

- Data organization and management software, including parallel and distributed file systems with global namespace, highly scalable (size and structure) relational databases, key-value pair (KVP) data stores, content management systems, graph databases, XML databases, object-oriented databases, dynamic

application data stores and caches, data integration, event-driven middleware, and others

- Analytics and discovery software, including search engines, data mining, text mining and other text analytics, rich media analysis, data visualization, and other related tools

- Applications software including business process or industry-specific applications such as for Web click-stream analysis, fraud detection, logistics optimization, and others

Services

- Business consulting, business process outsourcing, IT project-based services, network consulting and integration services, IT outsourcing, storage services, security services, software and hardware support, and training services related to Big Data implementations

Table of Figures

References

- 2025 Every Car Connected: Forecasting the Growth and Opportunity. SBD for the GSMA. February 2012.
- Ambient Insight's Worldwide Market for Digital English Language Learning: 2011-2016 Forecast and Analysis. Sam S. Adkins, April 2012.
- Big Data Analytics: Trends to Watch For in 2012, Harlan Smith, Manager, Hitachi Consulting, 2012.
- Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. 2011.
- Building Unified Access to Information with PointCross. IDC CUSTOMER SPOTLIGHT Sponsored by PointCross Inc. Susan Feldman. February 2012.
- CiLT/InterAct International, ELAN: Effects on the European Economy of Shortages of Foreign Language Skills in Enterprises, Dec 2006.
- Common Sense Advisory, multiple surveys and reports.
- Competing for 2020, IDC Predictions 2012.
- Delivering In-Vehicle Speech Applications with Computing Headroom to Spare. White Paper by Intel and Nuance Speech Technologies. 2011.
- Designing User Experience, Michael Fauscette, IDC Software Business Solutions Group.
- DRAFT REPORT of 20 July 2012 on the proposal for a regulation of the European Parliament and of the Council establishing the Connecting Europe Facility (COM(2011)0665/3 – C7-0374/2011 – 2011/0302(COD)) - Committee on Industry, Research and Energy, Committee on Transport and Tourism - Rapporteurs: Adina-Ioana Vălean, Dominique Riquet, Inés Ayala Sender.
- EUROBAROMETER, Business attitudes towards cross-border sales and consumer protection, 2008.
- EUROBAROMETER, Consumer Attitudes to Cross-Border Trade, 2011.
- EUROBAROMETER, Europeans and their Languages, 2006.
- EUROBAROMETER, User Language Preferences Online, 2011.
- European Commission, DG Translation Studies on Translation & Multilingualism, Crowdsourcing Translation, 2012.
- European Commission, DG Translation Studies on Translation & Multilingualism, Size of the Language Industry in the EU, 2012.
- Gartner Perspective. IT Spending Overview 2010. Gomulka, B, Gartner. 2010.
- Global Automotive Embedded Telematics Market 2011-2015, TechNavio 2012; covers Bayerische Motoren Werke AG, Ford Motor Co., General Motor Co., and Toyota Motor Corp. Other vendors mentioned in the report: Audi AG, Fiat S.p.A, Chrysler Group LLC, Peugeot S.A., Volvo AB, Nissan Motor Co. Ltd., Honda Motor Co. Ltd., Volkswagen AG, Hyundai Motor Co, and Daimler AG.
- Global Commercial Vehicle Telematics Market 2011-2015, TechNavio 2012; covers Qualcomm Enterprise Services, AirIQ Inc., MiX Telematics Ltd., and OnStar Corp., Wireless Matrix Corp., ATX Group Inc., ETAS Group, and Minorplanet Systems plc
- Global Electronic Health Record Systems Market 2011-2015, TechNavio 2012; covers Cerner Corp., GE Healthcare Ltd., McKesson Corp., and Siemens Healthcare Ltd.
- Global E-reader Market 2011-2015, TechNavio 2012; covers include Amazon.com Inc., Barnes & Noble Inc., Hanvon Technology Co. Ltd., Pandigital, and Sony Electronics Inc.
- Global Medical Devices Market 2011-2015, TechNavio 2012; covers Baxter International Inc., GE Healthcare, Johnson & Johnson, Medtronic Inc., Philips Medical Systems, and Siemens Medical.
- Global On-demand Enterprise Applications Software Market 2011-2015, TechNavio 2012; covers IBM Corp., Microsoft Corp., Oracle Corp., Salesforce.com Inc., Workday Inc., Infor Global Solutions, Epicor.
- Global Smartphone Device Market 2011-2015, TechNavio 2012; covers Apple Inc., HTC, RIM and Samsung.

- Global Tablet Computers market 2011-2015, TechNavio 2012; covers Amazon Inc., Apple Inc., Motorola Mobility Inc., and Samsung Electronics.
- How 'Big Data' Is Different. Management of Technology and Innovation. Davenport, T; Barth, P; Bean, R. July 30, 2012. IDC Predictions 2012: Competing for 2020. Frank Gens. December 2011.
- IDC 2011 Digital Universe Study: Extracting Value from Chaos. John Gantz (sponsored by EMC)
- IDC 2012 Digital Universe Study: Big Data, Bigger Digi tal Shadows, and Biggest Growth in the Far East. John Gantz and David Reinsel (sponsored by EMC)
- IDC Worldwide Authoring and Creation Market 2012–2016 Forecast
- IDC's Worldwide Big Data Technology and Services 2012–2015 Forecast
- IDC's Worldwide Business Analytics Software 2012–2016 Forecast and 2011 Vendor Shares
- IDC's Worldwide CRM Analytics Applications 2011–2015 Forecast and 2010 Vendor Shares
- IDC Worldwide Document Imaging Market 2012–2016 Forecast
- IDC's Worldwide Enterprise Social Software 2012–2016 Forecast
- IDC's Worldwide Mobile Phone 2012–2016 Forecast Update: September 2012
- IDC's Worldwide New Media Market Model 1H12 Highlights: Internet Becomes Ever More Mobile, Ever Less PC Based
- IDC Worldwide Ruggedized device market 2012–2016 Forecast
- IDC's Worldwide SaaS and Cloud Software 2012–2016 Forecast and 2011 Vendor Shares
- IDC Worldwide Search and Discovery Software 2012–2016 Forecast
- IDC's Worldwide Unified Communications and Collaboration 2012–2016 Forecast
- In-car Infotainment (ICI) Market - Global Forecast & Analysis by OEM & Aftermarket (2011-2016)In-car Infotainment (ICI) Market - Global Forecast & Analysis by OEM & Aftermarket (2011-2016). Markets & Markets, March 2012. Industry developments and models. IDC's Software Taxonomy; Heiman, R; Clute, S; Lawton, M. 2011.
- KickStarter Stats - http://www.kickstarter.com/help/stats
- Lionridge Technologies: Taking Part In Language Technology Growth. Seeking Alpha, September 28, 2011.
- LT markets model and forecast bases development. IDC Spain for LT-Innovate. 2012.
- LT-Innovate fieldwork collected insights from interviews, events, desk research and SIGs. 2012
- Medical Transcription Market in North America 2011-2015, TechNavio 2012; covers SPi Technologies Inc., Amphion Medical Solutions LLC, BayScribe Inc., American Transcription Solutions Inc., and 3M Health Information Systems Inc.
- Mystery Shopping Evaluation of Cross-Border E-Commerce in the EU, conducted on behalf of the European Commission, Health and Consumers Directorate-General, Final Report by Dr. Katja Meier-Pesti & Christian Trübenbach, 20 October 2009.
- Perspectives on In-Vehicle Infotainment Systems and Telematics. How will they figure in consumers' vehicle buying decisions?. Accenture. 2011.
- SDL 2011 Annual Report. Enabling Global Business to engage with their customers.
- Social Commerce Trends Report, Europe 2012, bazaar voice
- Social Media Revolution 2011, Socialnomics, June 2011.
- Software Corp., Sage Software Inc., QAD Inc., Google Inc., SAP AG, NetSuite Inc., and Cisco Systems Inc.
- Speech Recognition in Mobile Devices. ABBI Research. Morgan, M; Orr, J. August 1, 2012.
- Strategic Management Journal vol. 33, "Entry into platform-based markets", Zhu & Iansiti, 2012
- Strategic Research Agenda for Multilingual Europe 2020, META-NET, 2012
- Taus Annual Plan 2011. Translation Innovation Think Thank. Interoperability Watchdog.
- The Forrester Wave: Message Archiving Software, Q1 2011. Hill, B. March 4, 2011.

- The Future of Mobile Payments [INFOGRAPHIC]; July 2011.
- The New ICT Ecosystem: Implications for Policy and Regulation, Fransman, Cambridge University Press, 2010
- Top Factors for Big Data Success. IDC: Henry Morris. Computerworld. The Power of big data symposium. June 2012.
- Unified Communications & Collaboration (UC&C) research examining UC&C implementation, investment plans and vendor requirements, IDG Enterprise, 2012
- What Every Exec Needs to Know About the Future of eCommerce Technology, Brian Walker, Forrester, August 2010
- Worldwide Business Analytics Software 2012–2016 Forecast and 2011 Vendor Shares. Dan Vesset, Brian McDonough , Mary Wardley, David Schubmehl. June 2012.
- Geoffrey Moore, Systems of Engagement and the Future of Enterprise IT, AIIM White Paper, 2010
- Henry Chesbrough, Open Innovation: The New Imperative for Creating and Profiting from Technology (HBS Press, 2003); Open Business Models: How to Thrive in the New Innovation Landscape (HBS Press, 2006); Open Innovation: Researching a New Paradigm (Oxford, 2006)
- Francesc Estanyol Casals, The SME Co-operation Framework: a Multi-method Secondary Research Approach to SME Collaboration, International Proceedings of International Development & Research, 2010
- Data Science Revealed: A Data-Driven Glimpse into the Burgeoning New Field, EMC, Dec 2011

www.lt-innovate.eu
twitter.com/ltinnovate
twitter.com/langtechnews
ltinnovate.blogspot.com

LT-INNOVATE.EU

Funded by the ICT programme of FP7

SEVENTH FRAMEWORK PROGRAMME