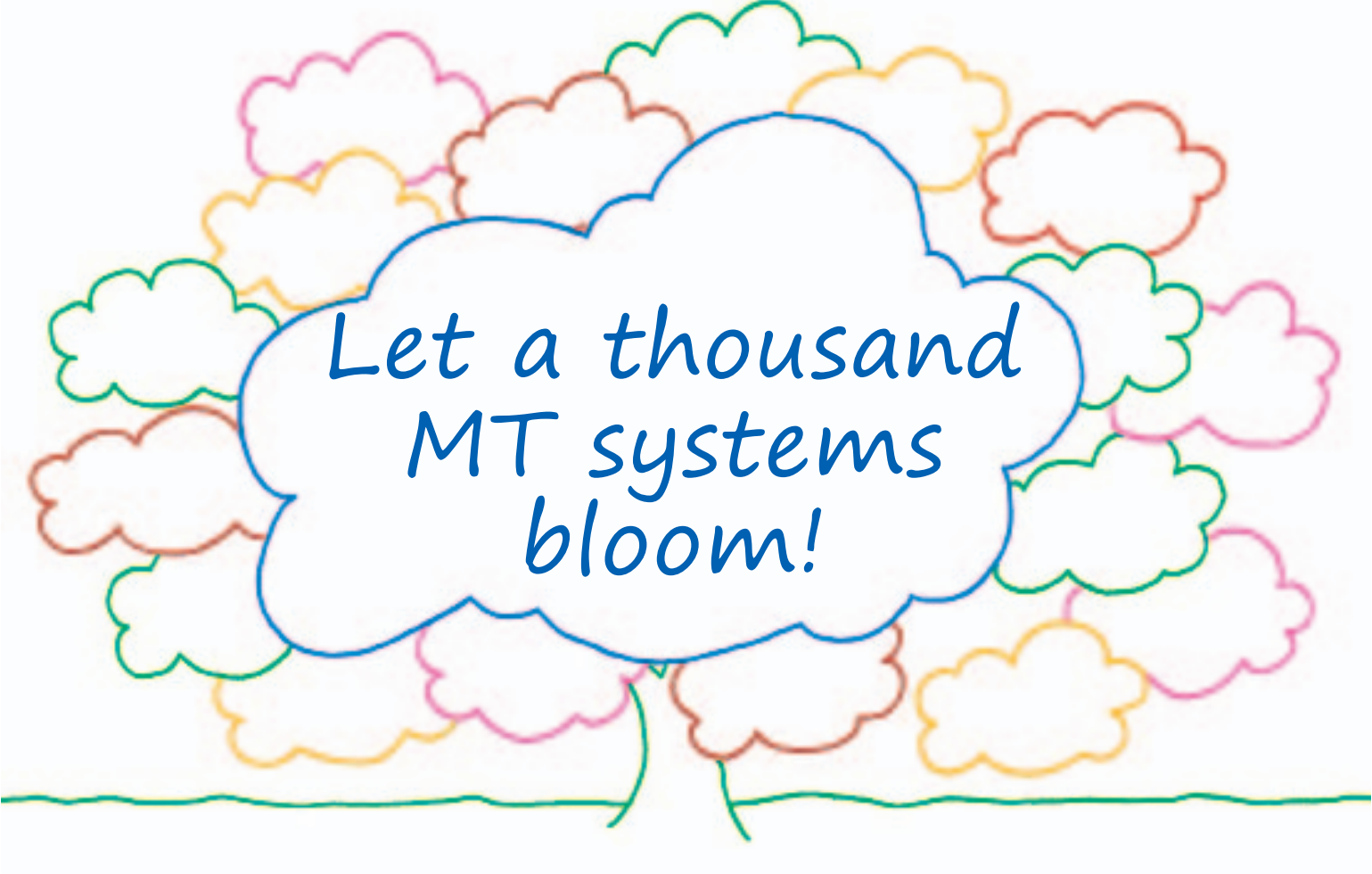


The profit of sharing



*Let a thousand
MT systems
bloom!*

Jaap van der Meer
TAUS User Conference 2009, Portland, 28-30 October

Looking into the future, I see a thousand MT systems blooming. I see fortune for the translation industry, and new solutions to overcome failed translations. I see a better world due to improved communications among the world's seven billion citizens. And the reason why I am so optimistic is that the process of *data effectiveness* is joining hands with the trend towards *profit of sharing*.

The first is somewhat hidden from view in academic circles; the other leads a public life in the media and on the internet. One is simply science at work, steadily proving that numbers count and synergies work. The other is part of the ongoing battle between self-interest and the Zeitgeist. And the Zeitgeist is destined to win.

Effectiveness of data

A good example of the irrefutable effectiveness of data is the Human Genome Project. Aimed at discovering the finest details of the human organism, this huge scientific project was launched in 1990. After thirteen years of research, it delivered a description of the sequence of the three billion chemical base pairs that make up human DNA.

To give an idea of the magnitude of this challenge, the Human Genome Project web site says: *"If the DNA sequence of the human genome were compiled in books, the equivalent of 200 volumes the size of a Manhattan telephone book (at 1000 pages each) would be needed to hold it all. This includes sequence data only and does not include data annotations and other information that can be associated with sequence data."*

The Human Genome Project has fueled tremendous growth in the biotechnology industry and significant advances in medical and biology science. A giant leap forward for humankind.

The overall structure of human language is far more complex than the human genome because it is the environment for human cognition and construction of social reality and it relies on yet undiscovered properties of human mind like consciousness and understanding. Therefore it requires a much larger effort in terms of data gathering and analysis. The first time a researcher from IBM applied statistics to the machine translation challenge in 1988, the audience was shocked. Language, we had all been taught, was based on intelligent design and not a subject for routine number crunching.

The data-driven approach was a slap in the face for linguists who have spent the previous 35 years researching machine translation in the traditional spirit of grammar rules. Statistical approaches do away with Chomsky's tree structures and the value of explicit linguistic knowledge of rules. Instead they depend on data to build a model of language structure and use.

The two key notions are *language model* and *translation model*. The language model provides the probabilities of fluent sentences on the basis of target language text samples. The translation model provides the parameters for the best possible translations. And this translation model requires lots of data.

The IBM research group that first developed this data-driven approach used the three million French-English sentence pairs from the proceedings of the Canadian Parliament. Initial results were mediocre, but more data tended to raise quality. Two decades later, statistical MT has become mainstream. And interestingly, further research is showing how parsing and rule-based techniques can further contribute to optimizing the training process. Factored translation models are now a standard part of the open SMT engine Moses and can be prepared by using 'part-of-speech' and 'lemma' features for aligned words in source and target. Even 'Chomskian' tree-to-tree or tree-to-string translation models become now increasingly popular, and can eventually reduce the size of training data needed to build good SMT systems.

Fifty-five years of MT research and twenty years of statistical MT research have resulted in a wealth of knowledge and tools; and academics and researchers around the world are sharing their results for various professional reasons to advance the cause. These open source tools include the BLEU metric, used for measuring and benchmarking machine translation output, GIZA++, a shareable tool for pattern matching and identifying co-locations of terms in parallel text corpora; and the Open Corpus Workbench and Tree Tagger to help developers add part of speech annotations to very large corpora of texts, and to run complex queries on multiple layers of this annotation. But models alone will not help: The key challenge now is to provide access to reliable – i.e. validated and curated - language data for specific genres, topics, subject domains and products.

At a TAUS Executive Forum in Beijing in 2007, Microsoft talked about their statistical MT system using training corpora of a few hundred million words. Then Google stood up and blew us all away, announcing a training corpus of billions of words that generated small yet systematic improvements in their BLEU scores.

Google researchers published an article earlier this year under the catchy title "The Unreasonable Effectiveness of Data". What is seen as 'unreasonable' in the Google perspective is that more data always seems better, even if the data is not a hundred percent grammatically correct and clean in terms of spelling.

This is yet another slap in the face for rule purists among linguists and language professionals twenty years after statistics caused the first shock. The Google observations highlight the fundamental question of whether the language we all use can really be controlled and normalized. We need to investigate more deeply and discuss more broadly the importance and exact definition of the *cleanness* of language data and text corpora for training and customizing MT systems. That is why this topic is on the agenda of the first TAUS User Conference and will no doubt feature again in future TAUS events.

The bottom line is this: Access to large volumes of language data helps increase translation productivity and improve the output quality of machine translation. Google is demonstrating this with their Translate button, and many TAUS use cases and pilot projects prove its relevance.

The TAUS Data Association provides an industry-owned platform for sharing billions and billions of words from all its members. And there is more to come. As a rule of thumb, an estimated 200,000 professional users of translation memory software almost certainly produce around 400 million words of quality translations every single day, or two billion words a week.

Our modest ambition is to host at least 20% of the output of the world's human translations, or 20 billion new words or so every year. This is not an unreasonable target. We only share translation memories from trusted and legitimate sources. We categorize the language data by industry, content type, data owner and provider, so as to make it easier to run domain-specific training and customization. This way, the TDA repository serves as a springboard for the translation industry to multiply its output and fuel growth and success.

The profit of sharing

'Sharing' is the name of the game these days. We meet it everywhere, from television, to the Internet and in advertising. "Win-win", is the typical catchphrase. It is core to web 2.0 in the form of social networking and peer-to-peer dynamics. Indeed, it is the Zeitgeist, the flavor of the month/year/decade.

Sharing – or reciprocity - has always been part of human transactions.

Humankind would never have come this far if we had not used this basic social instrument. Look at the way we use language on an everyday basis. Unless we share the same words and expressions, we would not be able to communicate at all. Yet sharing is never easy, as we always have to calculate the self-interest involved: *Am I giving away more than I get?*

This is the dilemma that members of the TAUS Data Association have to face, along with all other stakeholders in the global translation industry. If you share your translation memories, will you get more back than your initial investment?

Logically, the answer is 'yes', and here is how it works: If I am the only person in the world using a translation memory, I can only leverage words I have translated myself. If I share with one or more colleagues in my office, I get more benefit. If I share with the whole world, I gain almost unlimited benefits.

However, you can't extend this simplistic logic to the whole industry. Some will inevitably benefit more than others in a complex social and commercial environment once free resource sharing becomes the norm. So let's ask: Who would benefit most if translation memories were freely available to the whole world? Well, for one, translation buyers, especially those with a relatively small databases of own translations. But large translation buyers and owners also benefit, albeit in a different way. By sharing their large volumes of language data, they create a favorable market environment for their products and services as their terminology spreads more widely across new locales. Small and large language service providers benefit too, along with individual translators. In fact, when you think about it, everyone seems to benefit (yes, *win-win*, as they say in the commercials).

So let's ask a more relevant question: Who *loses* when translation memories are freely available? Probably those who use their translation memories to protect their market position. Language service providers, who often act as the *de facto* owners of large volumes of translation memories may be less inclined to share data. They may be interested by the even larger volumes of data they could access, but they are torn between this net benefit and the terrible fear of giving up what they have aggregated themselves. Freelance translators, on the other hand, don't worry that sharing translation memories would dry up the well and put their jobs at risk.

In response to those players who are afraid of sharing their translations, I can only say that the Zeitgeist will win this battle. Whether you like it or not, sharing is happening. There are tools out there that can scrape translations from the internet and align them to create new translation memories, and they are being used intensively. Cutting yourself off from this tendency will be dangerous. No one will understand why it is so important for you to keep your language resources for yourself. But if you were to share them, you would show you are working for the general good, just like the open source movement.

We must look beyond immediate fears, and focus on collaborating towards the Human Language Project. By openly sharing billions and billions of translated words, we will be able to give a tremendous boost to the global translation industry. Just as sharing data in the Human Genome Project was key to resolving one the biggest scientific challenges, and successfully grew the biotechnology industry.

The business impact

There is no question that the combination of *effectiveness of data* and *profit of sharing* will have a tremendous impact on the translation industry.

Easy access to large volumes of data in many different domains and languages will trigger the development of a range of dedicated MT engines. Customization and training of MT engines – still a specialized task today – will become faster and easier. The handful of MT development companies now operating will be joined by hundreds of new types of 'developers', for instance language service providers and corporate localization departments who could team up with universities to create their own dedicated engines. Post-editing will become a mainstream service, and word rates will increasingly be replaced by time-based or subscription and service-bundle pricing.

At the same time, the broader availability of much-improved – but not perfect – MT will stimulate the demand for high-value localization services that offer cultural and stylistic adaptation for specific customer groups, such as gamers or doctors or advertisers. Translation and language service providers will find new opportunities to specialize and innovate their offerings.

They will also face competition from new entrants with specializations in different vertical industries. These companies, both current and new, will find tremendous growth opportunities by integrating translation into almost any application used by consumers and business users.

TAUS USER CONFERENCE 2009 - LET A THOUSAND MT SYSTEMS BLOOM

The bridge between translation and speech technology, for example, will be critical to making translation-as-a-utility as widely available as the telephone or web today. So what might start as a threat will in the end turn into the greatest opportunity this industry could ever hope for.

TAUS roadmap

TAUS is expanding its role from missionary work to hands-on operations. We shall continue to drive the localization business innovation agenda, with a special focus on open translation platforms, collaboration networks, language data sharing, and translation automation.

In 2010 we shall be adding new services such as the **TAUS Tracker**. This will be an online directory of MT engines providing information about their use cases, business value and best practices. TAUS will also offer expert workshops on training and customizing MT engines, optimizing leveraging, and evaluating MT output.

The **TAUS Data Association** will focus on making translation memory sharing as friction-free as possible. Today members have to go to the TDA web site, log in, select a language pair, industry, content type, and translation tool before they can click on the upload button.

But TDA is already offering an API to enable members to integrate the TM sharing service into their own translation platform and editors. In future releases, we shall also be able to automate most of the current manual selections, (e.g. language ID, industry and genre (content type), classification and identification of the translation tool). This will deliver TM sharing at the click of a button.

The **TAUS Search release 2** and the **widget** now offer the world's translation memories to anyone's desktop. It is amazing what we can do already with just one billion words in the repository. We have integrated some of the great open source tools I mentioned above, allowing us to automatically translate of terms and phrases, include part-of-speech tagging and run reverse searches. As more and more data are being shared, the TAUS Search will be enriched with fully automatic synonym search, semantic classification, and crowdsourcing features.

Naturally TAUS and the TAUS Data Association are restricted to supporting their members and will not offer any commercial services.

A better world

Let me conclude by illustrating what I mean about translation failure, giving a historical perspective on how the world becomes better by sharing language data.

In the 12th century, the highly respected Gerard of Cremona, a Lombard translator of Arabic scientific works from the Moorish libraries of Toledo, Spain, came across the word *mumya* in a medical handbook.

He could not find any literary references to this foreign word, so he relied on his imagination and decided that *mumya* referred to a fluid found in mummies.

According to the Arabic tradition, *mumya* had a healing affect on scars and was a panacea for many diseases. Because of this translation, during the Crusades and for centuries afterwards, Egyptian mummies were robbed from graves and smuggled to northern Europe to harvest the *mumya* that was sold in pharmacies right up to beginning of the 20th century. It is only in recent times that it was discovered that *mumya* is in fact a mineral found on a specific mountain in Persia.

So I was fascinated a few weeks ago when I read about the discovery that the first sentence of the Bible – “*In the beginning God created the heavens and the earth.*” – is an inaccurate – or failed – translation from the original Hebrew. In the Hebrew text the word ‘*bara*’ means ‘separate’ and not ‘create’. So the opening of the Old Testament should therefore have been: “*In the beginning God separated the heavens from the earth.*” The original text does not refer to *creatio ex nihilo*, and also mentions multiple gods, (mis)translated’ as *angels* in the standard English Bible. You could argue that this erroneous translation has caused the deaths of millions of people. Yet Bible experts say they knew this already.

The Wall Street Journal reported last week on the pain that France’s General Commission for Terminology and Neology is going through to come up with a good French equivalent for ‘cloud computing’. This 17 member-group of professors, linguists, scientists and a former ambassador spent 18 months before coming up with the proposed term *informatique en nuage*, but in the final round of voting it was rejected and the team had to go back to the drawing board.

I looked up the term ‘cloud computing’ using TAUS Search and saw that EMC leaves it in English while Dell translates it as *infrastructures de traitement distribuées*.

The moral of this story is: Language will always be controlled by the ruling class, the most popular guy, or the latest fashions. That’s how it is built into our social evolution. So the question is not whether or not language conventions should be legislated and imposed by religion, government or committee. The real question is: how can we progress together and create balanced interpretations of the complexity of the universe.

Equal access to the world’s translation memories will be a giant leap forward for civilization. Gerard of Cremona would have loved to find the term *mumya* in the TAUS Search engine had it existed, and prevented the centuries of stealing dead bodies from the peace of their tombs.

We cannot blame the translators of the Old Testament from Hebrew to Greek to Latin to English for having lost word meanings on the way, when you realize how few tools and resources they had.

But the French Committee for Neologisms, and other bodies like it, has no excuse. They should embrace our super-cloud of language data right away and let community statistics decide.

TAUS is a think tank for the translation industry, undertaking research for buyers and providers of translation services and technologies.

Our mission is to increase the size and significance of the translation industry to help the world communicate better.

To meet this ongoing goal, TAUS supports entrepreneurs and principals in the translation industry to share and define new strategies through a comprehensive program of events, publications and communications.