Morten Pilegaard
Professor
Knowledge Communication Lab, Aarhus School of Business, Aarhus University, Denmark
Fuglesangs allé 4, DK 8210 Aarhus V, Denmark

Collaborative Repositories: An Organisational and Technological Response to Current Challenges in Specialized Knowledge Communication?

Background

Textual and communicative competence lies at the heart of the skills of professional mediators entrusted with the task of specialized knowledge communication, be they translators and writers of professional texts or subject matter experts. Such competence is particularly important in specialized writing that requires deep conceptual and contextual knowledge, especially when these mediators communicate across disciplines, languages and knowledge asymmetries. In today's transnational knowledge societies, the need for communicating deep knowledge within and across disciplines and languages is growing by the day.

Within linguistics, much of extant research takes an analytical, genre-based approach to knowledge communication, examining real life text-based communication breakdowns due to for example the 'intergeneric derivation process' whereby texts are created (Askehave & Kastberg, 2001), or it analyses texts from perspectives alternatively named hybridization (Fairclough, 1993), genre-mixing (Bhatia 1995, 1997) or recontextualization (Linell 1998). Remarkably little has been done to develop operational models or software capable of transforming the results of conceptual model analyses into systems that support the encoding and transmission of such 'deep' knowledge across disciplines, languages and knowledge asymmetries.

Specialized knowledge mediation is of particular relevance to translation as a discipline and as an industry. The translation industry is in a transition phase. The driving forces of this transition are both technological and organisational. The rapid advances and pervasiveness of information technology is thus changing the translation process fundamentally. The organisational response to the massive growth in global communication and the need for multilingual, often highly technical information is one of mergers and amalgamations creating large, dominant global language service providers (LSPs). At the other end of the organisational spectrum, small, mainly nationally based LSPs are becoming steadily more specialized. Both ends are responding to the megatrend of growing 'connectivity' (TAUS 2009a) by establishing or seeking membership of collaboration portals where language resources are shared to leverage existing translation memories (TMs). Such shared repositories have so far either been proprietary, like the internet-architected real-time Logoport of Lionbridge, or they have been narrowly focused in terms of membership and the nature of language data pooled like the 'TAUS super cloud' (TAUS 2009b). With their technical sophistication and organisational 'exclusivity', these new systems target a narrow, specialized fraction of those who are knowledge mediators, i.e. professional translators, and accordingly neither tap the by far richest possible source of specialized knowledge, namely that possessed by those knowledge mediators who are subject matter experts, nor place the TMs at the disposal of this large fraction of overall pool of knowledge mediators.

In principle, the Web gives unlimited access to a wealth of data[1] and information (Wiig, 1999) relevant to particular knowledge mediation processes. In practice, however, the creation of a single information space where information is accessible to those who need it for knowledge mediation processes just-in-time and in their mother tongue is hampered by technology barriers and ownership structures. The volume of quality parallel texts is considerable and growing rapidly, but generally such data, whether in the form of TMs or

[1] *Data* is here used simply to denote a string of letters, words or sentences. Following Wiig (1999) *information* may be defined as facts and data organized to characterize a particular situation, whereas *knowledge* may be defined as a set of truths and beliefs, perspectives and concepts, judgments and expectations, methodologies and know-how. Information is accordingly data that is made meaningful because it is put into a context; whereas knowledge is data made meaningful through a set of beliefs about the causal relationships between actions and their probable consequences. Such beliefs can be gained through either inference or experience.

aligned texts, are only available within the organisation where it originated and generally only accessible to the translator mediator as stated above. Recent years have seen the emergence of two initiatives which aim to make large pools of texts available, either for research purposes like the trillion-word corpus of the English language made available by Google to the research community in 2006 (Halevey, Norvig & Periera, 2009) and now currently used in the Google Translate service, or for commercial purposes like the TAUS Data Association (TDA). Google demonstrates the value of large amounts of data for the development of (statistical) machine translation engines. The Google Translate service performs surprisingly well, even if its data are entirely untagged and totally unclassified. The TDA, pursuing a similar end, works with trusted translations only and classifies data by owner, industry, domain and content type to allow its language service provider members to leverage the data in a targeted manner for translation sector efficiency and effectiveness purposes. The TDA will undoubtedly outperform Google on automated mediation/translation across languages within specialized domains. But only members may leverage the aligned text data. Flexible tools and organisational models to efficiently support web-based, inter-organisational sourcing and sharing of TMs and relevant text-based knowledge for non-translator subject matter experts within specialized or highly technical domains have so far not been developed.

These problems assume particular pertinence in the highly specialized domain of life sciences where subject matter experts act as knowledge mediators who must communicate across knowledge barriers, be they horizontal or vertical. These knowledge mediators often find themselves challenged when encoding messages carried in texts to recipients whose knowledge exists at less deep levels and who have less profound subject knowledge. In the pharmaceutical sector, for example, this challenge is felt by pharmacists when writing summaries of product characteristics (SPCs[2]) and patient information leaflets (PILs[3]). Both types of documents must be sent to and approved by the European Medicines Agency (EMEA) and national drug agencies in order for pharmaceutical companies to obtain a European and a national licence to market their products. This particular group is, indeed, facing a dual challenge: first, to structurally and linguistically gear the target text to maximum satisfaction of its communicative purpose within a social or discourse community context 'foreign' (SPC: national/international authorities; PILs: lay people) to that in which it was conceptually 'born' (expert domain); second, to cross this barrier in two languages. Often, these subject matter experts rely on non-subject matter experts, for example translators, for translation of L1 SPCs and PILs into L2. Yet, translators rarely possess the deep knowledge required of the subject matter expert to ensure smooth text-based knowledge mediation, here defined as properly localised L2 texts. Inversely, technological and proprietary barriers prevent subject matter experts from accessing the translators' TMs and aligned texts. Thus, both subject matter experts and a range of language service providers strive to solve nearly identical knowledge mediation (here narrowly conceived as translation).

The current situation represents a substantial unexploited potential in terms of specialized knowledge communication in general and translation productivity and quality in particular. What is needed is an architecture of cooperation and participation supported by a user-friendly web-based technology platform where collaboratively established language data may be leveraged to facilitate knowledge communication in its broadest sense within specialized domains where knowledge mediation is required.

The aim of the present paper is to report the use of the 'cluster' concept as a model for creating such an architecture of cooperation and participation within the specialized domain of life science and to present concepts and software systems for multilingual terminology- and text-based knowledge leveraging. The terminology system takes the form of a dynamic, multilingual specialized '*web-dictionary'.* Structured around the notion of genre, the text or corpus system takes the form of a multilingual, genre-based, meta-tagged *'web-corpus'* of medical text (genre) hierarchies. The focus is on leveraging of language knowledge items in multiple formats within a specialized domain via an online, integrated, interoperable and highly flexible web system offered to knowledge mediators across disciplines and trades.

[2] A summary of product characteristics is a text with a fixed (mandated) structure detailing the composition, form, clinical particulars, pharmacological properties and particulars of a medicinal product.

[3] A patient information leaflet is a text with a fixed (mandated) structure containing information about product details, directions for use, ingredients and warnings

Materials and Methods

Materials

The design of the *platform* containing the knowledge cluster's knowledge assets is inspired by the so-called *ba,* following Nonaka and Takeuchi, as discussed below (Nonaka & Takeuchi, 1995). *Ba* can be thought of as the knowledge cluster's shared space which assists cluster members in advancing individual and collective knowledge. According to Nonaka and Takeuchi, this space can be physical (e.g. offices), mental (e.g. shared experience) and virtual (e.g. software supporting knowledge exchange). This paper focuses on the virtual platform*,* i.e. the software, and how this software assists cluster members in advancing knowledge at the level of the individual, the group and the cluster. The platform is an SQL server 2008 database and an ISS web server as front end coded in .NET(C#).

Contents for the *web-dictionary* consist of already digitalised data, for example the Danish-English-Danish Dictionary of Medicine (Pilegaard & Baden, 2004), non-digital materials which are digitalised, data stored in company databases and data added on an ad-hoc basis by cluster members accessing the system. Data consist of source language terms and target language equivalents, definitions, collocations and examples. Contents for the bi-lingual *corpus* consist of SPCs, PILsinserts and labels for 360 products under the central EU procedure of the EMEA and the equivalent texts in Danish, approved by the regional (Danish) drug agency, plus a number of proprietary aligned texts provided by the users.

Setting. The Knowledge Communication Lab (Aarhus School of Business, Aarhus University, Denmark) is the cluster's "conceptual locomotive" that solicits public funds to design, develop and test knowledge mediation and subsequently facilitate software in a pre-market phase. During the initial test phases, cluster members (counting medical research communities, the language service provider industry and the pharmaceutical, the medical device and health information technology sectors) test the functionality and user friendliness, volunteer data to be contained in the software and provide feedback on software functionality and contents. Cluster members are successively tied into meaningful research and development (R&D)-oriented relationships with the KCL, where they can use the software and leverage its data in return for feedback, data donation and co-financing (time reported used on the software). They can thereby assure themselves that the software meets real needs and that repository contents can, indeed, be used to enhance their knowledge base. Their subsequent transformation into customers is undertaken by the techtrans unit TermShare Ltd, also at the Aarhus University.

Methods

Knowledge cluster. The present project draws on the theory of *industrial cluster,* which broadly defines clusters as groups of companies or institutions co-located in a specific geographic region and linked in interdependencies in providing a related group of products or services (Ketels 2003). For the present purpose, this concept was broadened to *knowledge cluster* [4] and extended in the sense that (a) participants were both producers of knowledge (researchers), producers of knowledge services (language service providers) and producers of tangible goods (pharmaceuticals and medical devices); (b) proximity was not one of geography, but one of activities and circumstances; and (c) interdependences sprang not from relatedness of products and services, but more from their complementary nature and from the collectively recognised need to establish collaborative and participatory structures to efficiently leverage these diverse competences. The knowledge cluster was designed to serve the dual purpose of driving university-based innovation and addressing select strategic challenges of the industrial arm of life sciences (pharmaceuticals, medical devices), namely enhanced inter- and intra-sector knowledge mediation and knowledge sharing for the industry to reap its full global potential. The cluster consists of the Danish health care, life science, pharmaceutical, medical device and translation sectors which were 'fused' to create a forum for knowledge dialogue between research communities and communities of (language) practice (Wenger, 1998) and for obtaining feedback on the virtual architecture, i.e. a collaborative knowledge repository facilitating merging and subsequent leveraging of the cluster members' data.

[4] Defined by Ketels 2003 as a local innovation system organised around research institutions and firms which tend to drive innovation and create new industries.

Collaborative knowledge repository. The present paper is not directly concerned with conceptual or theoretical models for collaborative knowledge building (Stahl, 2009); yet, it builds on the premise that underlying the theory of knowledge is a social epistemology, namely that knowledge is a socially mediated product. It adopts the view that the medium of knowledge – language – is grounded in the working life experiences of the individual, in the verbal interaction patterns of communities of practice and in their background knowledge. A knowledge repository is here "simply" a database containing aggregated knowledge assets captured in concepts and their terminological representation, on the one hand, and in prototypical, sanctioned exemplars of select text genres, on the other hand. These assets are systematically organised to facilitate searching, editing, retrieval and leveraging in its broadest sense. Leveraging the repository for knowledge mediation purposes feeds into individual and corporate knowledge creation in the sense that such leveraging is part of the process (see 'Virtuous knowledge circle' below), whereby the knowledge held by individuals is amplified and internalised as part of the organisation's knowledge base (Nonaka 1994). To create collaborative repositories, participants volunteer part of their data to the repository via a shared web platform. Cluster members volunteered two kinds of data: terminology within their respective main knowledge domains and aligned texts describing the physical products resulting from their knowledge activities.

'Virtuous' knowledge circle. The *web dictionary* is a dynamic, multilingual specialized web-based software that operationalises the different stages of the virtuous knowledge cycle proposed by Nonaka & Takeuchi, (1995). The system mirrors 'real life' working processes and allows repeat conversions of knowledge between its tacit and explicit forms and makes it possible for that knowledge to be codified and to spiral up from the individual to the collective level both within a group and further 'up' to the knowledge cluster level through a double validation loop[5].

Genre. As texts are seen as instruments of knowledge mediation, genre theory was an obvious choice and point of departure for the design of the *web corpus*. This theoretical point of departure was enriched by the insights brought by the focused study of specialized genres in general and the medical genres in particular, including studies on professional medical translation (Pilegaard, 1997), medical translation from a learner and learning-centred perspective (Resurrecció & Davies, 2007), production of particular genres like case reports (Wildsmith, 2003) and the more general aspects of information structuring in healthcare materials (Wright, 1999). The text repository is structured around the notion of genre following Bergenkotter & Huckin, (1995), Bhatia (1997a) and Swales (1990). The system is a multilingual corpus of medical text (genre) hierarchies (Bazerman,1994), i.e. texts essentially about the same topic, but in multiple formats within a particular specialized domain, in casu SPCs, PILs and labels. The genre approach to knowledge communication, borrowed from translation studies (Trosborg, 1997; Resurrecció & Davies, 2007), is considered particularly productive in the present project because it adds a socio-professional perspective that takes into account the insights, needs and practices of the non-linguistic mediators, the subject matter experts, often producing these genres.

Melting-pot and eclecticism. To overcome the weakness or intrinsic biases and the problems that come from single method, single-observer, single-theory studies, it was considered necessary to do 'theory triangulation', i.e. to use more than one theoretical scheme (cluster theory, knowledge management and functional linguistics). This had implications for *systems design.* The systems should simultaneously capture and represent significant idiosyncratic, departmental, corporate aspects and exemplars of particular concepts and their terminological representation in the web dictionary and genre exemplars in the corpus; at the same time, it should represent normative, sanctioned or mandated exemplars of these concepts and genres.

This also had implications for *data collection* in the sense that knowledge items and system design feedback were obtained from knowledge mediators from diverse sectors (life science, pharmaceuticals, medical devices, translation service providers), various professions (translators, pharmacists, medical doctors, biologists, medical writers) doing different writing or knowledge mediation tasks (L1 production, L2 production, L1-L2 translation, L2 revision) in various contexts and people with different different levels of linguistic competence (layman, semi-expert, expert). This melting pot-like approach was adopted to be able to tap knowledge items from so many sources, in so many forms and in so many communities of practice as possible.

[5] Compare Stahl 2009 for a conceptual model to guide software design supporting collective knowledge-building processes

Results and discussion

The knowledge cluster – an instrument for collaborative repository building. The main result of the present study is that a domain-specific knowledge cluster was created. The cluster counted members from the Danish health care research community, from life sciences, and from the pharmaceutical, medical device and translation sectors, who were contacted through the principal investigator's personal and professional network and by branching out through contacts to leading industry associations and professional organisations in these fields. The message sent was that by using the cluster's virtual *ba,* their members could improve knowledge mediation, reduce their translation and second language text production costs while simultaneously increasing the quality of their communication because they could leverage aggregated, validated, shared terminology and corpus resources in an easy to use knowledge sharing tool (Pilegaard, 2007). The knowledge cluster created a new collaborative and participatory architecture and thus demonstrated the feasibility of collaborative approaches both to technology design and to knowledge repository building where members volunteer their knowledge assets to a common pool. Generally, the concept of inter- and transdisciplinary knowledge sharing and the creation of a shared, domain-specific knowledge repository was welcomed both at policy and implementation levels.

It was initially assumed that creation of the cluster would promote beneficial feedback between the participants and would set in motion a chain-reaction-like process to the benefit of the participants and their (knowledge) production individually or group-wise (be they corporations, departments or simply ad hoc collaborative groups) and, eventually, result in (a) new forms of collaborative and participatory structures at other than cluster level and (b) new products at the cluster level[6]. Both assumptions held true: the first as evidenced by the growing number of cluster members using either or both tools, occasionally together forming ad-hoc groups across organisational boundaries; by the formation of loose organisational ties and collaborative activities between individual cluster members, like independent translators forming a company; and by a licensing agreement between the cluster hub, the KCL and a major health technology software company, the CSC. The latter assumption was also confirmed by the very fact that the web dictionary (www.medicinordbogen.dk) and the aligned genre corpus (www.laegemiddelkorpus.asb.dk) are now collaboratively established knowledge products in the Danish market, unique within the domain of life sciences and medical devices.

The collaborative repository built by the present knowledge cluster differs from repositories built by other known collaborative initiatives, for example the endeavour driven by the International Health Terminology Standards Organisation (IHTSDO, 2009) to establish collaborative spaces in the form of a systematised nomenclature of medical terms (SNOMED) accessible in multiple languages, including Danish (SST, 2009). However, the SNOMED does not map the terminological practices of the medical discourse communities. Its main concern is with ontology, nomenclature and prescription. The system is neither designed to continuously incorporate feedback on the psychological acceptability of terminology proposed; nor to tap terminological knowledge at the practitioner source. Hence, the system is a largely static, normative repository. Trade-specific efforts to establish collaborative repositories are also ongoing, for example the TAUS (TAUS 2009). Yet, this initiative is seriously challenged both by global initiatives driven by leading software companies like Google (Halevy, Norvig & Pereira, 2009) and by the highly fragmentary and competitive nature of the very business itself. Thus, the TAUS super cloud has just recently been released, and the major players in the translation industry seem only reluctantly to surrender their knowledge assets for fear of piggybagging (personal communication, unpublished data). Remarkably, despite fierce competition within the sector, the Danish cluster obtained participation from all major Danish translation service providers (including affiliates of global corporations), except one, which explicitly stated fear of piggybagging as its reason for not joining.

Successful creation of collaborative repositories would thus seem to require that such initiative be led by an unpartisan academic knowledge broker with a strong research base and no obvious trade anchorage. First, such a broker should be unshackled by commercial interests, which would seem to be a prerequisite for persuading knowledge owners to volunteer their knowledge assets to a common pool, especially in a

[6] Much along the lines of the Japanese Sapporoto IT Carrozzeria Project (http://www.it-cluster.jp/english/eng01.html) (23 June 2009)

competitive trade. Second, language service providers will only donate data if they have a positive return either qualitatively or quantitatively. Neither of these conditions is being met in the TAUS context. In a highly competitive business like translation, quality and validation is imperative. An academic broker with access to academic and business subject matter experts vouching for the quality of the knowledge items and for the fact that they do reflect current practices within their respective communities of practice through their validation activities is therefore in an ideal position to 'persuade' data owners to volunteer their knowledge assets. We may, indeed, hypothesize that the transdisciplinary nature of cluster membership lay at the root of the cluster's recruitment success.

Although cluster members differed in their priorities and often competed among themselves, at least within the translation industry, the incentives for sharing and the benefits from leveraging the shared knowledge outnumbered the risks and drawbacks. This provides for economy of knowledge, and the present case would seem to represent the first experience of an evolution from desktop to company server to industry-shared, even inter-disciplinary, language repositories.

Management and trust. Management of the cluster during the R&D phases was a two-tier system. First, a selection of the members formed a 'forum', namely a steering committee, to support knowledge exchange and dialogue between the different research communities and different communities of (language) practice (Wenger 1998). The steering committee also served as a deliberative forum for any issues arising from the merging of the members' data. Second, the knowledge cluster was led by the knowledge broker (the Knowledge Communication Lab) in a so-called 'hub-within-the-cluster' structure (Evers, 2008). This gave the cluster clear leadership and the needed clarity of purpose, scope and drive to set and to accomplish goals in the medium term. This structure was presumably also instrumental in nurturing the entrepreneurial culture of the research environment, in bringing into focus the uses to which various strands of research could be brought, and in softening the drive for economic benefit in the short term which could otherwise obstruct or jeopardise knowledge sharing. The hub was thus the principal body responsible for designing the mental, physical and virtual spaces, or *ba*s, and knowledge cluster members had clearly defined roles, as also recommended for industrial collaborative undertakings (Nonaka & Takeuchi, 1995).

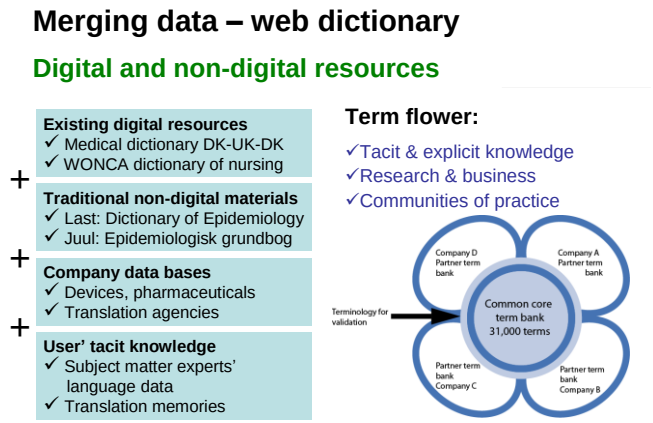
Like in industrial alliances (Gulati, 1995), a history of ties between the members forming the knowledge cluster generated the trust needed to obtain endorsement and active backing from major data owners, data donators and professional/interest groupings pursuing their own, occasionally different or even conflicting agendas. Cluster members were therefore recruited 'in waves': the first members through the principal investigator's personal and professional network. First wave members spread the word within their respective communities in an effort to recruit more members. It is speculated that this approach boded for the needed trust to recruit a sufficient number of members for the cluster. To ensure that the *ba* was designed to meet real needs, active participation and genuine knowledge exchange was sought at an early stage, and clear value propositions were presented to strengthen ties. We may argue that the initial trust based on a history of ties was supplemented with trust based on calculation by articulating a clear and calculated basis for mutual benefit.

Practicality – A second main result of the present project is that willingness to become a cluster member and to volunteer knowledge assets is also a question of practicality. Rather than supporting the traditional assumption that knowledge sharing is hampered by knowledge sharing parties' fear of being used or exposed to opportunistic behaviour, end users in the present project pointed to 'user friendliness' (Gulati, 1995; Borgman, 2009) as a necessary precondition for engaging in knowledge-sharing activities. If the tool is easy to use, it will be used. Whereas any significant extra burden on the user in the form of time consumed, lack of transparency or technical obstacles encountered is highly counterproductive. As long as the advantages of use outnumber the disadvantages, use is psychologically acceptable and the participants will accept the idea – without prejudice or fear. This confirmed our initial assumption that it is important to secure that the tool is easy to use for all knowledge mediators, also those unfamiliar with traditional computer-assisted translation technologies. User-friendliness is thus a precondition for reaping the benefits of knowledge sharing; moreover, we may indeed speculate that user friendliness will lie at the heart of *any* technology offered to transdisiplinary clusters whose members have diverse competences and technology proficiency levels.

Web dictionary: supporting the 'Virtuous knowledge circle'

An initial 800+ users were recruited during 2004-6, as described above, as test users of the web dictionary during the R&D phase. Among these, about one third became subsequent subscribers during the commercial phase which started in 2006 (Pilegaard, 2007). The web dictionary was the first Danish collaborative terminological knowledge repository, and it is now a fully commercial product licensed to the CSC. The web-dictionary can be accessed on a search-only basis by all medical professionals in Denmark, and it is being used on a daily basis as a terminology sharing tool by 65+ corporations and research institutes.

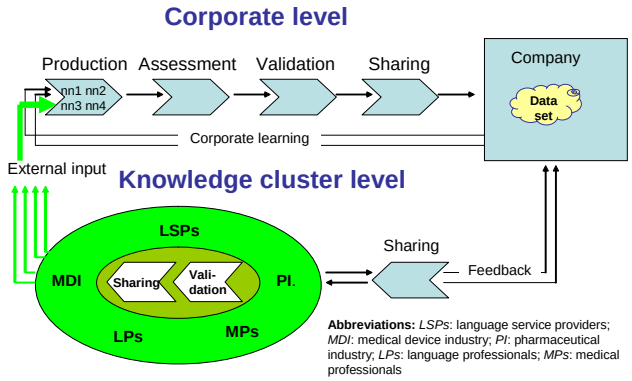
Figure 1



The knowledge base was built by merging existing digital and non-digital knowledge data (approximaely 25,000 equivalence pairs), by soliciting contents from company data bases and by tapping knowledge items (approx. 6,000 equivalence pairs) at the user base as shown in Figure 1. Knowledge data has grown quantitatively by about 35% from a total of about 31,000 equivalence pairs upon its launch in 2005 to nearly 40,000 equivalence pairs in 2009. The data is being qualitatively upgraded or improved on a continuous basis with a total addition and copy-and-edit turnover of an annual mean of 3,000 core knowledge items. The editorial board does not monitor contents build-up at corporate and group levels.

Figure 2

Overlapping virtuous knowledge circles



Abbreviations: *LSPs*: language service providers; *MDI*: medical device industry; *PI*: pharmaceutical industry; *LPs*: language professionals; *MPs*: medical professionals

The system is designed to reflect a 'life-cycle' of collaborative knowledge building, helping users to encode their knowledge, to discuss this knowledge with others, to exchange perspectives with others and to retrieve others' knowledge. The system thereby facilitates negotiation of shared understanding and formulation of knowledge in lasting representation forms. Individual users are part of 'groups' (i.e. a department, company or a research institute) or may themselves form a participate on an individual level. Groups are members of the cluster. Travelling along the 'corporate knowledge spiral' (upper half of Figure 2), when someone

reflects-in-action, for instance by reflecting on the terminological representation of a concept in one language (L1) during a knowledge mediation process in another language (L2), he/she becomes a researcher in the (language) practice of both contexts and is encouraged by the system to codify his/her knowledge (i.e. the knowledge products that emerge as a result are the terminological representation of concept *x* in L1 = *x1,* and its lexical representation in L2 is *x2*). In Figure 2, this first step is the *knowledge production* step. By codifying his or her knowledge, the user makes knowledge explicit and available to others' *assessment* (the second step). The system offers a slot for comments and negotiation of the meaning and form of knowledge items within the group. Any item must be assessed and explicitly approved within the group before it can be leveraged by other members. Assessment whether to accept (and discuss) the knowledge produced by an individual into the group's data set is undertaken by a group member (a validator) with relevant knowledge competences and system rights. If the knowledge item meets group quality and relevance criteria, i.e. if *validation* is successful (third step), the validator allows the item to become part of the shared *corporate data set* through the final step (*sharing*). Sharing involves making the knowledge accessible and leverageable for other members of the group, who can now literally absorb the knowledge, internalise, enrich or challenge it – the last two processes collectively named 'copy-and-edit'.

Users may edit knowledge items both from their own organisation as well as from the core material shared among all groups and their members. When a user performs this 'copy-and-edit' procedure, the system automatically creates a copy of the original knowledge item exclusively available to that user. This allows the user to continuously adapt existing knowledge items drawn from the group's private data set or from the core's shared data set to the unique context of his or her own practice. Copied and edited knowledge items are returned upon searches in a layered format where the user first obtains the private/group-edited item and with a simple click can access the original, common core item for comparison. Groups may have various motives for accessing the knowledge cluster's core data: a) to leverage data from the shared pool; b) to solicit knowledge cluster comments and validation of its own knowledge; c) to donate its own knowledge to the shared data set and hence enrich the shared repository; d) to challenge existing core knowledge items by offering comments and revisions. Any 'copy-and-edit' procedure marks the start of a new knowledge cycle at the corporate level where knowledge constantly builds up. The 'sharing-with-the-core' facility in turn allows knowledge to spiral 'up' to the knowledge cluster level in what we may term a double validation loop[7].

When an organisation decides to share data, they are forwarded to the knowledge cluster's editorial committee for comments, critique and validation. An advanced versioning system allows all entries in the shared pool to be edited by any user with basic editing rights and returned to the committee, which may determine to include the user's changes or simply return the item to the organisation for further processing. This ensures an ongoing dialogue. The system thus consists of two dynamic overlapping, seamless knowledge circles where the knowledge cluster pool continuously grows in quality and quantity owing to cluster member input and where groups can continuously access the most recent, upgraded material while simultaneously managing their own terminological data sets.

What this study adds. Most on-line web dictionaries are *static* representations of *normative* terminological knowledge. The present system differs from existing terminology systems, not only by its collaborative approach to repository building as discussed above, but also by offering a comparison perspective and by being highly dynamic. *Comparison* and *dynamism* are built-in system features. Thus, the web dictionary (a) aggregates input from various individuals and/or groups (the petals in the 'terminology flower', Figure 1), thereby creating a truly collaborative repository; and (b) allows for easy comparison between individual perspectives within groups, thereby facililtating adoption and adaption of ideas from other persons' perspectives, and between groups and the core. This fosters convergence and sharing of insights and interpretations at group level, which is critical to knowledge construction in any collaborative community (Stahl, 2000). For example, applying a new release of the software (TermShare), one of the cluster groups, the Via University College, reported improved knowledge sharing and better mutual understanding among ergotherapists, physiotherapists, nurses and doctors and better learning outcomes in clinical training courses. The software created a virtual *ba,* allowing group members from the different professions involved to explain key concepts relevant to all from their respective professional perspectives (unpublished data). Adding to intra-group comparison, users may also map their idiosyncratic knowledge items and items sanctioned at group level against items mandated or sanctioned at cluster level.

[7] Compare conceptual model to guide software design supporting collective knowledge-building processes in Gerry Stahl: A model of collaborative knowledge building.

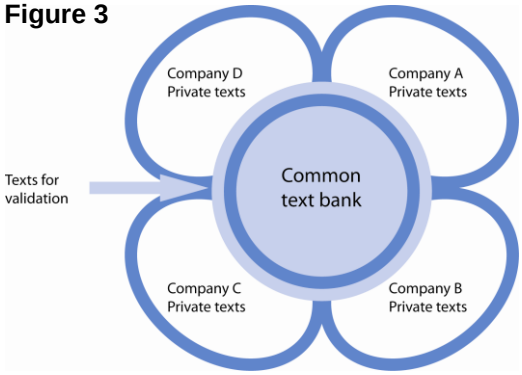
Viewed differently, we may argue that the comparison perspective is equivalent to an invitation to challenge existing knowledge, both at group and core level, even if the most delicate element in collaborative knowledge building is, perhaps, precisely the issue of negotiation of meaning because it involves power differentials of all kinds. Rather than resisting the renegotiation of meaning of established knowledge items as is the case in existing, normative systems (e.g. the SNOMED and existing web dictionaries), the present system by its very design welcomes any negotiation of change. The simple rationale underlying this is the assumption that negotiation of meaning is critical to help different perspectives converge on shared knowledge in a collaboratively established repository – and the fact that the domains of life sciences and medical devices, like most other specialized knowledge domains, undergo such rapid change in the face of technological innovation and accumulation of new insights that any system with the ambition of being relevant to specialized knowledge mediation must be highly dynamic and able to tap that knowledge 'just-in-time'. Such needs cannot readily be met by conventional dictionaries whose lead time is simply too long. What is needed today is web-based, dynamic and collaborative systems that simultaneously reflect static, normative representations of terminologies and capture and represent new conceptual and terminological developments as they surface.

Web corpus: genre hierarchy representation

The user base for the bilingual aligned text representation system was narrowly targeted to include language service providers and pharmaceutical companies. Users were recruited mainly among prior R&D users of the web dictionary. A total of 51 groups took part in the R&D phase, a majority of whom are now in the transition process to become commercial users of the software. The system is currently being used on a daily basis by 35+ groups or corporations.

Briefly, the system allows users to browse through 360 central procedure SPCs, PILs and leaflets in one or two languages[8]; to search Google-wise across documents either broadly or narrowly by language, procedure, document type, word or word string to check approved wording and/or find translation in the target language; to create own text repositories for section and/or company eyes only. Additionally, it grants users discrete rights to share, edit, upload and search in own text repositories[9] on needs and job function/qualification basis; to share access to 'private' text repositories across sections/departments and/or with industry repositories; and to search in shared or 'private' repositories separately or simultaneously.[10]

Figure 3



A particular challenge addressed in the text-representation system is (a) to facilitate knowledge communication across contexts and knowledge asymmetries and (b) to offer the same opportunity for

[8] All texts have been approved by the EMEA and regional drug agencies and are, therefore, assumed to satisfy the criteria of contents and form of such genres within these particular contexts. We may see them as protypical genre exemplars

[9] A facility currently used by Pfizer, PharmAdvice among others

[10] Other corpora have been built. One such corpus, serving the knowledge needs of a 'law cluster', includes the entire Acquis Communautaire in all European languages

comparison at text level as the web dictionary provides at the level of concept and term. The former (a) was made possible by allowing users a choice between different data sets and by meta-tagging these data sets by language, origin, genre, subject, time and other relevant features. Thus, users can access and compare different representations of the same knowledge items, for example descriptions of adverse side effects in expert-to-expert tenor in SPCs versus corresponding descriptions in expert-to-lay tenor in PILs. Following the 'terminology flower' principle as illustrated in Figure 3, users can also compare core text data sets with their own, private data sets. In other words, the core contains prototypical exemplars of textual products (Izquierdo, 2008) accomplishing particular, contextually determined communicative purposes. They reflect the conventions, restrictions and typicality governed by principles and practices of interaction (Bazerman, 1999) within the pharmaceutical sector. The 'petals', on the other hand, contain user or private variants of the prototypical forms, reflecting that genres are non-static and may exist in forms that defy prototypicality, even in mandated genres (Askehave & Zethzen, 2008)[11]; indeed, we may argue that the texts contained in the petals are reflections of textual practices that continually evolve and change in response both to demands extraneous or peripheral to the 'true' communicative purpose they should serve. A modified version of the "copy-and-edit-principle" of the web-dictionary was applied in the corpus: the system only allows users to access, manage and continuously edit texts in their 'private' text repositories, while at the same time accessing and leveraging texts on a copy-paste basis contained in the common or shared text bank.

What this study adds. Most existing corpora are either minimally or untagged, monolingual data repositories (like WordNet, the British National Corpus), tagged corpora intended mainly for research and didactic purposes (Izquierdo, 2008) or specialized aligned data sets contained in computer-assisted tools available mainly to language service providers to leverage prior work for business purposes. The present genre-based corpus represents one of the first attempts to assist knowledge mediation across knowledge and language barriers by pooling large amounts of parallel texts within a highly specialized knowledge domain and by making these texts equally available to translators and subject matter experts within a knowledge cluster. For the language service industry, which has evolved into a business where economic survival is crucially dependent first and foremost on the quality and the quantity of the data it owns and controls, and secondarily on its comparative advantages in terms of translation process optimization and customer loyalty building, the knowledge cluster offers crucial quality inroads. Language service providers who do not use these tools are likely to forego the opportunities of reaping the qualitative benefits from capturing the terminological and textual knowledge of the non-linguist expert users of specialized languages who are the principal knowledge mediators in this domain. They may also forego the opportunities arising from cross-sectoral and cross-institutional knowledge sharing in general.

Existing commercially available web-based computer assisted translation systems do not support such knowledge sharing in a broad sense; indeed, by their complexity they are counterproductive to that purpose. Nor do they support company-external, interdisciplinary knowledge sharing. The main strengths of the present project compared with proprietary systems like TRADOS and LOGOPORT therefore lie, firstly, in the inclusiveness that springs from its participatory and collaborative nature as opposed to the closed/restricted nature of the proprietary systems. Secondly, a main difference is the nature and quality of the data. The present system's core data consist of existing, approved language data from central and national drug agencies that are aligned and the specialized, validated texts from quality data owners in the field. The public and "private" data is constantly benchmarked against one another, so overall quality is presumed to be better than in purely proprietary systems.

We may argue that the present case has demonstrated that the creation of a knowledge cluster and collaborative text repositories has strengthened participating language service providers by creating a knowledge pool where they may leverage subject matter expert knowledge; indeed, all contacted language service providers working within the domain joined the cluster.

Conclusion

[11] If texts are not written in conformity with a standard set by the EMEA or the regional drug agency, marketing authorisation is withheld

The creation of an trans- and interdisciplinary architecture of cooperation and participation supported by a user-friendly, dynamic web-based technology platform with a collaboratively established community of practice language data and a two-tier validation structure may facilitate specialized knowledge communication in general and may raise translation productivity and quality in particular.

References

Askehave, I & Kastberg, P, 2001: Intergeneric derivation: On the genealogy of an LSP text. *Text* 21(4), pp 489-513.

Askehave, I & Zethsen, KK, 2008: Mandatory genres: the case of European Public Assessment Report (EPAR) Summaries. *Text & Talk* 28-2, pp 167-191.

Bazerman, C, 1994 Systems of genres and the enactment of social intentions. In A Freedman & P Medway (eds.). *Genre and the New Rhetoric*, London, Taylor & Francis Ltd., pp 79-101.

Bazerman, CH, 1999: Introduction: changing regularities of genre. *IEEE Transactions on Professional Communication.* Vol 42(1), pp 1-3.

Berkenkotter, C & Huckin, TN, 1995: *Genre knowledge in disciplinary communication: Cognition/culture/power.* Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Bhatia, V, 2004. *Worlds of Written Discourse: A Genre-Based View*. London & New York. Continuum International Publishing Group.

Bhatia, V, 1995: Genre-mixing in professional communication. The case of private intentions v. socially recognized purposes. In *Explorations in English for Professional Communication,* P Bhruthiaux, T. Boswood and B. Du-Babcock (eds), 1-19, Hong Kong: University of Hong Kong.

Bhatia, V, 1997a: Genre-mixing in academic introductions. *English for Specific Purposes* 16(3): 181-195.

Bhatia, V, 1997b: Translating legal genres. In Anna Trosborg (ed.) *Text Typology and Translation,* Amsterdam, John Benjamins.

Borgman, CL, 2009: Towards a definition of user friendly: A psychological perspective. University of Illinois. USA. 1986. *Clinic on Library Applications of Data Processing 23rd* *https://www.ideals.uiuc.edu/bitstream/handle/2142/758/Borgman.pdf?sequence=2* (10 June 2009).

Fairclough, N, 1993: Critical discourse analysis and the marketization of public discourse: The universities. *Discourse and Society* 4(2):133-168.

Gulati, R, 1995: Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances. *Academy of Management Journal,* 38 (February 1995), 85-112.

Halevy, A, Norvig, P & Pereira, P, 2009: The unreasonable effectiveness of data. IEEE Computer Society in their March/April journal.
http://www.computer.org/portal/cms_docs_intelligent/intelligent/homepage/2009/x2exp.pdf (6 June 2009)

IHTSO, 2009: The International Health Terminology Standard Organization. http://www.ihtsdo.org/about-ihtsdo/collaborative-space/ (10 July 2009).

Izquierdo, IG 2008: A multidisciplinary approach to specialized writing and translation using a genre based multilingual corpus of specialized texts. *LSP & Professional Communication.* 8(1): 39-63.

Ikujiro, N & Takehuchi, H, 1995. *The Knowledge-Creating Company,* Oxford University Press, EEUU.

Ketels, C, 2003: The development of the cluster concept – present experiences and further developments. Harward Business School. http://www.isc.hbs.edu/pdf/Frontiers_of_Cluster_Research_2003.11.23.pdf (5 June 2009).

Linell, P, 1998: Discourse across boundaries: On recontextualizations and the blending of voices in professional discourse. *Text* 18(2): 143-157.

Nonaka, I. & Takeuchi, H, 1995. *The knowledge-creating company; how Japanese companies create the dynamics of innovation*, New York; Oxford University Press.

Pilegaard, M, 1997. Translation of medical research articles. In Anna Trosborg (ed.) *Text Typology and Translation*. Amsterdam and Philadelphia: John Benjamins, 159-184.

Pilegaard, M, 2007: Value added lies in sharing. *Language at Work*, vol. 3, 2007.

Pilegaard, M & Baden, H, 2004: Medicinsk-odontologisk Ordbog. Dansk-engelsk, engelsk-dansk. København. Gyldendal. Nyt Nordisk Forlag.

Resurrecció, VM & Davies, GM, 2007: *Medical Translation Step by Step.* St. Jerome Publishing, Manchester.

Stahl, G, 2009: A model of collaborative knowledge building. *Proceedings of ICLS.* Barry J Fishman & Samuel F O'Connor-Divelbiss (http://www.google.com/books?hl=da&lr=&id=cDyHSJhkC0gC&oi=fnd&pg=PA70&dq=collaborative+knowledge+repository+theory&ots=mAcIx3NMLw&sig=YwMN7CGnkz0w2JJUr58_L1fRWF8) 9 July 2009.

Swales, J, 1990: *Genre Analysis. English in Academic and Research Settings.* Cambridge, Cambridge University Press.

SST, 2009. Sundhedsstyrelsen. SUNDTERM-projektet. En samlet dansk sundhedsterminologi – et sprog for klinisk dokumentation i EPJ. Sundhedsstyrelsen. København. http://www.sundhedsstyrelsen.dk/Indberetning%20og%20statistik/Terminologi/SNOMED_CT/Sundterm_pilot.aspx (10 juni 20009).

TAUS, 2009a: Translation Automation User Society: Roadmap for innovation and interoperability in translation industry. http://www.translationautomation.com/images/stories/downloads/taus-innovation-roadmap-may-2009.pdf (18 June 2009).

TAUS, 2009b: Translation Automation User Society TAUS Releases First Version of Translation Memory Data Exchange
Donald A. DePalma 5 June 2009. Filed under (Translation Technologies)
http://www.globalwatchtower.com/2009/06/05/taus-tda-v1/ (11 June 2009).

Trosborg, A, 1997. *Text Typology and Translation*. Amsterdam and Philadelphia: John Benjamins.

Wenger, E. *Communities of Practices, learning, meaning and identity*, Cambridge University Press, 1998.

Wiig, KM, (1999) "Introducing knowledge management into the enterprise", in: *Knowledge management handbook*, edited by J. Liebowitz. pp. 3.1-3.41. NY: CRC Press.

Wildsmith, J, 2003. How to write a case report. In G.M.Hall (ed.) *How to Write a Paper.* London: British Medical Journal Books, 85-91.

Wright, P, 1999. Writing and information design of healthcare materials. In Christopher N Candlin and Ken Hyland (eds.). *Writing: Texts, Processes and Practices,* London and New York: Longman 85-98.