World Class Language Technology - The Process of developing a Language Technology Strategy for Danish

Sabine Kirchmeier[1], Philip Diderichsen[2], Peter Juel Henrichsen[3], Sanni Nimb[4], Bolette S. Pedersen[2]

[1]Kirchmeier.dk, [2]Department of Nordic Studies and Linguistics, University of Copenhagen, [3]Danish Language Council
[3]Society for Danish Language and Literature,
[1]sabine@kirchmeier.dk, [2]{cph, bspedersen}@hum.ku.dk, [3]pjh@dsn.dk, [4]sn@dsl.dk

Abstract

Although Denmark is one of the most digitized countries in Europe, no coordinated efforts have been made in recent years to support the Danish language regarding language technology and artificial intelligence. In March 2019, however, the Danish government adopted a new, ambitious strategy for LT and artificial intelligence. In this paper, we describe the process behind the development of the language-related parts of the strategy: A Danish Language Technology Committee was constituted and a comprehensive series of workshops were organised in which users, suppliers, developers, and researchers gave their valuable input based on their experiences. We describe how, based on this experience, the focus areas and recommendations for the LT strategy were established, and which steps are currently taken in order to put the strategy into practice.

Keywords: Danish, language strategy, language policy

1. LT in Denmark

Denmark is one of the most highly digitised societies in Europe. Several ambitious governmental IT strategies during the last ten years have led to the digitization of most processes in the public space. (Digital Strategy 2016-2020). All citizens have a digital secure login to public services and other services. All citizens and companies are obliged to regularly consult their government-provided virtual mailbox (the so-called e-Boks, cf. eboks.dk) while paper-based correspondence is no longer supported by the public sector. Schools are obliged to provide educational services online, and all public service providers must serve the citizens via online interfaces. Central recommendations have been made for the development of IT systems for state institutions. Here, the development of clear terminology and semantic descriptions of metadata and content play a crucial role.

However, governmental initiatives for artificial intelligence and especially for LT have been rather scarce. In the META NET white papers (Pedersen et al. 2012), Danish was among the lowest ranked languages in Europe, and even much smaller language communities such as Iceland and Latvia have invested more in LT in recent years.

This is mainly due to five factors:

1. The lack of freely available linguistic resources
2. The specific characteristics of the Danish language
3. The small size of Denmark, both as a language community and as a market (5.8 mill. citizens).
4. Lack of coordination in the development, distribution and use of Danish LT
5. Too modest investment in research and training in Danish LT in recent years.

2. Existing LT for Danish

Danish differs markedly from English both in the way words and sentences are formed, and in the way, words are pronounced. Danish is exceptionally vowel rich (it has three times more vowels than English in IPA-counting). Danish pronunciation is characterized by a lot of phonetic reductions - which foreigners often experience as mumbling. This makes the Danish language difficult to understand and produce – not least for speech technology applications.

Over the years, a number of language databases and tools for Danish have been created based on public and private initiatives, but the linguistic resources for Danish are still somewhat scattered and efforts have not been sufficiently coordinated. In particular, insufficient attention has been paid to the fact that these building blocks, often costly to produce, generally need upscaling, validation, and maintenance, and last but not least that they should be made freely available. This means that many data sets are limited or not available at all.

Speech technology is currently the most discussed technology. There are a few international and Danish technology providers, but most of them are based on Nuance proprietary technology, and there are no freely available speech corpora of sufficient size and quality for developing state-of-the-art speech recognizers and synthetic voices.

Machine translation is supplied by international providers such as Microsoft and Google, and no Danish MT provider has been entering the market so far.

A Danish FrameNet, a semantically annotated corpus, a Danish wordnet and several other datasets, wordlists, corpora and tools exist (for details, see sprogtek2018). Common for many of these collections, however, is the fact that they are neither complete nor validated to a sufficient degree since they are typically developed under projects with limited budget and duration.

Only a few smaller companies offer tailored services such as semantic search, taxonomy based indexation and the development of chat bot services.

3. Developments in Knowledge Modelling, Terminology and Public IT Architecture

In a highly digitized society, the need for consistent IT systems that smoothly support human workflows, for instance in relation to electronic patient records or the document management systems of municipalities, is of vital importance.

Over the years, thousands of public IT systems have been developed, but they are not communicating in a common language, and there has not previously been a common public plan for how IT systems securely and efficiently can exchange data and become part of a coherent process.

Therefore, the government, municipalities and regions have agreed that, as part of the public digitization strategy for 2016-20, a common public architecture for secure and efficient data sharing and the development of processes that connect public services must be established.

A number of general principles have been developed that all publicly funded IT projects must adhere to:

1. IT architecture must be managed at the right level according to a common framework
2. IT architecture must promote coherence, innovation and efficiency
3. IT architecture and legal provisions must support each other
4. Security, privacy and trust must be ensured
5. Processes must be optimized across different fields
6. Good data must be shared and reused
7. IT solutions must collaborate effectively
8. Data and services must be delivered reliably.

In particular, principles 3 and 6 point to the need to create a common language and a clear understanding of concepts and the importance of data sharing. A uniform and consistent description of concepts, both in legislation and IT architecture, is a cornerstone of the Danish IT architecture principles emphasising the need for a closer integration of terminology, concept modelling and semantic description in the public sector.

4. Academic and Political Initiatives to promote LT

Over the last 10 years, several academic and political initiatives were taken in order to promote LT for Danish. Already in 2009, the Danish Language Committee under the Minister of Culture stressed the need for more research into and development of LT in its report Sprog til tiden (2009).

Even more strongly, the report of the Danish Language Council, on the status of the Danish language, Dansk sprogs status (2012), stressed the importance of a national plan for the development of LT and a national terminology and knowledge bank.

The status report and the findings of the META Net project, which were published the same year, and a number of follow-up articles in the daily newspapers, finally lead to a proposal for more funding for Danish LT and a terminology bank by the Danish People's Party in the Danish parliament. However, as the proposal also contained other language measures, such as restrictions on the use of English at Danish Universities and the naming of public institutions, which were unacceptable for the other parties in parliament, the entire proposal was rejected.

During 2015, the Danish Language Council published a number of articles explaining the principles behind a number of LT applications such as spell checkers, speech recognition, MT and concept modelling in relation to terminology (Diderichsen 2015-1; Diderichsen & Kirkedal 2015; Diderichsen 2015-2; Diderichsen 2016; Madsen & Hoffmann 2016). Together with leading researchers in the field, another article was published in the newspapers illustrating the problems that Denmark would face, if no measures were taken to improve Danish LT (Diderichsen et al. 2016).

This led to a new proposal in parliament by the Danish People's Party in 2016: to establish a Danish language committee. The proposal was well received, but not immediately accepted. Instead, it was decided to call a public hearing about LT in the parliaments committee of culture in the beginning of 2017. This hearing, finally, led to the establishment of a language technology committee under the Ministry of Culture led by the Danish Language Council in the beginning of 2018.

5. The Work of the Danish Language Technology Committee

5.1 Organisation and Composition

The Language Technology Committee was established by the Minister of Culture and organised by the Danish Language Council. The work started on January 1st 2018. The purpose of the committee was to clarify the perspectives and challenges of LT in a Danish context and to produce a set of recommendations on how to stimulate research and development in Danish LT as used in artificial intelligence, teaching, public services, and related areas. The committee was also required to investigate the perspectives for a national term bank.

The committee consisted of 14 members representing

- Large international IT companies (IBM and Google), Danish LT companies (Ankiro, Dictus, MIRSK and Unsilo)
- User representatives (the municipalities of Odense and Ballerup)
- Researchers and language resource developers (the Centre for Language Technology at Copenhagen University, the DANTERM Centre at Copenhagen Business School and the Danish Society for Language and Literature)
- Public stakeholders (the Danish Digitization Agency and the Danish Language Council).

This composition was chosen to ensure the largest possible support and accept of the conclusions of the committee.

5.2 Working Method

The committee decided to assess the current situation for Danish LT from four different perspectives:

1. LT users groups/professional LT users
2. LT suppliers/vendors
3. LT engineers/developers
4. Researchers and teachers in the field of LT and AI.

The main instrument for collecting knowledge was a series of workshops focusing on the four perspectives, complemented by 2 workshops on specific areas, machine translation and terminology. Prior to each workshop, the committee sent out a questionnaire to stakeholders identified by the committee members. Through structured discussion papers based on the results of the questionnaires the workshop participants (20-30 participants in each workshop) identified the problem areas and proposals for the most effective actions.

A final seminar was held in January 2019, allowing all workshop participants to discuss the recommendations of the committee before publication.

In total, representatives of about 120 different Danish companies, organisations, public authorities and educational institutions contributed to the committee's investigation.

5.3 Dissemination and Communication

The committee's work was disseminated through a blog, sprogtek2018.dk. A short video was produced explaining the basic facts about the work of the committee and distributed via social media.

The committee maintained close contacts with Nordic and European organisations (ASTIN/Nordic Language Councils, ELRC, META-NET, EFNIL) to share knowledge across borders and to explore opportunities for cooperation.

5.4 The Danish LT situation as uncovered by the Committee

The participants at the 6 workshops contributed substantially to a detailed picture of the situation for Danish LT.

End users, especially public sector organisations, expressed that they have little influence on the design and implementation of the basic LT products (e.g. speech recognition and translation tools), often developed by international companies. Many indicated that they provided language data for development purposes, but that they have no access to their own data once incorporated into a product. This leads to a high level of dependence on the providers and restricts the ability of end-users to significantly improve the performance of the service. End users also reported problems with the quality of the systems and, as a consequence, difficulties in making effective use of the products.

The suppliers, both Danish and foreign companies, describe the market as small but well organised (in particular the public sector) and with proximity to end users. It is considered an advantage that customers in state institutions and municipalities often participate actively in the development work and in securing project funding.

Among the threats is the fact that small businesses are often out of choice as suppliers in public tenders, particularly affecting companies with their main customer base in Denmark. The lack of qualified staff (Danish LT specialists) is also becoming more and more evident.

The main problem is, however, that all companies, regardless of their size, need to invest substantial amounts in collecting and developing the basic linguistic resources for Danish as the foundation for their products. Many indicate that it would encourage the development of new products and enhance quality significantly if a set of basic resources of high quality was freely accessible and maintained on a continuous basis.

The developers, typically employees in Danish and international companies and Danish research institutions, also stress the importance of free access to a Danish language repository, including text and speech data free from restrictions such as copyright and limitations due to GDPR regulations. They also point out the lack of access to structured word databases with semantic information on the Danish vocabulary. The developers suggest the establishment of an independent service organisation that can continuously develop, distribute and maintain language resources and disseminate knowledge and provide training and guidance on methods and technologies for handling the Danish language - i.e., a language bank.

Scientists in particular highlight the lack of language engineers with strong expertise in the Danish language and thus stressed the need for more education and training. They also express the need for an organisation that can plan the development of new basic resources, the education of experts and the dissemination of language technology resources and products in the Danish society (again: a language bank)

Professional MT/TM users stress the fact that it is necessary in particular for public institutions to raise awareness of the value of linguistic data, such as translated texts and terminology databases or lists, for AI projects and for the development of LT for Danish. Shareable data should be identified and freely shared, both for language technological applications and for AI purposes.

Public institutions are seen as crucial players in order, for example, to improve the EU translation programme eTranslation, which is accessible free of charge to all public institutions in Europe. Several Danish public agencies, however, have recently outsourced their translation assignments, making it more difficult to reuse translation memories.

Users of Danish terminology in the public and private sectors unanimously point to the need for better coordination of the national terminology in Denmark by setting up a national term bank. Although most of the terminology work takes place in relation to communication and translation projects, the language technology committee's investigations have shown that about 20 % of the terminology work is carried out in the context of IT system development and the digitisation of legislation. There is not only a need for terms, definitions and translations, but also for structured knowledge about the relations between terms.

5.5 Assessment of Danish Language Resources

During the work of the committee, 124 language resources for Danish could be identified (Danish Language Resources 2019).

This overview is, to the best of our knowledge, the most comprehensive list of Danish language resources for NLP development compiled to date. It is a critical observation, then, that very few of the items are at all relevant for serious

NLP development. For instance, not a single source[1] possesses all of these properties: (*i*) unrestricted access (CC0 license or freer), (*ii*) up-to-date data (≤10 years), (*iii*) significant size (100M+ text words *or* 100h+ spoken words *or* 100k entries for dictionaries); (*iv*) high-quality annotation and metadata; (*v*) platform-independent format (XML or similar). Danish start-ups wanting to develop TTS or ASR product lines thus look in vain for existing language resources to begin from.

Furthermore, these 124 resources are spread over 72 different web sites and portals, some of them with different versions and copies in several places.

5.6 Recommendations

On these grounds, the committee proposed:

1. The creation of an organisation that coordinates the development of resources for Danish LT;
2. The creation of a Danish language bank that supports the development and maintenance of databases and software for LT and AI applications by making Danish language resources and tools freely available. This includes a Danish terminology bank.
3. University programs for Danish LT.
4. More research funding for Danish LT.

The Committee also suggested that public institutions should have more focus on making linguistic data available to ensure the use of Danish in all areas of society, and for instance, to make translated texts available to improve MT technologies. Further recommendations include a change in copyright legislation in favour of LT and AI applications.

5.7 Results

During 2018, a close contact between the Danish Agency for Digitization and the Language Technology Committee was established. The Danish government was preparing a new strategy for the digitization of the public sector and the first public AI strategy for Denmark. In both strategy development processes, LT was included.

In October 2018, the Danish government's new digitization strategy for the public sector contained a chapter on LT and a strategic goal to develop "world class language technology for Danish" (World Class Digital Service 2018).

In March 2019, the Danish government presented the first Danish national strategy for artificial intelligence (Denmark's National Strategy for Artificial Intelligence 2019). The strategy also included a chapter on LT and recommended the establishment of a Danish language resource repository. At the same time, the Danish government allocated 35 million DKK (4.7 million EURO) to the project.

The recommendations of the committee were finalized in January 2019. The committee report was launched in April 2019 (Kirchmeier et al. 2019).

6. Conclusions and Prospects for the Future

The Danish Agency for Digitization is commissioned with the task of implementing the recommendations of the committee as part of the national AI-strategy.

The project is currently at an early stage, focused on planning and organizing, and on delivering short term wins while still keeping a focus on the long-term agenda. The 35 million DKK are to be spent over a period of 6 years. However, there is a clear understanding that further investments will be necessary, that LT for Danish must be further developed and Danish language data continuously maintained.

The overall aim is to develop a technical platform containing various free open databases and software for use in the NLP industry.

The next steps include incorporating and upgrading existing Danish digital dictionaries, terminological and lexical resources, developing and implementing a time-coded general purpose Danish speech corpus, and identifying and developing new language resources and technologies in close cooperation with the LT and AI communities.

7. Bibliographical References

Dansk sprogs status (2012). Danish Language Council. https://dsn.dk/udgivelser/sprognaevnets-udgivelser/sprognaevnets-skriftserie-1/det-danske-sprogs-status/view

Denmark's National Strategy for Artificial Intelligence (2019). Danish Digitization Agency. https://en.digst.dk/policy-and-strategy/denmark-s-national-strategy-for-artificial-intelligence/

Diderichsen, P. (2015-1). Stavekontrol. *Nyt fra Sprognævnet*, (1), p. 5-8.

Diderichsen, Philip & Kirkedal, A. S. (2015). Talesyntese og talegenkendelse. *Nyt fra Sprognævnet*, (2), p. 6-11.

Diderichsen, P. (2015-2). Maskinoversættelse. *Nyt fra Sprognævnet*, (3), p. 8-15.

Diderichsen, P. (2016). Hverdagens sprogteknologi nu og i fremtiden. *Nyt fra Sprognævnet*, (1), 8-13.

Diderichsen, P., Henrichsen, P.J., Kirchmeier, S. & Pedersen, B. S. (2016, September 26th). *Sproget er digitaliseringens sorte guld*. Politiken.

Digital strategy 2016-2020. Danish Digitization Agency 2016. https://en.digst.dk/media/14143/ds_singlepage_uk_web.pdf

Kirchmeier, S., Henrichsen, P.J., Diderichsen, P. and Hansen, N. B. (2019). *Dansk Sprogteknologi i verdensklasse.* Rapport fra sprogteknologiudvalget under Dansk Sprognævn nedsat af Kulturministeriet. Dansk Sprognævn. https://dsn.dk/nyt/nyheder/2019/dansk-sprogteknologi-i-verdensklasse-rapport-fra-sprogteknologiudvalget/?searchterm=dansk%20sprogteknologi http://www.lr-coordination.eu/News/New-Report-on-Language-Technology-for-Danish

[1] with the possible exceptions of Europarl and Folketingstidende.

Madsen, B. N. & Hoffmann, P. L. (2015), Sprogteknologi til håndtering af fagsproglig viden. *Nyt fra Sprognævnet* (4). p. 1-6.

Pedersen, B. S, Wedekind, J., Kirchmeier-Andersen, S., Nimb, S., Rasmussen, J.E., Larsen, L.B., Bøhm-Andersen, S., Henriksen, P., Kjærum, J.O., Revsbech, P., Thomsen, H.E., Hoffensetz-Andresen, S. & Maegaard, B. (2012). *Det danske sprog i den digitale tidsalder*. In: Rehm, G. and Uszkoreit, H., (eds) META-NET White Paper Series: *Europe's Languages in the Digital Age.* Springer.

Sprog til tiden (2009). Danish Ministry of Culture. https://kum.dk/publikationer/2008/sprog-til-tiden-rapport-fra-sprogudvalget/

World Class Digital Service (2018). Danish Government. https://en.digst.dk/news/news-archive/2019/january/new-direction-for-reform-to-create-world-class-digital-services/

8. Language Resource References

Overview of Danish language resources (2019). Danish Language Council. https://dsn.dk/nyt/nyheder/2019/sprogteknologiudvalgets-resurseoversigt