Chapter 22
Conclusions and Outlook to the Future

Jan Odijk

22.1 Introduction

The preceding chapters have sketched the context of the STEVIN programme and detailed descriptions of some of the results of the scientific projects carried out in the STEVIN programme. In this chapter I will briefly and very globally describe the impact of the STEVIN programme as a whole to human language technology (HLT) for the Dutch language in the Low Countries (cf. Sect. 22.2). In Sect. 22.3 I will identify a number of research data and topics that are, despite the STEVIN programme, still insufficiently covered but needed. Here I will also take into account international developments in the field of HLT that are relevant in this context. In Sect. 22.4, I identify recent international trends with regard to HLT programmes, assess the position of the Dutch language, and the prospects for funding of programmes and projects that are natural successors to the STEVIN programme. I also make some suggestions to government administrations for future policy actions. In Sect. 22.5, I summarise the major conclusions of this chapter.

22.2 Results of the STEVIN Programme

In this section we briefly and globally discuss the results of the STEVIN programme and its impact on HLT for the Dutch language in the Low Countries. More details about the results of the STEVIN programme can be found in Chap. 1, page 1.

It is clear from this book that the main objectives of the STEVIN programme have been achieved. Firstly, an effective digital language infrastructure for Dutch,

J. Odijk (✉)
UiL-OTS, Trans 10, 3512 JK Utrecht, The Netherlands
e-mail: j.odijk@uu.nl

P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-30910-6_22, © The Author(s) 2013

based on the BaTaVo priorities[1] has been realised. An impressive amount of properly documented language resources (both data and software) has been created in the STEVIN programme, together with guidelines, conventions, and best practices for creating these and similar language resources. These language resources and their documentation are stored and maintained at the HLT Agency, and made available to all researchers and developers in a non-discriminative way, either via resource-specific licenses or under an Open Source License (cf. Chap. 21, page 381). The intellectual property rights (IPR) surrounding the language resources have been explicitly dealt with, so that the resources can actually be used by researchers and developers.

Second, strategic HLT research has been carried out in the STEVIN programme in a range of projects, selected, again, on the basis of the BaTaVo priorities.

Thirdly, the STEVIN programme has enormously stimulated network creation among players in HLT. There has been close cooperation between researchers from academia and industrial developers. There has also been close cooperation between partners from the Netherlands and partners from Flanders.[2] The STEVIN programme has consolidated the HLT activities in the Low Countries. It has created and supported a wide range of events where researchers, industrial developers, policy makers, potential users from industry and governments could meet, exchange ideas, and discuss important technical and policy issues. The STEVIN programme has educated new experts:

- Recently graduated students and post-docs as a side effect of the research and resource creation projects they were participating in;
- Students and potential students via educational activities;
- Policy makers and decision makers from government and companies via master classes.

STEVIN has also contributed significantly to the transfer of knowledge from academia to industry and vice-versa through the various projects in most of which academia and industry collaborated, as well as via several publications on the STEVIN programme that have appeared in journals published by ministries and the journal of the NOTaS[3] organisation DIXIT.[4]

The STEVIN work programme lists a number of potential (classes of) HLT applications as illustrations. The STEVIN projects have contributed to the realisation of such applications in many ways: in some cases indirectly, e.g. by creating resources required for the development of the technology underlying the application or by

[1] The STEVIN priorities have been listed in Chap. 1, Table 1.1 (page 2) and have been derived from the BaTaVo priorities [7]. 'BaTaVo' is an acronym standing for **Ba**sis**Ta**al**Vo**orzieningen (Basic Language Resources).

[2] Over 330 binary cooperation link occurrences in the STEVIN projects alone witness to the extent of collaboration in the STEVIN programme.

[3] NOTaS is a professional organisation for HLT in the Netherlands. See http://www.notas.nl/

[4] http://www.notas.nl/en/dixit.html

doing strategic research on underlying technologies; in other cases more directly by carrying out application oriented research. Some projects dealt with the creation of the application itself, e.g. in demonstration projects. Multiple STEVIN projects have contributed to the priority example applications, which were: information extraction from speech, detection of accent and identity of speakers, extraction of information from (monolingual or multilingual) text, semantic web, automatic summarisation and text generation, automatic translation, and educational systems.

The STEVIN programme has been evaluated halfway and at the end of the programme by external experts. In both cases the evaluations were very positive [13, 15].

Summarising, it can be concluded that the STEVIN programme has been very successful and has largely achieved its objectives in an excellent way.

22.3 Desiderata for the Near Future

The STEVIN programma has largely achieved its objectives, as described in the preceding section. However, this does not mean that the field of HLT for Dutch is now fully covered.

Firstly, there are a number of topics that have not been covered at all or only to a limited degree within STEVIN. This includes corpora for and research into multimedia, and corpora for and research into speech synthesis. For these areas, this was in part intentional. STEVIN attempted to avoid overlap with the concurrently running IMIX programme,[5] which covers some aspects of multimedia. But even with IMIX, resources for multimedia are largely absent and research into it has been very limited.[6] Research into speech synthesis was considered not so useful because the state-of-the-art systems at the time were closed commercial systems.[7] Semantic analysis and semantic corpora were part of STEVIN (inter alia in D-Coi and SoNaR, cf. Chap. 13, page 219), but were represented there only to a very small degree. The lexical semantic database Cornetto (cf. Chap. 10, page 165) created in the STEVIN programme is evidently relevant for applications related to the semantic web, but no application-oriented research project had the semantic web as its focus. Annotation of discourse and rhetoric relations was not completely absent in STEVIN, but was not addressed in a systematic manner or on a sufficiently large scale. Morphological analysis of derivation and compounding is lacking completely. Corpora for and research into robust speech recognition were well represented in STEVIN, but, of course, not all problems of robust speech recognition are solved with it. There

[5] http://www.nwo.nl/imix

[6] Some (Dutch) organisations (TNO, Twente) were involved in the European FP6 projects AMI and AMIDA, which deal with multimedia. See http://www.amiproject.org/

[7] The work done in the Autonomata project (Chap. 4, page 61), however, is surely relevant to speech synthesis.

are many situations were adverse conditions make automatic speech recognition a challenge, and only some of them were addressed in STEVIN. And this also holds for other research areas. Because of limitations of budget and time, many of these areas could be covered only in part. The large number of excellent project proposals (that would qualify for funding if there were enough money) witness to the fact that there are still many areas in which excellent research can be carried out that unfortunately could not take place in the STEVIN programme.

Secondly, STEVIN has, by its success, yielded new data that enable researchers to address existing research questions in a better way (e.g. in a more data-intensive way). STEVIN also made it possible to address completely new research questions. Though many of the newly created insights, data and software have already been used (sometimes in preliminary versions) by other projects in the STEVIN programme, the complete data sets of e.g. SoNaR (Chap. 13, page 219) and Lassy (Chap. 9, page 147) have become available only in a rather late stage in the programme. Therefore, the potential that they create for research in HLT as well as for other fields (e.g. research in various subdisciplines of linguistics) has hardly been exploited. A critical remark on the research carried out in STEVIN has been that it was largely incremental in nature, and that in some projects the research replicated earlier research but now for Dutch [6]. This may well be true, and is not unexpected given the nature of the STEVIN programme. But STEVIN has prepared the grounds for ground-breaking and cutting edge research in HLT, so that new research programmes and projects are now needed to tap into this potential and optimally exploit the rich resources of data and software that STEVIN has yielded.

Thirdly, significant developments on an international level have occurred as well. Firstly, IBM has shown, with its *DeepQA* approach, that language technology can be used to robustly extract precise answers to specific questions from a mix of structured (e.g. databases) and unstructured data (e.g. texts on the world wide web).[8] Though IBM demonstrated the capabilities of the Watson DeepQA in a game context (the Jeopardy! Game), it is now setting up systems based on the same principles for commercial applications (for example in the medical domain). With this, IBM has set the extraction of precise answers from unstructured data central on the HLT agenda. Analysis of unstructured documents is also needed in the context of business intelligence, e.g. in determining the perception of a company's public image and of specific commercial products and services by clients and prospects on the basis of an analysis of unstructured text resources in modern social media such as blogs, product review web sites, Twitter, etc. This topic is of great importance for big companies such as Philips and SAP [12]. They currently carry out such analysises still largely manually, but that is becoming infeasible with the exponentially growing digital content. The same techniques are also needed and increasingly demanded in the area of humanities research: the sharp increase in the number of available digital documents and digital audiovisual material makes it impossible to work in the traditional manner. The analysis of large document collections and audiovisual

[8] http://www.research.ibm.com/deepqa/deepqa.shtml

materials requires the use of sophisticated HLT as auxiliary tools for research into linguistics, literature, history, culture, and political sciences. In the area of speech dialogue, Apple's SIRI on iPhone[9] has brought speech dialogues to a new level. This is only possible thanks to two factors: firstly, the speech recogniser contained in it is extremely robust to background noise of a large variety of environments, to make its use on a mobile phone feasible. Secondly, SIRI maximally uses contextual information (largely available on the owner's iPhone) to interpret (usually ambiguous and underspecified) utterances spoken by the phone owner in the dialogue.

STEVIN has not specifically addressed the issue of Question Answering on the basis of unstructured data, though it has prepared many components (e.g. the tools used and further developed to create the SoNaR and Lassy corpora; some components developed in the DAISY (Chap. 19, page 339) and DAESO (Chap. 8, page 129) projects). Within STEVIN there was some research in the area of opinion mining (in the DuOMAn project, Chap. 20, page 359), but the large industrial interest justifies investing more in this area. Robust speech recognition in adverse conditions has explicitly been addressed in STEVIN, but as stated above, surely not exhausted. Research into spoken dialogue and a maximal use of context in interpreting the utterances and guiding the dialogue has been largely absent in STEVIN, but surely deserves more attention.

In short, with the STEVIN programme finished, the Dutch and Flemish HLT researchers are in an excellent position to deepen the research and to extend it to new areas, some of which are of big importance to industry and other scientific areas such as the humanities. They look forward to new opportunities to maximally exploit the insights gained and the materials created in STEVIN in new research programmes and projects.

22.4 Future

Unfortunately, the prospects for funding a successor project to STEVIN are grim. In the Netherlands, consultation meetings with the NOTaS organisation have been held,[10] and in Flanders a round table meeting with some 40 players from the field [11]. A policy document sketching the outlines of a research programme in the area of the extraction of information from unstructured data (both textual and audiovisual) is available [2]. But obtaining funding for such a research programme is not easy. There is firstly the fact that the funding opportunities have to come both from the Netherlands and from Flanders. However, the priorities of the two governments differ, the instruments are different and the timing is difficult to synchronise. Furthermore, in the Netherlands there are no opportunities for

[9] http://www.apple.com/iphone/features/siri.html

[10] See e.g. http://taalunieversum.org/taal/technologie/taalinbedrijf/documenten/notas.pdf

discipline-specific research programmes. All research is organised via what are called *Top Sectors*, a series of sectors identified by the Dutch government as specifically important and with high potential for the Dutch economy. HLT research fits in very well in the Creative Industry Top Sector [14, 16]. For example, HLT can obviously be used in many applications in the areas of social media, publishers, TV and radio, gaming, museums and cultural heritage, and in mobile applications. But any research programme in this top sector will cover multiple disciplines in which HLT must try to find its place. In Flanders, the situation is different but [11] also concludes that opportunities must be sought to embed HLT research in broader initiatives. One possibility is to focus research and development effort around an integrated demonstrator that requires research into and development of technology from multiple disciplines, HLT being one of them.

In order to improve the chances of obtaining funding, the visibility and strength of the HLT sector can and must be further improved. Some HLT organisations in the Netherlands are united in NOTaS but certainly not all. Though CLIF[11] unites the scientific HLT community in Flanders, [11] argues in favour of the creation of a structure that unites and promotes the whole HLT sector.[12]

On the other hand, there are other developments that are directly relevant and make one more optimistic.

Firstly, some researchers from linguistics and HLT, in particular from the Netherlands, have, for several years now, been arguing for the need of setting up a distributed technical research infrastructure for humanities research. These efforts have led to a proposal for such an infrastructure called CLARIN (Common Language Resources and Technology Infrastructure). CLARIN has been put on the ESFRI roadmap in 2006. The ESFRI-funded European CLARIN preparatory project has been successfully executed.[13] CLARIN has been put on the national Dutch roadmap for large infrastructures in 2008. Since 2009 the national research infrastructure project CLARIN-NL[14] is running. The targeted research infrastructure will contain, inter alia, a range of HLT data and tools. These data and tools must be adapted to make them user friendly and easy to use for humanities researchers without an HLT background. Though CLARIN in Flanders has so far only been awarded funding for preparatory activities, a modest budget from Flanders could be secured for cooperation with the Netherlands. Together with funding from the CLARIN-NL project, a small-scale cooperation project between the Netherlands and Flanders could be set up to make the tools developed in the STEVIN programme (especially the ones created or extended in the D-Coi, SoNaR and Lassy projects) cooperate seamlessly with each other as web services in a work flow system. This project, called TTNWW, runs from 2010 to 2012.[15] A decision on larger scale

[11] Computational Linguistics in Flanders, http://clif.esat.kuleuven.be/

[12] The need for this is also felt at the European level, see below.

[13] http://www.clarin.eu

[14] http://www.clarin.nl

[15] More information on this project can be found here: http://www.clarin.nl/node/76#TTNWW

activities for CLARIN in Flanders is expected in the course of 2012. If that decision would be positive, it might open up new opportunities for collaboration between the Netherlands and Flanders in HLT, though the focus will be on applying HLT in a research infrastructure for humanities researchers, but not on research into HLT itself.

Secondly, a range of projects is working on the research agenda for HLT at the European level. For example, the FLaReNet project[16] has consulted the HLT community in Europe and beyond to formulate recommendations for the policy of the European Union with regard to language resources. It has resulted, inter alia, in the *FLaReNet Strategic Language Resource Agenda* [5] and the FLaReNet recommendations for language resources [4].

Of particular importance in this respect is the fact that the European Union keeps expanding, and is becoming increasingly more multilingual. The multilinguality of Europe is on the one hand considered a valuable cultural asset. On the other hand it also is a burden because it makes communication more difficult and especially more costly. The European Commission is seeking ways of reducing these costs. HLT has the potential to significantly reduce the costs for Europe's multilinguality and even to turn this into an economic asset. Google has already proved that large-scale machine translation for several tens of language pairs is feasible for certain applications and services where the translation quality is of secondary importance.[17] Several policy makers in the European Commission believe that the HLT community in Europe has the potential to improve machine translation significantly, so that it becomes useful for applications where translation quality does matter. However, this is only possible if the research community is united, joins forces in a common strategic research agenda, and receives sufficient means to carry out ground-breaking research. These are pre-conditions for achieving the goals set and to avoid dependency on foreign commercial companies. Several projects in the EU ICT programme have been started up (arguably in part as a result of the FLaReNet project) to work towards such a situation, which hopefully can become part of Europe's Horizon 2020 Programme. These projects include META-NET[18] and various related projects such as META-NORD,[19] CESAR,[20] and METANET4U.[21] The META-NET project is carrying out some research, in particular it is building bridges to relevant neighbouring technology fields via its META-RESEARCH activities. But even more importantly, the META-VISION part of the project has consulted a wide range of players in the field, including researchers, commercial and non-commercial HLT developers, commercial users of HLT, language professionals, and others, to inventory needed and desired functionality, important research areas, commercial

[16] http://www.flarenet.eu/

[17] http://translate.google.com/

[18] http://www.meta-net.eu/

[19] http://www.meta-nord.eu/

[20] http://www.cesarproject.eu/

[21] http://metanet4u.eu/

potential for HLT, etc., to develop a vision on the future of HLT for the next decade. This vision has been created and laid down in a vision paper [8]. Currently, this vision is being developed into a strategic research agenda. At the same time, META-NET has assessed the status of HLT for the European Union languages, and it has described properties of each individual language that pose specific challenges to HLT for that language. This has resulted in an impressive range of *language white papers*, preliminary versions of which are already available,[22] including one for Dutch [9]. Not surprisingly, the Dutch language scores very well here, and plays in the same league as big European languages such as French and German. For a large part, the STEVIN programme is to be credited for this.[23]

The META-NET project is also working on improving the pre-conditions for carrying out excellent research and efficient technology development. In particular it aims at facilitating the sharing and exchange of language resources via the open distributed META-SHARE language resource exchange facility [10].[24] Obviously, though META-SHARE has a different goal and a different target group, there are commonalities with the CLARIN infrastructure, there is close collaboration between the two projects, and shared use of certain technology (e.g. both make use of the CMDI framework for metadata[25]). In this context it is also important to see how the status of the Dutch HLT Agency is developing. It is natural that it would develop into a data centre not only in the CLARIN infrastructure (which it is already working towards) but also in META-SHARE.

In these European developments, one can discern a parallel with the developments in the Netherlands and Flanders 10 years ago: the language white papers can be considered as the equivalent of the BaTaVo report [7] for the Dutch language, but now for all languages of Europe and with a special focus on multi- and cross-linguality; the META-SHARE facility that is being prepared can be considered the equivalent of the HLT Agency, though again now on a European scale; and the META-NET vision paper and the strategic research agenda that is being developed correspond to the policy documents made in the Netherlands (e.g. [1]) that have contributed to the positive decision to fund the STEVIN programme.[26]

It remains to be seen whether these efforts will indeed lead to a common strategic research agenda and sufficient funding to carry out the research and development that will be necessary to execute this research agenda. However, the fact that these efforts are being made turn one optimistic in believing that it will create

[22] http://www.meta-net.eu/whitepapers

[23] The final versions of the language white papers will be come available Mid 2012.

[24] http://www.meta-net.eu/meta-share

[25] Component-based MetaData Infrastructure [3].

[26] Though these parallels are real, I do not intend to claim that the European developments have been inspired completely by the developments in the Netherlands. Other projects, e.g. Euromap (http://www.2020-horizon.com/EUROPEAN-OPPORTUNITY-MAPPING-(EUROMAP)-s50899.html), have undertaken similar activities at the European level already more than 10 years ago, for the situation at that time.

opportunities for European HLT researchers in general, and Dutch and Flemish researchers in particular.

22.5 Concluding Remarks

I briefly summarise the current situation as sketched in this chapter. It is very unlikely that there will be a successor programme that is similar in nature to the STEVIN programme. At the moment, there are also no concrete opportunities for such a programme. Nevertheless, there is at least one concrete example where funding has been obtained for an (admittedly small-scale) successor project (CLARIN TTNWW). In addition, there are many opportunities for carrying out research in HLT, not only in the Netherlands and Flanders (separately), but also at the European level. However, these are mainly opportunities for individual projects, not for programmes. The future will tell whether these opportunities are real and will materialise into concrete cutting edge HLT research projects.

Acknowledgements I would like to thank Peter Spyns and an anonymous reviewer for valuable comments on an earlier version of this chapter.

References

1. Akkermans, J., van Berkel, B., Frowein, C., van Groos, L., Compernolle, D.V.: Technologieverkenning Nederlandstalige taal- en spraaktechnologie. Report, Ministry of Economic Affairs, The Hague (2004)
2. Boves, L.: Enterprise language processing: Een aanzet voor een nieuw programma. Report, Nederlandse Taalunie, The Hague (2011)
3. Broeder, D., Kemps-Snijders, M., Uytvanck, D.V., Windhouwer, M., Withers, P., Wittenburg, P., Zinn, C.: A data category registry- and component-based metadata framework. In: Calzolari, N., Maegaard, B., Mariani, J., Odijk, J., Choukri, K., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), pp. 43–47. European Language Resources Association (ELRA), Valetta (2010)
4. Calzolari, N., Bel, N., Choukri, K., Mariani, J., Monachini, M., Odijk, J., Piperidis, S., Quochi, V., Soria, C.: Language Resources for the Future – The Future of Language Resources. CNR – Istituto di Linguistica Computazionale A. Zampolli, Pisa (2011). http://www.flarenet.eu/sites/default/files/FLaReNet_Book.pdf. FLaReNet Final Deliverable
5. Calzolari, N., Quochi, V., Soria, C.: The FLaReNet Strategic Language Resource Agenda. CNR – Istituto di Linguistica Computazionale A. Zampolli, Pisa, Italy (2011). http://www.flarenet.eu/sites/default/files/FLaReNet_Strategic_Language_Resource_Agenda.pdf

6. Daelemans, W.: What did STEVIN do for HLT research? http://taalunieversum.org/taal/technologie/stevin/documenten/stevin_results_r_28112011.pdf (2011). Presentation held at the STEVIN Final Event, Rotterdam, the Netherlands
7. Daelemans, W., Strik, H.: Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basisvoorzieningen. een rapport in opdracht van de Taalunie. Report, Nederlandse Taalunie, The Hague (2002). http://taalunieversum.org/taal/technologie/docs/daelemans-strik.pdf
8. META-NET Consortium: The future European multilingual information society. Vision paper for a strategic research agenda. META-NET report, DFKI, Berlin (2011). http://www.meta-net.eu/vision/index_html/reports/meta-net-vision-paper.pdf
9. Odijk, J.: Languages in the European information society – Dutch (early release edition). META-NET white paper series, META-NET, Berlin (2011). http://www.meta-net.eu/whitepapers/download/meta-net-languagewhitepaper-dutch.pdf
10. Piperidis, S.: META-SHARE: An open resource exchange infrastructure for stimulating research and innovation. Presentation held at METAFORUM 2011, Budapest, Hungary (2011) http://www.mt-archive.info/META-FORUM-2011-Piperidis.pdf
11. Spyns, P.: TST-rondetafel 07/09/2011 – STEVIN-roadmapworkshop. Report, EWI, Brussels, Belgium (2012). Version 4 April 2012 http://http://taalunieversum.org/taal/technologie/stevin/vl_rondetafel/
12. Taal in Bedrijf: Panel on business intelligence. http://taalunieversum.org/taal/technologie/taalinbedrijf/programma_2011/ (2011)
13. Technopolis$_{|group|}$: Eindevaluatie STEVIN programma: Eindrapport. Report, Nederlandse Taalunie, The Hague (2011). http://taalunieversum.org/taal/technologie/stevin/documenten/stevin_eindevaluatierapport.pdf
14. TopTeam Creatieve Industrie: Creatieve industrie in topvorm: advies topteam creatieve industrie. Report, Ministry of Economic Affairs, Agriculture and Innovation, The Hague, the Netherlands (2011). http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2011/06/17/creatieve-industrie-in-topvorm.html
15. Uszkoreit, H., Prószéky, G., Moore, R., Calzolari, N., Heisterkamp, P., Adda, G., Abeillé, A., Piperidis, S.: STEVIN mid term review. Report, Nederlandse Taalunie, The Hague (2008). http://taalunieversum.org/taal/technologie/stevin/documenten/iap_tussentijdse_evaluatie.pdf
16. van den Bosch, A., Odijk, J., Cucchiarini, C., van den Bosch, L.: Taal- en spraaktechnologie voor het Nederlands: toptechnologie voor de topsectoren. http://www.clarin.nl/system/files/TST-Topsectoren.pdf (2011)