Chapter 1
Introduction

Peter Spyns

1.1 Context

The STEVIN ("STEVIN" is a Dutch acronym for "Essential Speech and Language Technology Resources for Dutch") programme aimed to contribute to the further progress of Human Language Technology for Dutch (HLTD) in the Low Countries (i.e., Flanders and the Netherlands) and to stimulate innovation in this sector. The major scientific goals were to set up an effective digital language infrastructure for Dutch, and to carry out strategic research in the field of language and speech technology for Dutch.[1] Consortia could submit project proposals in response to calls for proposals. Several calls were issued, and they included three open calls and two calls for tender as well. The thematic priorities for each call were determined in line with the overall STEVIN priorities and the state of their realisation before each call. The STEVIN thematic priorities, based on what is called the Basic Language Resource Kit (BLARK) for Dutch [20], are summarised in Tables 1.1 and 1.2. A BLARK is defined as the set of basic HLT resources that should be available for both academia and industry [13].

STEVIN advocated an integrated approach: develop text and speech resources and tools, stimulate innovative strategic and application-oriented research, promote embedding of HLT in existing applications and services, stimulate HLT demand via

[1] We refer the reader to Chap. 2 for more details.

P. Spyns (✉)
Nederlandse Taalunie, Lange Voorhout 19, 2514 EB Den Haag, Nederland
e-mail: pspyns@taalunie.org

Vlaamse overheid – Departement Economie, Wetenschap en Innovatie, Koning Albert II-laan 35, bus 10, B-1030 Brussel, België
e-mail: Peter.Spyns@ewi.vlaanderen.be

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_1, © The Author(s) 2013

Table 1.1 Summary of STEVIN scientific priorities – resources and research

Speech	Resources (I)	Strategic research (II)
	1. Speech and multimodal corpora for:	1. Robustness of speech recognition
	1. (a) Computer-Assisted Language Learning applications	
	1. (b) Applications in which names and addresses play an important role	
	1. (c) Call centres question and answer applications	
	1. (d) Educational applications	
	2. Multimodal corpora for applications of broadcast news transcription or person identification	2. Output treatment (inverse text normalisation)
	3. Text corpora for the development of stochastic language models	3. Confidence measures
	4. Tools and data for the development of:	4. Adaptation
	4. (a) Robust speech recognition	5. Lattices
	4. (b) Automatic annotation of corpora	
	4. (c) Speech synthesis	
Text	Resources (IV)	Strategic research (V)
	1. Richly annotated monolingual Dutch corpora	1. Semantic analysis, including semantic tagging and integrating morphological, syntactic and semantic modules
	2. Electronic lexicons	2. Text pre-processing
	3. Aligned parallel corpora	3. Morphological analysis (compounding and derivation)
		4. Syntactic analysis (robust parsing)

Table 1.2 Summary of STEVIN scientific priorities – application oriented

	Embedding HLTD
Speech (III)	1. Information extraction from audio transcripts created by speech recognisers
	2. Speaker accent and identity detection
	3. Dialogue systems and Q&A solutions, especially in multimodal domains
Text (VI)	1. Monolingual or multilingual information extraction
	2. Semantic web
	3. Automatic summarisation and text generation applications
	4. Machine translation
	5. Educational systems

demonstration projects and encourage cooperation and knowledge transfer between academia and industry. As all these aspects were targeted in one and the same programme, the structure and goals of STEVIN were based on the theoretical notion of a *stratified innovation system*. The main idea behind a stratified innovation

Table 1.3 Distribution of the STEVIN scientific projects (=the HLTD supply side) over the layers of a stratified innovation system – demonstrators, representing the application layer (=the HLTD demand side), are not included

	BLARK for Dutch	HLTD R&D	Embedding HLTD
Speech	Autonomata	Autonomata TOO	
	SPRAAK	MIDAS	DISCO
	STEVINcanPRAAT		
	JASMIN-CGN	NBest	
Text	D-Coi		
	LASSY		DuOMAn
	COREA	DAESO	Daisy
	Cornetto		
	DPC	PACO-MT	
	IRME		
	SoNaR		

system is that the strata or layers of an innovation system[2] do not exist in isolation, but build on one another [1, p. 63]. Consequently, each layer requires a proper set of government support measures that nevertheless have to be consistent and reinforce one another. For example, the STEVIN third open call, which focussed on application oriented projects, favoured proposals that used results (basic resources) of earlier STEVIN projects.

Modern theory on innovation systems states that no central entity can "steer" a sector or domain, but that knowledge (not limited to scientific expertise but also legal knowledge, business expertise etc.) of the sector or domain is distributed over the various actors in an innovation system. Hence, interactions and connections between these (different types of) actors had to be considered as well. Therefore, in addition to scientific projects, STEVIN also funded networking and dissemination activities. Depending on the focus of the projects, they are situated in a different layer of the innovation system. Table 1.3 shows all the STEVIN scientific projects (cf. Sect. 1.2.1) situated in the appropriate layer of a stratified innovation system. Four layers are distinguished:

- "BLARK": create basic HLT for Dutch (HLTD) resources – e.g., compile a large annotated corpus of written Dutch;
- "R&D": perform HLTD research[3] – e.g., investigate methods and build components that enhance the noise robustness of a speech recogniser;
- "embedding": enhance the functionality of applications thanks to HLTD – e.g., integrate a speech component in a computer-assisted language learning (CALL) system;
- "applications": create end-user HLTD applications – e.g., a speech therapy application for persons with a cochlear implant.

[2]Cf. the column labels of Table 1.3 and see [2] for the definition of an innovation system.

[3]In the case of STEVIN, it concerned strategic research, not fundamentel research.

In total, 14 demonstrators (cf. Sect. 1.2.2) were built mainly by small and medium sized enterprises (SMEs) (and hence represent the HLTD demand side), while 19 scientific projects (cf. Sect. 1.2.1) were carried out mainly by academic partners (the HLTD supply side).

1.2 STEVIN Projects

The most salient results of the various STEVIN projects are summarised below. Section 1.2.1 contains the main results of the scientific projects. In order to be complete, we enlist in Sect. 1.2.2 the other STEVIN projects as well. As their main goal was rather to create visibility for HLTD in the Low Countries than to achieve scientific progress, these projects are not further described in this volume.

1.2.1 STEVIN Scientific Projects

The numbers of the enumeration also refer to Fig. 1.1 of Sect. 1.3.2.

1. Autonomata (*Automata for deriving phoneme transcriptions of Dutch and Flemish names* – cf. Chap. 4) built two resources: (1) a grapheme-to-phoneme (g2p) conversion tool set for creating good phonetic transcriptions for text-to-speech (TTS) and automatic speech recognition (ASR) applications with a focus on phonetic transcriptions of names [27], and (2) a corpus of 72,000 spoken name utterances supplied with an orthographic and auditorily verified phonetic transcription [26]. These resources have been used in the Autonomata TOO project (cf. project 12).
2. SPRAAK (*Speech Processing, Recognition and Automatic Annotation Kit* – cf. Chap. 6) re-implemented and modernised a speech recognition tool kit and provided demo recognisers for Dutch. The SPRAAK tool kit combines many of the recent advances in automatic speech recognition with a very efficient decoder in a proven hidden Markov model (HMM) architecture (cf. project B in Fig. 1.1) [8]. SPRAAK is flexible modular tool kit meant for speech recognition research and a state of the art recogniser with an extensive programming interface.
3. STEVINcanPRAAT (cf. Chap. 5) extended the functionality of the widely used PRAAT open source package for doing phonetics by computer (cf. project A) [3]. In particular a Klatt synthesiser, a vowel editor and some under the hood improvements were added to the PRAAT system. The updated software is freely available via the regular PRAAT distribution channel (www.praat.org).
4. JASMIN-CGN *(Extension of the CGN with speech of children, non-natives, elderly and human-machine interaction)* – cf. Chap. 3 extended the Spoken Dutch Corpus (CGN – cf. project A in Fig. 1.1) with 115 h of read speech

and human-machine dialogues of children, non-natives and elderly people in the Netherlands (2/3) and Flanders (1/3). All recordings were delivered with a verbatim orthographic transcription, a transcription of the human-machine interaction (HMI) phenomena, part of speech (POS) tagging and an automatic phonetic transcription [6].

5. D-Coi (*Dutch Language Corpus Initiative* – cf. Chap. 13) was a preparatory project that created a blueprint for the construction of a 500-million-word corpus of contemporary written Dutch (SoNaR – cf. project 11) [16]. A set of annotation protocols and other reports useful for corpus building have been made available. A 54-million-word pilot corpus was compiled, parts of which were enriched with linguistic annotations. The corpus exploitation tool of the CGN (cf. project A) was adapted to cope with written text data.
6. LASSY (*LArge Scale SYntactic annotation of written Dutch* – cf. Chap. 9) created a large one-million-word corpus of written Dutch texts (LASSY small) that was syntactically annotated and manually corrected [23]. In addition, a 1.5-billion-word corpus (LASSY Large) was annotated automatically with part-of-speech and syntactic dependency information. Various browse and search tools for syntactically annotated corpora as well as the Alpino parser (cf. project D in Fig. 1.1) [24] were extended. These were used by DPC (cf. project 9) and SoNaR (cf. project 11).
7. COREA (*COreference Resolution for Extracting Answers* – cf. Chap. 7) implemented a robust tool to resolve coreferential relations in text and to support annotation activities by humans [11]. It is relevant for a range of applications, such as information extraction, question answering and summarisation. A corpus (in Dutch) was annotated with coreferential relations of over 200,000 words. In addition, general guidelines for co-reference annotation are available.
8. Cornetto (*Combinatorial and Relational Network as Tool Kit for Dutch Language Technology* – cf. Chap. 10) built a lexical semantic database for Dutch by combining and aligning the Dutch WordNet and the Reference File Dutch (Referentiebestand Nederlands). It includes the most generic and central part of the Dutch vocabulary and a specialised database for the legal and finance domains [31]. In total the Cornetto database contains more than 70,000 concepts, 92,000 words and 120,000 word meanings. Also a tool kit for the acquisition of new concepts and relations was implemented. This tool kit facilitates the tuning and extraction of domain specific sub-lexica from a compiled corpus. It was used in e.g., the FP7 Kyoto project [30].
9. DPC (*Dutch Parallel Corpus* – cf. Chap. 11) is a ten-million-word parallel corpus comprising texts in Dutch, English and French with Dutch as central language [17]. It consists of two sentence-aligned bilingual corpora (Dutch-English and Dutch-French) with a portion aligned at a sub-sentential level as well. The corpus has four translation directions (at least two million words per direction) and is a balanced corpus including five text types. A user friendly interface (parallel web concordancer) to query the parallel corpus is available on-line.

10. IRME (*Identification and Representation of Multi-word Expressions* – cf. Chap. 12) carried out research into sophisticated methods for automatically identifying MWEs in large text corpora and into a maximally theory-neutral lexical representation of MWEs. With an identification method derived from the research, a list of MWEs and their properties were automatically identified and formed the basis for the corpus-based DuELME Dutch lexical database of MWEs [10]. This DuELME database was later (not in the STEVIN context) adapted to be compliant with the Lexical Mark-up Framework (LMF).
11. SoNaR (*STEVIN reference corpus for Dutch* – cf. Chap. 13) constructed a 500-million-word reference corpus of contemporary written Dutch texts of various styles, genres and sources. The entire corpus was automatically tagged with parts of speech (POS) and lemmatised. In addition, for a one-million-word subset of the corpus different types of semantic annotation were provided, viz. named entity labels, co-reference relations, semantic roles and spatial and temporal relations. Tools and materials from other STEVIN projects (D-Coi, LASSY, COREA – cf. projects 5–7 respectively) were re-used. An important aspect of the project consisted of clearing the IPR for the various types of corpus material and documenting the acquisition process [19].
12. Autonomata TOO (*Autonomata Transfer of Output* – cf. Chap. 14) tackled the problem of spoken name recognition in the context of an automated Point of Interest (POI) providing business services [18]. New solutions were found by exploiting and extending the phoneme-to-phoneme (p2p) learning tools that were developed in the Autonomata project (cf. project 1). Autonomata Too delivered a demonstrator of a POI providing service and p2p converters for POI name transcription. Furthermore, it produced a corpus of read-aloud POI names from Belgium and the Netherlands. This corpus consists of 5,677 sound files and corresponding manually created phonetic transcriptions.
13. MIDAS (*MIssing DAta Solutions* – cf. Chap. 16) tackled the noise robustness problem in automatic speech recognition by missing data techniques, which enables masking out "unreliable" parts of the speech signal (due to noise etc.) during the recognition process [9]. The missing information is reconstructed by exploiting the redundancy in the speech signal. The algorithms were implemented and integrated in the SPRAAK speech recognition tool kit (cf. project 2).
14. NBest (*Dutch Benchmark Evaluation of Speech Recognition Technology* – cf. Chap. 15) developed an evaluation benchmark for large vocabulary continuous speech recognition in Dutch as spoken in Flanders and the Netherlands. It defined four primary tasks based on transcriptions of broadcast news and conversational telephony style speech in Northern and Southern Dutch [12]. The project defined evaluation protocols and training material, and collected evaluation data sets. Seven academic speech recognition systems – including SPRAAK (cf. project 2) – participated in the benchmark evaluation [28].
15. DAESO (*Detecting And Exploiting Semantic Overlap* – cf. Chap. 8) implemented tools for the automatic alignment and classification of semantic relations (between words, phrases and sentences) for Dutch, as well as for a Dutch

text-to-text generation application that fuses related sentences into a single grammatical sentence. The project also built a two-million-word monolingual parallel corpus [14]. In addition, three specific corpus exploitation tools were implemented as well as a multi-document summariser for Dutch.

16. PACO-MT (*Parse and Corpus based Machine Translation* – cf. Chap. 17) built a hybrid machine translation system for Dutch-English and Dutch-French (in both directions) integrating linguistic analysis and a transfer component based on syntactic structures into a data-driven approach [29]. Some specific components were implemented, such as a node aligner, a grammar rule inducer, a decoder and a target language generator. In addition, more than 48 resp. 45 million source words of parallel texts for Dutch-English resp. Dutch-French were collected.
17. DISCO (*Development and Integration of Speech technology into COurseware for language learning* – cf. Chap. 18) developed an ASR-based Computer-Assisted Language Learning (CALL) prototype for training oral proficiency for Dutch as a second language (DL2). The application optimises learning through interaction in realistic communication situations and provides intelligent feedback on various aspects of DL2 speaking, viz. pronunciation, morphology and syntax [21]. It uses the SPRAAK tool kit – cf. project 2.
18. DuOMAn (*Dutch Online Media Analysis* – cf. Chap. 20) developed a set of Dutch language resources (including sentiment-oriented lexica) and tools for identifying and aggregating sentiments in on-line data sources [22]. The tools support automated sentiment analysis, parsing, entity detection and co-reference resolution with an emphasis on robustness and adaptability. An on-line demonstrator is available.
19. Daisy (*Dutch lAnguage Investigation of Summarisation technologY* – cf. Chap. 19) developed and evaluated technology for automatic summarisation of Dutch informative texts. Innovative algorithms for topic salience detection, topic discrimination, rhetorical classification of content, sentence compression and text generation were implemented [7]. A demonstrator was built and the Alpino parser (cf. project D in Fig. 1.1) was extended with a text generation and fluency restoring component. In addition, a tool that segments and classifies the content of Web pages according to their rhetorical role was implemented.

1.2.2 Other STEVIN Projects

The "other" projects mainly include demonstration projects. They served to convincingly illustrate the feasibility of applying HLT in end-user applications and services in Dutch. The purpose was to stimulate the uptake of HLTD by industry. Therefore, the main applicant had to be a private company. Two "educational projects" had to increase the interest of students for HLT and HLT related studies. Two master classes targeted decision makers in industry and government to increase their awareness of the potentialities of adopting HLT in their organisation. Four

of these "other" projects (labelled i–iv) are included in Fig. 1.1 as they build on resources of earlier STEVIN scientific projects. We refer the interested reader to the "STEVIN programme: project results" booklet[4] for more detailed descriptions of these projects.

1. The "*licence plate line*" ("Kentekenlijn") allows Dutch police officers on the road (in a car, on a bicycle, on foot) to check registration plates of vehicles in a hands-free (and eyes-free) manner using the NATO alphabet. Various types of information on the car are read aloud using speech synthesis. As a result, fewer officers are needed in the police control room to manage this type of calls and requests. Hence, they can spend more time on more urgent and higher priority calls. And, more requests for licence plate numbers can be processed.
2. The Dutch information portal for legal professionals ("*Rechtsorde*") provides a more user-friendly access to information about laws and local government regulations as available in official legal publications. The system corrects spelling errors and suggests synonyms and closely related terms based on compound decomposition and inflectional analysis.
3. *"CommuneConnect!"* ("GemeenteConnect") is a phone dialogue system that allows for free speech input that provides the caller with information on legislation and procedures that apply in a commune (question-answering). It uses a combination of state-of-the-art speech recognition, classification and computational linguistics based dialogue management.
4. A *spell check chatbot* ("Spelspiek") provides the correct spelling of pseudo-phonetically spelled words that present spelling difficulties. If several spelling alternatives exist, extra explanation is added. It consists of an automatic conversational agent that behaves as an expert in Dutch spelling. The core of the system consists of a one-million-word vocabulary, a spelling error database and smart spelling error recognition algorithms. The chatbot also knows how to respond to "unexpected" input by exhibiting some sense of humour. Currently, the service is available through a webpage and Twitter.
5. "*SonaLing*" ("Klinkende Taal") is a dynamic jargon detection system to avoid administrative and complicated language in documents and written customer communication by local governments. It focusses on readability and revision advice. The project resulted in several commercial product offers, including a freely accessible web tool, and a large user base.
6. *WebAssess* allows for the automatic pre-selection of call centre agent candidates during the recruitment process. The total set-up includes an e-learning module via the internet and a speech interactive voice response system that mimics a customer calling a contact centre. The system checks the replies of the candidate on the presence of need-to-have-answers and nice-to-have-answers, and adapts the dialogue flow accordingly.

[4] http://www.stevin-tst.org/english/

7. *Primus* adapted Microsoft's Dutch spelling and grammar checking tools for use by dyslectic users. Adapted grammar rules provide more useful correction suggestions and a text-to-speech module pronounces the suggestions.
8. A Flemish editor offers a daily audio edition "*Audio Newspaper*" ("Audiokrant") for visually disabled persons of two popular newspapers. A daily production process using speech synthesis generates CDs that contain a complete spoken newspaper. The CDs are compliant with the international DAISY (digital accessible information system) standard that allows for navigation over the newspaper content.
9. The "*NeOn*" project (labelled "iii" in Fig. 1.1) combines speech segmentation, speech recognition, text alignment and sentence condensation techniques to implement a less labour intensive semi-automatic tool to produce Dutch subtitles for certain television shows (for the Dutch NPO and Flemish VRT broadcasting organisations). This resulted in a reduction of up to 40 % in processing time compared to the method traditionally used. A follow-up project has been initiated by the VRT.
10. The "*Justice recognition*" ("Rechtspraakherkenning") application produces transcriptions of recordings made in the courtrooms in the Netherlands. The recordings are made searchable to enable retrieval of relevant items from the fully recorded lawsuits. In addition, a spoken summary of the trial can be generated. Even if the transcriptions are not completely accurate, the application significantly reduces the human effort in producing full transcriptions. Also the search and retrieval function is well appreciated. Several governmental organisations in the Netherlands have shown interest in this application.
11. *Woody* is a self-correcting talking word prediction system built for dyslectic users. Word lists and word prediction algorithms form the core of the tool. The project was the basis for a subsequent commercially available product called Wody.
12. The *"literacy plan foreign speakers* ("Alfabetisering Anderstaligen Plan" or AAP – labelled "ii" in Fig. 1.1) demonstrator can be used to train knowledge of the Dutch language, pronunciation of Dutch, and Dutch literacy. It uses speech recognition software and can be integrated in an existing language learning application for second language learners with a very limited level of literacy and limited knowledge of Dutch to produce speech feedback.
13. The *Hatci* project (labelled "iv" in Fig. 1.1) resulted in an automatic speech assessment system that can support a speech therapist in helping a patient with a cochlear implant to learn to speak. The tool plays an audio file (and/or a video file to allow for lip reading) to prompt a patient. A speech recogniser analyses the accuracy of the reproduction by the patient and hence assesses his/her hearing and speech reproduction abilities.
14. The "*YourNews*" news brokerage service uses language technology to collect, summarise and classify more than 1,000 newspaper articles per minute in accordance with the International Press and Telecom Council (ITPC) classification standard.

15. Two master classes were prepared and organised: one on ICT and dyslexia, and a second one on a general introduction on HLT for decision makers of mainly public organisations.
16. *Kennislink* is a website popular in the Low Countries mainly used by students and teachers to find information about recent scientific developments. Ninety-three articles on STEVIN projects and HLT in general were added to the Kennislink website. In addition, two perception studies were organised amongst students: one to rate the Kennislink HLT articles and one about their familiarity with and interest in HLT.
17. The *DiaDemo* "educational" application (labelled "i" in Fig. 1.1) can detect on the spot to which main dialect group a Flemish person belongs on basis of a few utterances.

1.3 Mission Accomplished

1.3.1 Addressing the STEVIN Priorities

To know the extent to which STEVIN has achieved its targets as defined at the start of the programme, the STEVIN priorities are compared to the topics and output of the various projects. Table 1.4 shows the distribution of the 19 scientific projects (cf. Sect. 1.2.1 for their short descriptions) over the STEVIN priorities as detailed in Tables 1.1 and 1.2 (cf. Sect. 1.1). The subsequent chapters of this volume provide ample details of each STEVIN scientific project.

The SPRAAK project, in combination with the MIDAS project, covered the development of a robust speech recogniser with additional features for noise robustness (priorities II.1–5). SPRAAK is re-used by the DISCO project that itself is a computer-assisted language learning application (priority VI.5). Autonomata and Autonomata TOO address the issues regarding the correct synthesis (priority I.4.c) and robuster recognition (priority II.1) of proper nouns, street names and names of points of interest (priority I.1.b), which is highly relevant for (car) navigation systems and call centre applications (priorities I.1.c). STEVINcanPRAAT basically is an improvement of the PRAAT (phonetic) signal processing tool (priority I.4.c). The JASMIN-CGN project extended the already available Spoken Dutch Corpus, in a manner useful for CALL applications (priorities I.1.a and I.1.d), and built automatic speech corpus annotation tools (priorities I.4.a–b).

Many STEVIN scientific projects obviously dealt with the creation and annotation of a corpus for written Dutch: D-Coi, LASSY, IRME and, of course, SoNaR that built a reference corpus of written Dutch of 500 million words (priorities IV.1 and I.3). The SoNaR corpus was annotated automatically using pre-processing tools and syntactico-semantic annotation tools and tagging schemas resulting from the D-Coi corpus pilot project (priorities IV.1). Also the COREA co-reference tools were used to annotate the SoNaR corpus. Lexica (priorities IV.2) were built by the IRME, Cornetto and DuOMAn projects. The DAESO tools focused on alignment

Table 1.4 STEVIN scientific projects mapped on the STEVIN priorities (cf. Tables 1.1 and 1.2) they mainly address – empty cells represent priorities not covered

Speech	Resources (I)	Strategic research (II)	Applications (III)
	1.(a) JASMIN-CGN	1. Autonomata TOO, SPRAAK, MIDAS	1.
	1.(b) Autonomata, Autonomata TOO		
	1.(c) Autonomata, Autonomata TOO		
	1.(d) JASMIN-CGN		
	2.	2. SPRAAK	2.
	3. JASMIN-CGN, Autonomata	3. SPRAAK, MIDAS	3.
	4.(a) JASMIN-CGN	4. SPRAAK	
	4.(b) JASMIN-CGN	5. SPRAAK	
	4.(c) Autonomata, Autonomata TOO, (STEVINcanPRAAT)		
Text	Resources (IV)	Strategic research (V)	Applications (VI)
	1. COREA, LASSY, SoNaR	1. DAESO, Daisy, DuOMAn	1. Daisy, DuOMAn
	2. IRME, Cornetto, DuOMAn	2. Daisy, DuOMAn, PACO-MT,	2.
	3. DPC	3.	3. DAESO, Daisy
		4. PACO-MT	4. PACO-MT
			5.

of semantic relationships (at the sentence level) and sentence fusion (priority V.1). These are useful for QA applications, information extraction and summarisation (priorities VI.1 and VI.3). These latter two topics, albeit on the discourse level, were also addressed by the Daisy project. DuOMAn produced (web) opinion mining tools (priority VI.1). DPC built a trilingual parallel corpus (priority IV.3) that can be useful for machine translation systems, such as Paco-MT (priority VI.4). Many corpus projects used the Alpino parser to produce syntactic annotations. Hence, even if no single project aimed at re-implementing a robust parser, as the SPRAAK project did for a robust speech recogniser, the Alpino robust syntactic parser has been improved and extended in several ways by various STEVIN projects (priority V.4).

Still, not all the priorities could be addressed: e.g., the lack of a tool for morphological analysis for derivation and compounding (priority V.3) and the absence of a text-based educational system (priority VI.5) are considered as lacunas. Also, more projects related to the semantic web (priority VI.2) would have been welcome, even if Cornetto, which created a lexical semantic database, is surely of high relevance for semantic web applications in Dutch. The BLARK for Dutch report [20] also listed the creation of benchmarks as an important action line (cf. Chap. 2, Sect. 2.2.3, p. 24). The STEVIN work programme did not retain this

topic as a priority item. However, the work programme text did state that projects could receive funding for the creation of test suites, test data and benchmarks if used for project internal systematic evaluation. Some of these data sets and test suites are available and can serve as reference data for other researchers. Only one specific project dedicated to creating a benchmark, c.q. for speech recognisers, was proposed (and awarded): NBest – cf. Chap.15, p. 271.

Some of the STEVIN priorities have been achieved by other research programmes, amongst others, the IMIX programme (Interactive Multimodal Information eXtraction.[5]) The IMIX programme, solely financed by the Netherlands (NWO), focussed on multimodal dialogue management for a medical QA system [25] and a non domain specific QA system called Joost [4] (priority III.3). IMIX started in 2002 while STEVIN was still under preparation. Therefore, it was later on agreed amongst the funding organisations that STEVIN projects could benefit from IMIX results, and that STEVIN would not explicitly target the IMIX topics. Funding agencies continued to award national projects that dealt with monolingual multimodal information extraction: AMASS++ (IWT-SBO)[6] (priorities V.1, VI.1, VI.1.3 and III.1), searching in radio and television multimedia archives: BATS (IBBT & ICT-Regie Im-Pact)[7] (priorities I.2 and III.1–2), compiling the CHOREC corpus of speech of children (IWT-SBO)[8] (priorities I.4.1, I.4.3, II.1, II.4 and VI.5), semantic web applications such as Kyoto[9] (EC-FP7) (priority VI.2) etc.

Thus, all in all, the STEVIN priorities are to a very large extent achieved. Concerning the creation of a digital language infrastructure, STEVIN is even cited as "probably the best example of a BLARK initiative for a tier 2 languague" [5, p. 1805]. Nevertheless, the topics not addressed during STEVIN have to be retained as themes for subsequent R&D funding initiatives, or at least their priority status is to be reconfirmed.

1.3.2 Improving the Scientific Capacity

Not only the coverage of each scientific priority in isolation constitutes a success indicator for STEVIN, but also the degree of "convergence" between the project results highly matters. For example, it would not be sensible to improve a syntactic parser for annotation purposes if the annotation schema used (strongly) differs from annotation schemas used by corpus projects. Also, technological components have to be (backwards) compatible and easily integratable with other lingware modules or larger frameworks. It is quite irrelevant and useless to implement a proper

[5] http://www.nwo.nl/nwohome.nsf/pages/NWOP_5ZLCE8_Eng

[6] http://www.cs.kuleuven.be/groups/liir/projects/amass/

[7] http://www.ibbt.be/en/projects/overview-projects/p/detail/bats

[8] http://www.esat.kuleuven.be/psi/spraak/projects/SPACE/

[9] http://cordis.europa.eu/fp7/ict/content-knowledge/projects-kyoto_en.html

noun speech synthesis module that can only function in a standalone way. In such a case, that particular scientific STEVIN priority could have been well covered, but the value for the overall HLTD community (academia and industry) might be fairly limited. The build-up of scientific capacity, including a digital language infrastructure, is not about re-inventing the wheel but rather about "standing on the shoulders of giants".

Figure 1.1 shows all the STEVIN scientific projects (cf. Sect. 1.2.1), four earlier (important) HLTD resources (projects A–D),[10] and four of the demonstrators (cf. Sect. 1.2.2 – DiaDemo (i), AAP (ii), NeOn (iii) and Hatci (iv)) that integrate STEVIN scientific results. The figure shows that STEVIN scientific projects do not constitute "islands", but that the results are shared, re-used and improved by other scientific projects and even – if the time lines permitted – integrated into end-user applications.

Figure 1.1 shows that more *resources* (the BLARK for Dutch layer) for speech have been created prior to STEVIN. The CGN, the Spoken Dutch Corpus developed earlier (project A [15]),[11] is the current reference corpus for spoken Dutch. Therefore, efforts on speech resources could be limited to extending the CGN corpus for specific target groups (JASMIN-CGN – project 4). The HMM speech recogniser (project B by the KU Leuven) has been upgraded into the SPRAAK package (project 2). The open source software of Praat (project C by the University of Amsterdam) has been extended in the STEVINcanPRAAT project (project 3).

Regarding textual resources, some catching-up had to be done. Hence, quite some STEVIN projects have created annotated textual corpora and lexica. The many connections between all the STEVIN corpus projects (cf. Fig. 1.1) show a high degree of interrelatedness. In particular, SoNaR (project 11) with its pilot project (project 5), is meant to become *the* reference written language corpus for Dutch. All these corpus efforts additionally resulted in extensive expertise in what is usually considered to be "trivial" issues such as data acquisition, IPR clearing and licence handling. These issues are in fact far from trivial (cf. [19]). On the contrary, the subsequent exploitation and dissemination of a corpus crucially depend on it. This kind of knowledge surely can be considered as a valuable resource for a digital language infrastructure – albeit of a different nature. The Alpino syntactic parser (project D by the University of Groningen) open source package has been used, adapted and extended by several STEVIN projects, mainly LASSY (project 6) and PACO-MT (project 16).

The pre-STEVIN materials already established themselves as the reference resource or tool (of their kind) in the Low Countries. Also their extensions (JASMIN-CGN, STEVINcanPRAAT and the various Alpino adaptations) will most probably "inherit" the success of the ancestor. And Fig. 1.1 clearly illustrates the importance of the SPRAAK tool kit for the field.

[10] Pre-STEVIN projects are shown in grey.

[11] The CGN is owned by the Dutch Language Union and maintained and made available by the HLT Agency – cf. Chap. 21.

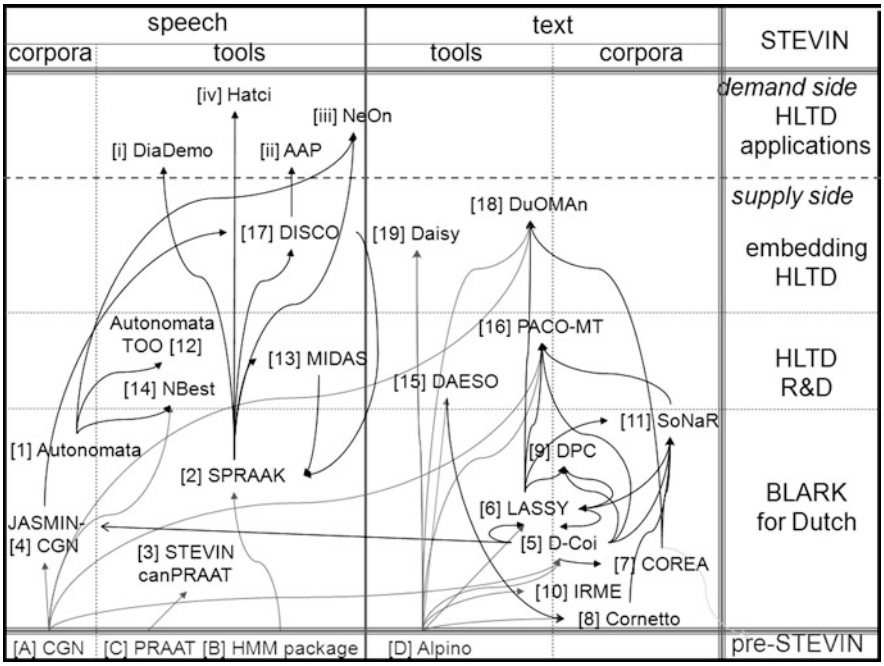

Fig. 1.1 A dependency graph showing how (pre-)STEVIN (scientific) projects are interrelated – projects are classified according to their most important results

The *HLTD R&D* layer presents a different situation for speech vs. text. In the speech processing area, several commercial TTS-engines (offering Dutch) exist (albeit as proprietary systems and "closed" source). The focus was put on improving the robustness of a speech recogniser and the treatment of proper nouns. The additional modules for proper noun pronunciation implemented by Autonomata (project 1) and Autonomata Too (project 12) can be used on top of a major standard commercial TTS package. Components of MIDAS have been integrated into SPRAAK to enhance the robustness to noise of the speech recognition tool kit. In the text domain, parsers and tagger/lemmatisers already exist in greater number. The research focus for STEVIN was thus placed on areas such as hybrid machine translation (PACO-MT – project 16), sentence fusion and detection of semantic overlap (DAESO – project 15).

STEVIN's *HLT embedded* text projects and applications (DuOMAn – project 18 and Daisy – project 19) were building to a lesser extent on previously developed STEVIN basic resources than is the case for the speech domain (DISCO – project 17) due to timing conflicts, even if some re-usage of materials did occur. Also, both in Flanders and the Netherlands, more research groups are working on text technology, all having their own tools based on different methods and principles (e.g., hand crafted rules vs. rules generated by machine learning techniques). In many cases, these have been adapted to Dutch so that the variety of tools available

is higher. Less de facto standard software packages exist in this domain – the Alpino parser being a notable exception.

However, it is to expected that in the future more tools and standards will establish themselves as de facto reference material. By the intermediary of CLARIN-NL,[12] standard file formats will most probably become widely used by the HLTD community, which will enhance the exchangeability of data between the various tools. Actually, the CLARIN-VL-NL[13] project with the name TTNWW, jointly funded by Flanders and the Netherlands, precisely aims at establishing standard formats to ensure the interoperability of tools during the execution of HLTD work flow processes. Many STEVIN materials are re-used in various CLARIN-NL projects. Hence, it is valid to state that STEVIN materials have effectively and substantially contributed to the build-up of HLTD capacity in the Low Countries. And it is equally safe to expect that STEVIN materials will remain important for the field in the near future. We refer the reader to the overall concluding chapter (Chap. 22, p. 395) for more reflections on the international context of STEVIN and for an outlook for future HLTD activities.

1.4 Organisation of This Volume

The remainder of this volume is organised as follows. A more detailed account from a policy point of view on the STEVIN programme is offered in the second chapter of this volume (Part I). The chapters on the STEVIN scientific projects are grouped into three parts in line with Table 1.3: resource related (II), technology or research related (III) and application related (IV). In a separate chapter, the HLT Agency, which is responsible for the IPR management, the maintenance and distribution of the STEVIN results, presents itself. Together with a concluding and forward looking chapter, it constitutes Part V.

References

1. Akkermans, J., van Berkel, B., Frowein, C., van Groos, L., Van Compernolle, D.: Technologieverkenning nederlandse taal- en spraaktechnologie. Techincal report., M&I/Partners & Montemore, Amersfoort & Leuven (2004) (in Dutch)
2. Arnold, E., Kuhlman, S.: Rcn in the norwegian research and innovation system. In: Background report no. 12 in the Evaluation of the Research Council of Norway, Fraunhofer ISI, Karlsruhe (2001)

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

[13] http://www.ccl.kuleuven.be/CLARIN/pilot.htm – in Dutch

3. Boersma, P.: PRAAT, a system for doing phonetics by computer. Glot Int. **5:9/10**, 341–345 (2001)
4. Bouma, G., Muri, J., van Noord, G., van der Plas, L.: Question-answering for dutch using dependency relations. In: Proceedings of the CLEF 2005 Workshop, Vienna (2005)
5. Boves, L., Carlson, R., Hinrichs, E., House, D., Krauwer, S., Lemnitzer, L., Vainio, M., Wittenburg, P.: Resources for speech research: present and future infrastructure needs. In: Proceedings of Interspeech 2009, Brighton (2009)
6. Cucchiarini, C., Driesen, J., Van hamme, H., Sanders, E.: Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech (2008)
7. De Belder, J., Moens, M.F.: Integer linear programming for Dutch sentence compression. In: Gelbukh A. (ed.) Proceedings of CICLing 2010, Lasi. Lecture Notes in Computer Science, pp. 711–723. Springer, Berlin, Heidelberg (2010)
8. Demuynck, K., Roelens, J., Van Compernolle, D., Wambacq, P.: SPRAAK: an open source speech recognition and automatic annotation kit. In: Proceedings of the International Conference on Spoken Language Processing, Jeju, Korea (2008)
9. Gemmeke, J., Van hamme, H., Cranen, B., Boves, L.: Compressive sensing for missing data imputation in noise robust speech recognition. IEEE J. Sel. Top. Signal Process. **4**(2), 272–287 (2010)
10. Grégoire, N.: Duelme: a Dutch electronic lexicon of multiword expressions. J. Lang. Resour. Eval. Special issue on Multiword Expressions. **44**(1-2), 23–40. Springer (2010)
11. Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.M., Van Der Vloet, J., Verschelde, J.L.: A coreference corpus and resolution system for Dutch. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech (2008)
12. Kessens, J., van Leeuwen, D.: N-Best: the northern and southern Dutch evaluation of speech recognition technology. In: Proceedings of Interspeech 2007, Antwerp, pp. 1354–1357 (2007)
13. Krauwer, S.: The Basic Language Resource Kit (BLARK) as the first milestone for the Language Resources Roadmap. In: Proceedings of the International Workshop Speech and Computer 2003, Moscow (2003)
14. Marsi, E., Krahmer, E.: Detecting semantic overlap: a parallel monolingual treebank for Dutch. In: Proceedings of Computational Linguistics in the Netherlands (CLIN 2007), Nijmegen (2007)
15. Oostdijk, N.: The design of the Spoken Dutch Corpus. In: Peters P., Collins, P., Smith, A. (eds.) New Frontiers of Corpus Research. Rodopi, Amsterdam/New York (2002), pp. 105–112
16. Oostdijk, N., Reynaert, M., Monachesi, P., van Noord, G., Ordelman, R., Schuurman, I., Vandeghinste, V.: From D-Coi to SoNaR: a reference corpus for Dutch. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech (2008)
17. Paulussen, H., Macken, L., Truskina, J., Desmet, P., Vandeweghe, W.: Dutch Parallel Corpus a multifunctional and multilingual corpus. Cahiers de l'Institut de Linguistique de Louvain **32.1-4**, 269–285 (2006)
18. Réveil, B., Martens, J.P., van den Heuvel, H.: Improving proper name recognition by adding automatically learned pronunciation variants to the lexicon. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta (2010)
19. Reynaert, M., Oostdijk, N., De Clercq, O., van den Heuvel, H., de Jong, F.: Balancing SoNaR: IPR versus processing issues in a 500-million-word written Dutch Reference Corpus. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta (2010)
20. Strik, H., Daelemans, W., Binnenpoorte, D., Sturm, J., de Vriend, F., Cucchiarini, C.: Dutch HLT resources: from BLARK to priority lists. In: Proceedings of ICSLP, Denver (2002)

21. Strik, H., Cornillie, F., Colpaert, J., van Doremalen, J., Cucchiarini, C.: Developing a CALL system for practicing oral proficiency: how to design for speech technology, pedagogy and learners. In: Proceedings of the SLaTE-2009 Workshop, Warwickshire (2009)
22. Tsagkias, E., Weerkamp, W., de Rijke, M.: News comments: exploring, modeling, and online predicting. In: Proceedings of the 2nd European Conference on Information Retrieval (ECIR 2010), pp. 109–203. Springer, Milton Keynes, UK (2010)
23. van Noord, G.: Huge parsed corpora in LASSY. In: Van Eynde, F., Frank, A., De Smedt, K., van Noord G. (eds.) Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories (TLT 7), Groningen (2009). LOT Occasional Series
24. van Noord, G.: Learning efficient parsing. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009), Athens (2009). Association for Computational Linguistics
25. van den Bosch, A., Bouma, G. (eds.): Interactive Mulit-modal Question-Answering. Theory and Applications of Natural Language Processing. Springer, Heidelberg/New York (2011)
26. van den Heuvel, H., Martens, J.P., D'Hoore, B., D'Hanens, K., Konings, N.: The Autonomata spoken name corpus. design, recording, transcription and distribution of the corpus. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech (2008)
27. van den Heuvel, H., Martens, J.P., Konings, N.: Fast and easy development of pronunciation lexicons for names. In: Proceedings of LangTech, Rome (2008)
28. van Leeuwen, D., Kessens, J., Sanders, E., van den Heuvel, H.: Results of the N-Best 2008 Dutch speech recognition evaluation. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009), Brighton. International Speech Communication Association, pp. 2571–2574 (2009)
29. Vandeghinste, V.: Scaling up a hybrid mt system: from low to full resources. Linguist. Antverp. **7**, 65–80 (2008)
30. Vossen, P., Agirre, E., Calzolari, N., Fellbaum, C., Hsieh, S., Huang, C., Isahara, H., Kanzak, K., Marchetti, A., Monachini, M., Neri, F., Raffaelli, R., Rigau, G., Tescon, M.: Kyoto: a system for mining, structuring and distributing knowledge across languages and cultures. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech (2008)
31. Vossen, P., Maks, I., Segers, R., van der Vliet, H., van Zutphen, H.: The Cornetto Database: the architecture and alignment issues of combining lexical units, synsets and an ontology. In: Proceedings of the Fourth International GlobalWordNet Conference, Szeged (2008)