

# Results of the STEVIN programme

STEVIN Final Event, Rotterdam, Nov 28 2011



Jan Odijk

 | STEVIN

# Overview

- **STEVIN Objectives**
- Digital Language Infrastructure
  - Creation
  - Resource Management
  - IPR
- Strategic Research
- LST Community Consolidation
- Various Statistics

# STEVIN Objectives

- Digital Language Infrastructure (DLI)
- Strategic Research (SR)
- LST community consolidation (CC)

# Overview

- STEVIN Objectives
- **Digital Language Infrastructure**
  - Creation
  - Resource Management
  - IPR
- Strategic Research
- LST Community Consolidation
- Various Statistics

# Digital Language Infrastructure

- Creation
- Resource Management
- IPR

# Overview

- STEVIN Objectives
- Digital Language Infrastructure
  - Creation
  - Resource Management
  - IPR
- Strategic Research
- LST Community Consolidation
- Various Statistics

# DLI: Creation

- Priorities for written language:
  - a. A large corpus of written Dutch
  - b. An electronic lexicon
  - c. Parallel corpora

# Realisation: Written (1)

- D-COI + SONAR: 500M word corpus (a)
- LASSY: 1M word Treebank (a)
- CORNETTO: 40k entry lexical semantic database (b)
- DPC: 10M word parallel corpus D-E / D-F (c )

## Realisation: Written (2)

- COREA: co-reference corpus (a)
- IRME: 5k MWE lexical database (b)
- DAESO: 1M word monolingual parallel corpus (c)
- DAISY (a)
- DUOMAN (a)
- PACO-MT (a,c)

# Creation: Priorities Speech (1)

- a. speech and multimodal corpora for CALL, NAW, CCQA applications
- b. multimodal corpora for
  - broadcast news transcription or
  - person identification;
- c. text corpora for stochastic language models;

## Creation: Priorities Speech (2)

- d. tools and data for the development of
  - robust speech recognition;
  - automatic annotation of corpora;
- e. speech synthesis;

# Realisation: Speech (1)

- Automata (a, NAW; e)
- JASMIN-CGN (a, CALL)
- D-COI + SONAR (c )
- SPRAAK (d)
- STEVINcanPRAAT (d)

## Realisation: Speech (2)

- Missing
  - (b) Multimodal corpora
- But partially covered by other projects
  - EU: AMI, AMIDA (U Twente)
  - NL: IMIX

# Overview

- STEVIN Objectives
- Digital Language Infrastructure
  - Creation
  - **Resource Management**
  - IPR
- Strategic Research
- LST Community Consolidation
- Various Statistics

# DLI: Resource Management

- HLT Agency set up
- See presentation by Remco van Veenendaal

# Overview

- STEVIN Objectives
- Digital Language Infrastructure
  - Creation
  - Resource Management
  - **IPR**
- Strategic Research
- LST Community Consolidation
- Various Statistics

# DLI: IPR

- Systematic attention for IPR & Ethical Issues from the start
  - Not easy but
  - The only way to ensure usage of LRs by the R&D community in a legal manner
- Specific regulation on how to deal with IPR in the STEVIN programme and projects

# Overview

- STEVIN Objectives
- Digital Language Infrastructure
  - Creation
  - Resource Management
  - IPR
- **Strategic Research**
- LST Community Consolidation
- Various Statistics

# Strategic Research

- Will be dealt with by Walter in his presentation
- Work programme lists examples of applications
  - how do STEVIN projects contribute to such applications (directly or indirectly)

# SR: Applications (1)

- **Information extraction from Speech:**
  - Rechtspraakherkenning, NEON, and SNRT
  - AUTONOMATA, JASMIN-CGN, SPRAAK, STEVINcanPRAAT, N-BEST, AUTONOMATA TOO and MIDAS.
- **Detection of accent and identity of speakers.**
  - JASMIN-CGN, SPRAAK, DISCO, Diademo, Rechtspraakherkenning

## SR: Applications (2)

- **Extraction of information from (monolingual or multilingual) text.**
  - DAESO, DUOMAN, Gemeenteconnect and YourNews.
  - COREA, IRME, D-COI, SONAR, DPC, LASSY, CORNETTO, and PACO-MT
- **Semantic web:**
  - CORNETTO, D-COI and SONAR

# SR: Applications (3)

- **Dialogue systems and Q&A solutions**
  - DAISY, DUOMAN, Gemeenteconnect, Web Assess.
- **Automatic summarization and text generation**
  - DAESO, Web Assess
  - D-COI and SONAR,

# SR: Applications (4)

- **Automatic Translation**
  - DPC, PACO-MT
  - D-COI, SONAR, LASSY, IRME, COREA, CORNETTO
- **Educational systems**
  - DISCO, SpelSpiek, Primus, HATCI, WooDy, AAP
  - All resource creation projects

# Overview

- STEVIN Objectives
- Digital Language Infrastructure
  - Creation
  - Resource Management
  - IPR
- Strategic Research
- **LST Community Consolidation**
- Various Statistics

# LST Community Consolidation

- Create networks
- consolidate LST activities
- educate new experts
- promote discussion
- promote transfer of knowledge

# LST Community Consolidation

- Set aside a specific budget and a dedicated WG
- joint KI/SME and NL/FL projects preferred
  - 330 binary cooperation link occurrences
- demonstration projects stimulated companies to participate

# LST Community Consolidation

- Educational projects (3)
- Master classes (2)
- Networking events organized
  - brokerage events, “Taal in Bedrijf” (‘language@work’), STEVIN programme meetings, etc..
- Networking events supported
  - e.g. CLIN, InterSpeech2007, ICT-Delta

# Overview

- STEVIN Objectives
- Digital Language Infrastructure
  - Creation
  - Resource Management
  - IPR
- Strategic Research
- LST Community Consolidation
- **Various Statistics**

# Money Distribution

- R&D (76.0%)
- Demonstration ( 8.5%)
- Supporting Activities ( 6.0%)
- HLT Agency ( 2.5%)
- STEVIN Management ( 6.5%)

# Strata Coverage

- Basic resources for LST (51.1%)
- Basic Research (23.3%)
- Application-oriented Res. (15.4%)
- Demonstration projects (10.2%)

## NL / FL Proportion

- R&D Projects 63%:37%
- Demonstrator projects 66%:34%
- Overall 64%-36%
  
- Educational projects (3) 68%:32%
- Master classes (2) 100%:0%

## KI / SME Proportion

- Money 83%: 17%
- R&D projects by project 19 : 13
- R&D projects by #participations 80%: 20%
- Demonstration projects 15%: 85%
- Master classes 0%:100%
- Education activities 83%: 17%

# Language / Speech

- Money: 53.1%:46.9%

## Funded v. Submitted

- R&D count 1 19/52 (36.5%)
- R&D count 2 19/68 (27.9%)
- Demonstration 14/41 (30.0%)
- Educational 3/ 5 (60%)
- Master Classes 2/ 3 (66.6%)
- Most proposals were very good
  - So many more could and should be done



Thanks for your Attention!