



The
Dutch-Flemish
HLT Programme STEVIN:
Essential Speech and Language Technology Resources

Elisabeth D'Halleweyn - Jan Odijk - Lisanne Teunissen - Catia Cucchiarini

Why?

Who?

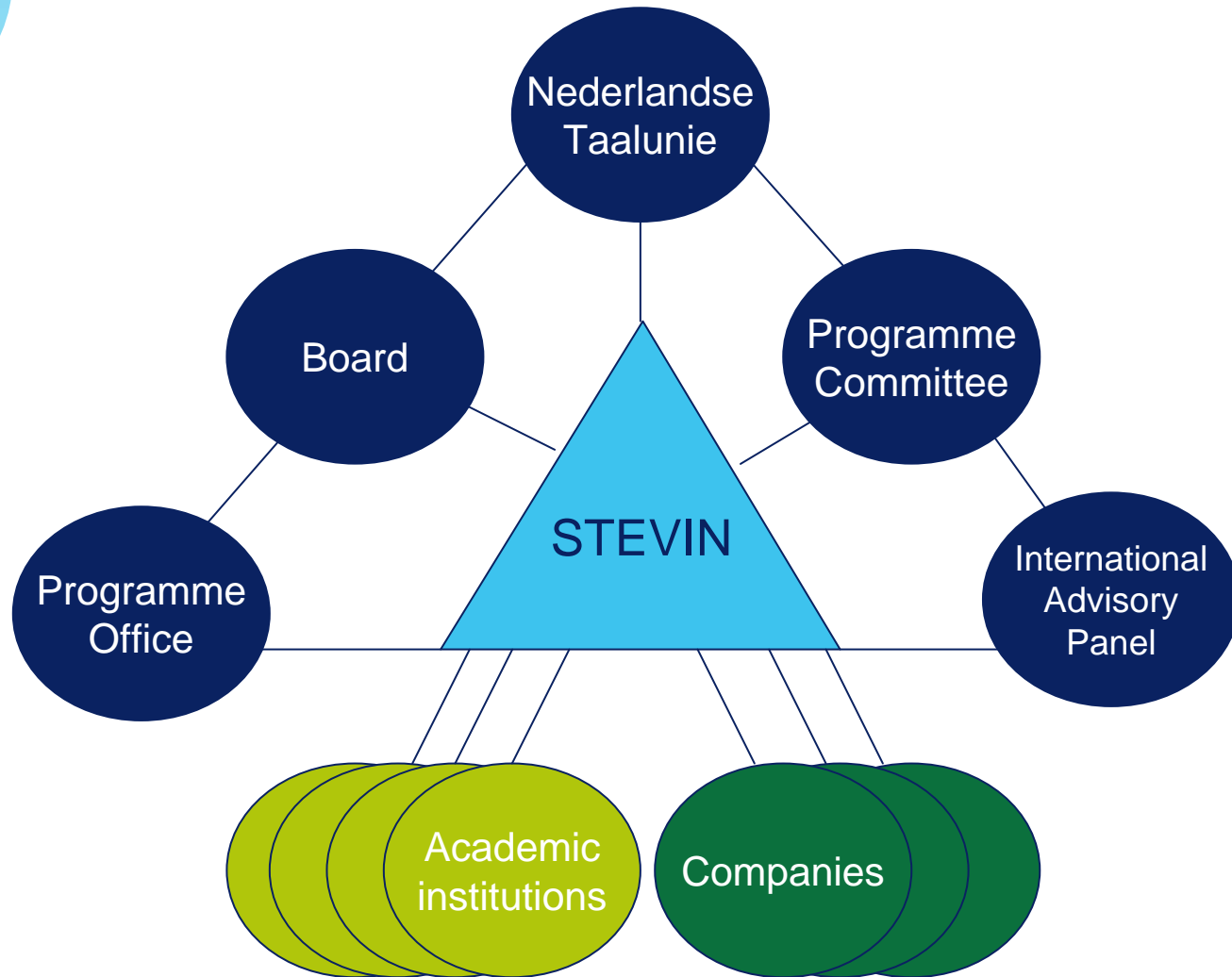
When?

What?

Why?

- HLT enables communication in the information society
- improve / secure position of Dutch: language of 21 million people
- stimulate technology sector

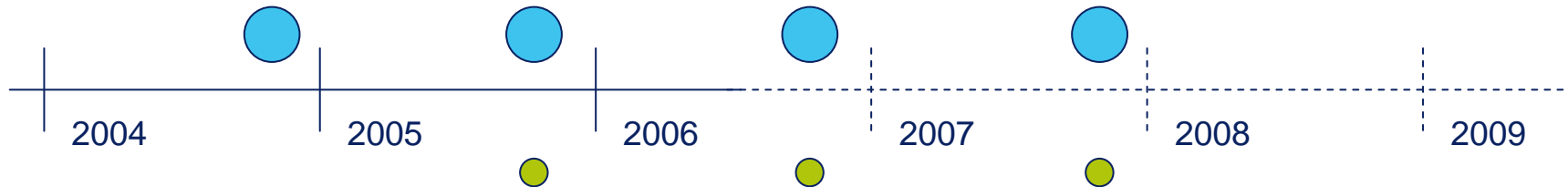
Who?



When?

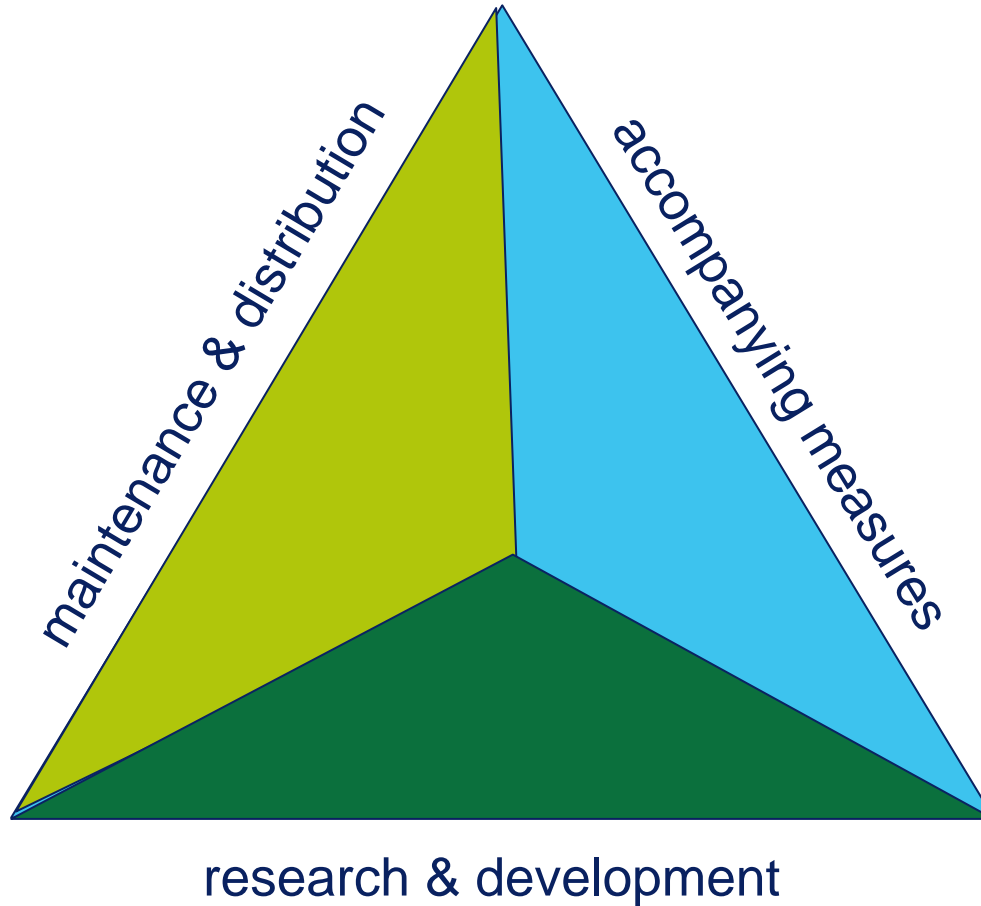
- duration: 2004-2009

calls for R&D projects



calls for demonstration projects

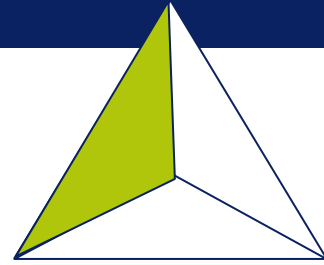
What?



total budget:
11.4 M €

What?

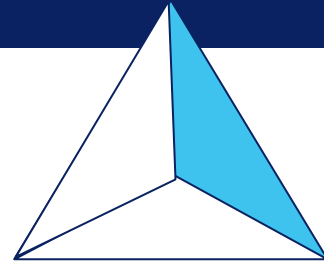
Maintenance & distribution



- HLT Agency
 - ⇒ *several other talks & workshops at LREC*
 - one-stop-shop supplier of resources
 - Intellectual Property Rights

What?

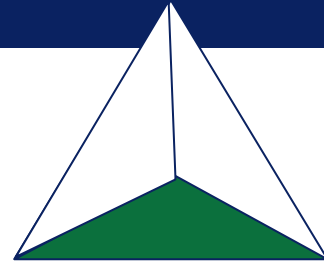
Accompanying measures



- raising awareness & stimulating demand:
 - conferences & seminars (e.g. Taal in Bedrijf)
 - demonstration projects
 - market studies
 - website & newsletters
 - ...

What?

Research & development

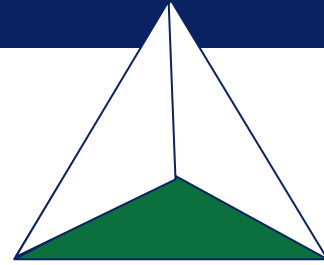


Projects (first call):

AUTONOMATA	improving G2P modules for names
COREA	resolution of coreference relations
D-Coi	preparatory project for Written Dutch Corpus
IRME	acquisition and representation of MWEs
JASMIN-CGN	speech database for children, non-natives and elderly people

What?

Research & development

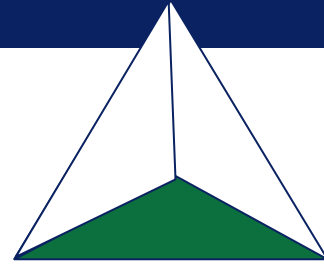


Projects (second call):

DAESO	detecting and exploiting semantic overlap
DPC	Dutch Parallel Corpus
LASSY	Large Scale Syntactic Annotation of Written Dutch
MIDAS	noise robustness in ASR systems
NBest	evaluation benchmark for large vocabulary ASR
STEVINcanPRAAT	improvements and extensions of PRAAT tool

What?

Research & development

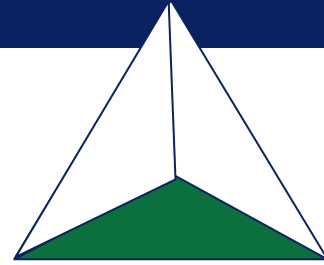


Projects (tender):

SPRAAK	speech recognition toolkit incl. acoustic models
Cornetto	semantic lexical resource for Dutch

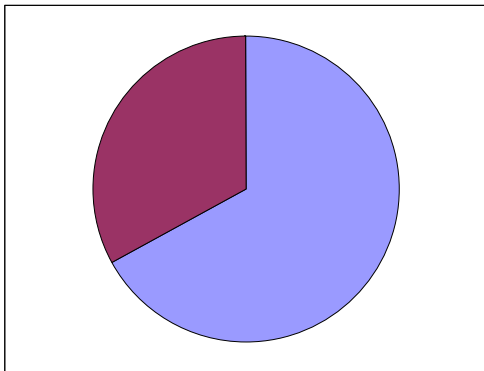
What?

Research & development



Balance: Netherlands vs Flanders

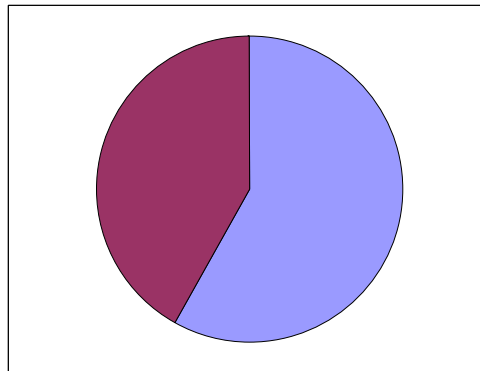
#



Netherlands: 67%

Flanders: 33%

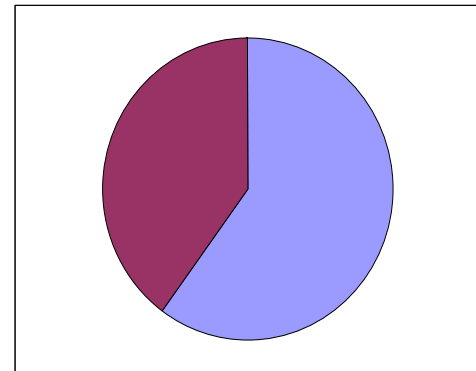
MM



Netherlands: 58%

Flanders: 42%

€

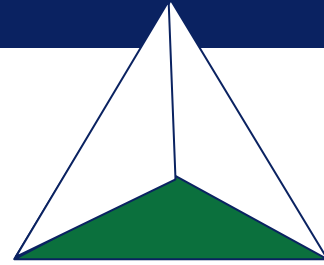


Netherlands: 60%

Flanders: 40%

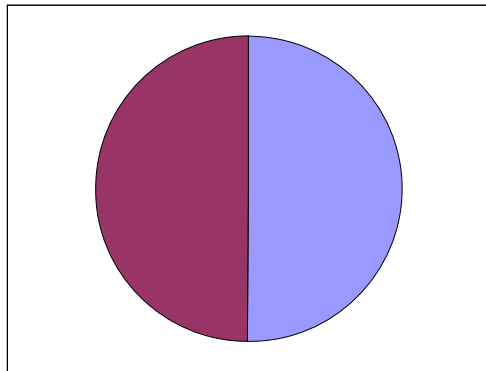
What?

Research & development



Balance: language vs speech

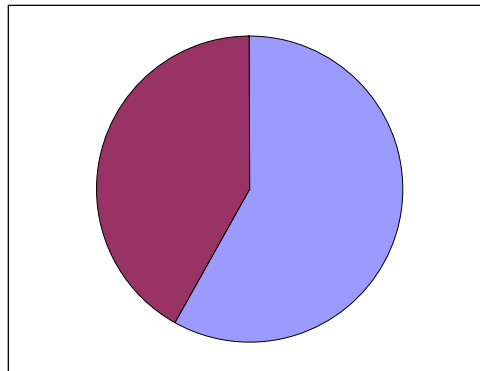
#



language: 50%

speech: 50%

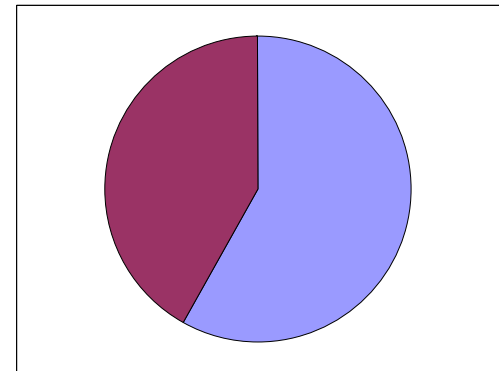
MM



language: 58%

speech: 42%

€

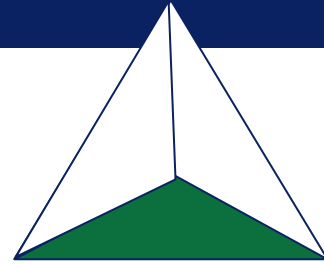


language: 58%

speech: 42%

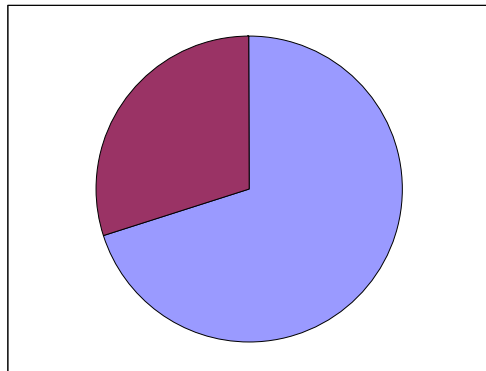
What?

Research & development



Balance: academia vs industry

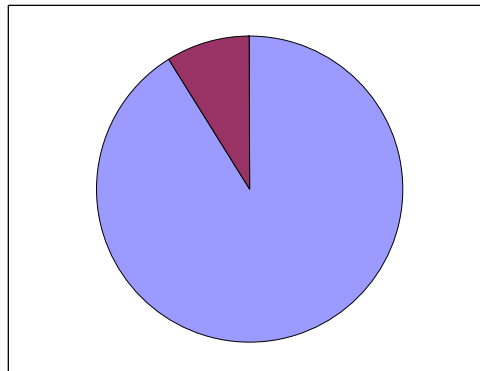
#



academia: 70%

industry: 30%

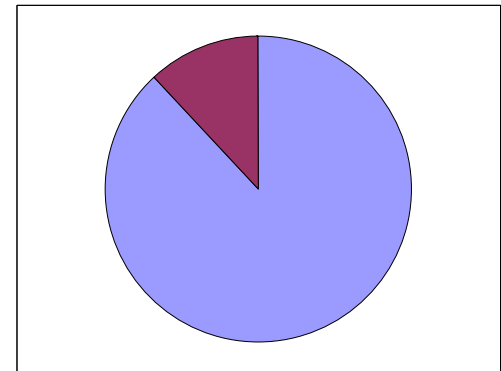
MM



academia: 91%

industry: 9%

€

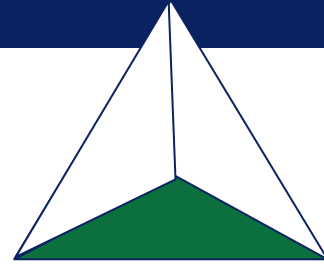


academia: 88%

industry: 12%

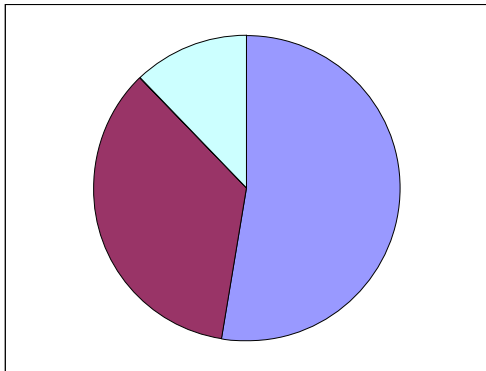
What?

Research & development

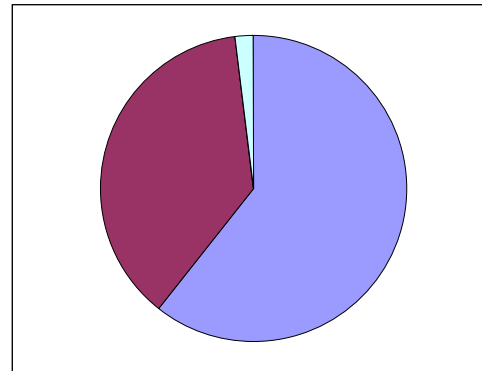


Balance: type of project

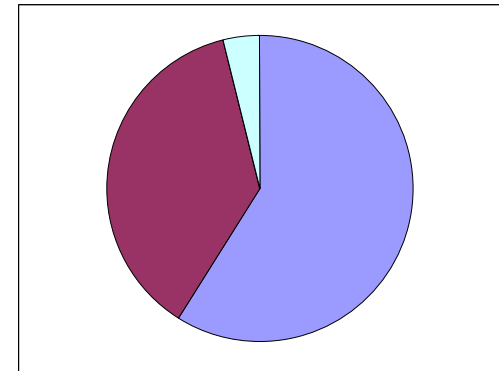
#



MM



€

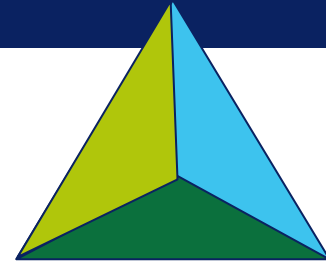


resource creation: 52%
strategic research: 35%
user: 12%
appl. development: 0%

resource creation: 60%
strategic research: 37%
user: 2%
appl. development: 0%

resource creation: 59%
strategic research: 37%
user: 4%
appl. development: 0%

Conclusions



halfway through the programme:

- significant progress towards full-fledged HLT infrastructure
- combining 3 aspects of triangle is fruitful and effective
- well-balanced in most respects; remaining imbalances will be straightened out

<http://taalunieversum.org/stevin/>