# VOORTGANGSVERSLAG VOOR STEVIN PROJECTEN

| | |
|---|---|
| **Projectnaam** | **Large Scale Syntactic Annotation of Written Dutch (LASSY)** |
| **Projectnummer** | **STE05020** |
| **Rapporteringsperiode** | **aanvang project – 1 september 2007** |
| **Deelnemers** | **KU Leuven en Rijksuniversiteit Groningen** |

**Uitvoerders van het onderzoek**

## Samenvatting

## 1.1. Deliverables (+tijdstip) volgens het projectvoorstel

Tot op heden zouden volgens planning de deliverables 1.1 en 2.1 afgerond moeten zijn. Hierover berichten we hieronder in meer detail, en geven daarna een overzicht met betrekking tot de vorderingen voor de overige deliverables.

## Deliverable 1.1 (na 3 maanden)

Deze deliverable bevat de precieze selectie van het kleine LASSY corpus van 1 miljoen woorden.

Deze deliverable is nog niet opgeleverd. De precieze vaststelling van de opbouw van het corpus is afhankelijk van de uiteindelijke versie van het D-Coi corpus en de daarin geannoteerde onderdelen. Het D-Coi corpus is naar verluidt nu bijna zover dat het kan worden opgeleverd, en dan kan in enkele dagen onze selectie worden gecompleteerd.

Het kleine LASSY corpus (handmatig te corrigeren) is 1miljoen woorden groot en omvat het materiaal dat in D-Coi wordt geannoteerd (dit is 500 duizend woorden geannoteerd met POS en lemma, en 200 duizend woorden syntactisch geannoteerd). Voor LASSY wordt dus nog een selectie van 500 duizend woorden gemaakt die nog zowel POStag annotatie als syntactische annotatie nodig hebben. Daarnaast worden 300 duizend woorden syntactisch geannoteerd die in D-Coi al voorzien zijn van POStag maar nog niet van syntactische annotatie. Voor dit laatste deel van 300 duizend woorden is dus geen selectie nodig.

In grote lijnen is de selectie van de 500 duizend nieuwe woorden al gemaakt. Na raadpleging van de gebruikersgroep (begin 2007) is besloten om ons grotendeels te beperken tot zakelijke tekst. Meer in het bijzonder bestaat de selectie uit een groot deel Wikipedia teksten en een groot deel autocues. De Wikipedia teksten zijn met name interessant voor toepassingen op het gebied van question answering, informatie extractie etc. Daarnaast proberen we voldoende materiaal van Vlaamse oorsprong aan de selectie toe te voegen. Besprekingen zijn gaande met het DPC project. Mocht dit niet tot het gewenste resultaat leiden, dan worden uit Wikipedia met name onderdelen geselecteerd met een duidelijk Vlaamse relatie, waarbij hopelijk vaak ook de auteur van Vlaamse origine was (dit is niet makkelijk te achterhalen voor Wikipedia).

## Deliverable 2.1 (na 6 maanden)

Deze deliverable bestaat uit de handmatig te corrigeren toekenning van POS-tag en lemma's voor 250.000 woorden. Omdat de precieze selectie nog niet definitief is (zie deliverable 1.1) is ook deze deliverable nog niet helemaal afgerond. Het goede nieuws is, dat de hoeveelheid woorden die voorzien zijn van POStag en lemma al ongeveer 400 duizend woorden groot is, dus royaal meer dan volgens planning beschikbaar zou moeten zijn.

## Overige deliverables

Het annoteren van POStag, lemma, en syntactische dependentiestructuren verloopt volgens planning. Tot nu toe zijn rond 400 duizend woorden geannoteerd met POStag en lemma (van de te annoteren 500 duizend). Syntactische dependentiestructuren zijn toegevoegd voor zo'n 300 duizend woorden (van totaal 800 duizend).

Voor de deliverables 5 en 6 op het gebied van XML technologie en de case studies is wel enige vooruitgang geboekt, maar vertraging kan optreden omdat het tot op heden niet is gelukt om in Groningen een post-doc voor deze taken aan te trekken. Momenteel worden de opties onderzocht hoe een en ander kan worden opgevangen. Omdat de geplande opleverdata van deze deliverables pas in

de tweede helft van het project zijn gepland, doen we nu nog geen concrete voorstellen op welke wijze de planning moet worden aangepast.

In het overzicht van publicaties beneden worden overigens twee publicaties genoemd die juist voor dit onderdeel van het project belangrijk zijn, ten bewijze dat ondanks deze problemen inderdaad vooruitgang op dit onderdeel van het project werd geboekt. Bouma en Kloosterman tonen aan hoe Xquery gebruikt kan worden om informatie te extraheren uit grote hoeveelheden syntactische dependentiestructuren van het type dat door LASSY wordt geconstrueerd. In het artikel van van Noord wordt aangetoond dat grote hoeveelheden automatisch toegekende dependentiestructuren gebruikt kunnen worden om selectie-restricties te leren, die vervolgens de prestaties van een automatische parser (Alpino) kunnen helpen verbeteren. Deze studie kan gebruikt worden als beginpunt van een van de voorgestelde case-studies.

## 1.2. Reeds eerder afgewerkte deliverables

## 1.3. Gewenste wijzingen (inhoud/tijdstip van deliverables)

## 1.4. De personeelsinzet in relatie tot het oorspronkelijke plan

Afgezien van de problemen met betrekking tot het werven van de post-doc in Groningen, verloopt de personeelsinzet volgens het oorspronkelijke plan.

## 1.5. Disseminatie van resultaten (publicaties, lezingen, …)

**Website**

Er is een begin gemaakt met de website voor Lassy met daarop zoveel mogelijk de beschikbare resourcen. http://www.let.rug.nl/~vannoord/Lassy/

**Publicaties**

Gertjan van Noord, Using Self-trained Bilexical Preferences to Improve Disambiguation Accuracy. In: Proceedings of the Tenth International Conference on Parsing Technologies. IWPT 2007. Prague. Pages 1-10. http://www.let.rug.nl/~vannoord/papers/

Gosse Bouma, Geert Kloosterman. Mining Syntactically Annotated Corpora using Xquery. In: Proceedings of the Linguistic Annotation Workshop. LAW 2007. Prague. http://www.let.rug.nl/~gosse/papers/law07.pdf

**Lezingen**

Gertjan van Noord, Large Scale Syntactic Annotation for Dutch: How and Why. Katholieke Universiteit Leuven, 15 februari 2007.