

Kort verslag workshop E: STEVIN-tender voor Lexicale Semantiek
Tweede STEVIN-bijeenkomst, Antwerpen, 2 maart 2005, 15:30 – 16:30u
Door Jan Odijk, met input van Werner Ceusters

- Werner Ceusters leidde de workshop in. Zie zijn presentatie (ook beschikbaar op de STEVIN-website).
- Commentaar uit de zaal:
 - de formulering van de tender is wel erg breed. Bvb. om te kunnen redeneren vereist een andere opzet/vulling van de databank dan bvb. voor zoeken;
 - het beoogde budget (waarschijnlijk 800 KEUR / 2 = 400 KEUR) en de voorziene doorlooptijd zijn niet in overeenstemming met de beschrijving zoals die er nu staat;
 - zou het niet ingeperkt moeten worden door andere talen als voorbeeld te nemen? Wat bestaat er in andere talen en heeft zijn nut bewezen, en kunnen we dat repliceren voor het Nederlands?
- Inventarisatie van wie de (potentiële) gebruikers zijn en wat hun behoeftes zijn van (potentiële) gebruikers. Genoemd werden:
 - retrieval, inferencing, word sense disambiguation:
 - kan bvb. ingevuld worden met
 - een verbeterde en uitgebreide EuroWordnet voor het Nederlands;
 - met name inclusief verticale relaties;
 - retrieval;
 - diambiguering voor parsing, incl. semantische selectierestricties;
 - disambiguering voor MT;
 - er moet een link met syntactische eigenschappen (bijv. syntactische selectie) zijn;
 - wordt er gedacht aan de mogelijkheid om named entities in de databank te hebben (de wenselijkheid daarvan werd niet beklemtoond)?
 - multi-word expressies moet er ook een plaats in vinden;
 - frequentie van betekenissen in corpora;
 - voor de representatie moet er in ieder geval een voldoende krachtige uitdrukkingstaal ontworpen worden, mogelijk gebaseerd op XML, en dat moet onderdeel van het project zijn;
 - de expansie van woorden moet eenvoudig groter of kleiner gemaakt kunnen worden;
 - de deliverable moet ook software bevatten om bruikbaarheid en uitbreidbaarheid te vergroten;
 - koppeling aan Engels Wordnet is zeer gewenst;
- “Conformereren aan standaarden”, wat betekent dat precies?
 - Bvb. zich richten naar “best practices” zoals Wordnet, EUROWordnet;
 - Wat is de relatie met Semantic Web, bvb. RDF/OWL etc.?
 - Streven naar een theorie-neutrale inhoud en representatie. Maar is dat mogelijk?
- Hoe wordt de databank gevuld?
 - Enkelen benadrukken dat dit menselijke interventie vereist (beroep doen op de kennis van de taal aanwezig bij de mens);

- Volgens anderen is volledig automatische vulling haalbaar, en levert mogelijk zelfs betere kwaliteit (grotere consistentie) dan manuele vulling;
- In ieder geval moet de vulling op de meest efficiënte manier gebeuren, met behoud van zo hoog mogelijke kwaliteit. Mogelijk kan dit door als eerste set een speciale groep zorgvuldig geselecteerde woorden te kiezen, die de vulling voor nieuwe woorden buiten deze set vergemakkelijken/versnellen, etc.;
- Het project moet als hoofddoel hebben het creëren van een lexicaal-semantic database, en geen onderzoeksproject naar methodes voor acquisitie van lexicaal-semantic eigenschappen worden;
- Hoe groot moet de database zijn? Hier moet waarschijnlijk een afweging gemaakt worden tussen kwantiteit (aantal entries) en uitgebreidheid van de eigenschappen per entry.
- Hoe de bruikbaarheid te toetsen?
 - Er moet voorkomen worden dat een database ontwikkeld wordt die volledig getuned is op het type applicatie waarmee getoetst gaat worden;
 - Er werd voorgesteld meerdere applicatietaken als toetssteen te definiëren, en volgens meerdere aanwezigen is dat haalbaar;
 - Het voorstel zoals gepresenteerd door Werner spreekt van een toetsing in een specifiek domein. Maar er werd betwijfeld of dat nu per se noodzakelijk is. Een toetsing per type applicatie/gebruik is voldoende, ook als het niet op een specifiek domein werkt.
- Hoe kunnen we de definitie inperken?
 - Bvb. door specifieke keuzes te maken;
 - Bvb. door ervan uit te gaan dat een redelijk theorie- en applicatieneutrale database gemaakt kan worden (dat werd echter betwijfeld door enkelen);
 - Bvb. door te komen tot een consensus tussen geïnteresseerde partijen (potentiële gebruikers). Joop van Gent zal met andere bedrijven overleggen om te proberen tot een gemeenschappelijk wensenpakket te komen dat toch realistisch is gezien het beschikbare budget en de voorziene doorlooptijd;
 - Er werd verwezen naar <http://openmind.org>, wat mogelijkheden zou bieden op relatief goedkope wijze automatisch database aan te maken. Zouden er dan ook meerdere (goedkopere) projecten gehonoreerd kunnen worden?
 - Bvb. door geen keuzes te maken in de tender, maar eenvoudigweg het beste voorstel te kiezen uit de ingediende voorstellen. “Het beste” voorstel dat goed scoort op de STEVIN-criteria en dat met name het best kan aantonen dat er daadwerkelijk grote behoefte is aan het type lexicaal database dat het voorstel beoogt te maken (en meer dan aan andere types), en dat aannemelijk kan maken dat de beoogde database goed uit te breiden is met andere types semantic informatie die meer geschikt zijn voor andere applicaties), zal het project gehonoreerd krijgen.
- IPR
 - Er waren wat korte vragen over IPR van de resulterende database. De algemene STEVIN-voorwaarden gelden hier.