1. Project Title & Acronym and Abstract

Title: Converting DUELME into LMF format
Acronym: DUELME-LMF
Abstract: The goal of the project is develop a converter from DUELME into LMF format, and vice-versa (for a subset of lexicons in LMF-format), to apply the first converter to create a curated DUELME resource fully compliant with standards supported by CLARIN. A mapping will be defined between DUELME-specific data categories and (possibly newly created) ISOCAT data categories to ensure semantic interoperability of the curated resource with other resources and tools. A document will be produced in which limitations of and desiderata for the LMF standard, ISOCAT and other infrastructural aspects will be described.
Target Start Date: 1 January 2010
Target End Date: 30 June 2010
Type: Resource Curation Project

2. Coordinator

Name: prof.dr. J.E.J.M. Odijk
Function: Professor of Language and Speech Technology
Organization: UIL-OTS, Utrecht University
Address: Janskerkhof 13, 3512 BZ Utrecht
E-mail: j.odijk@uu.nl
Tel: +31 30 253 6730
Fax: +31 20 253 6000
Role(s): User

3. Composition of the Project Team

Name: Drs. R. van Veenendaal
Function: Project leader Flemish-Dutch HLT Agency (TST-Centrale, Centrale voor Taal- en Spraaktechnologie)
Organization: Institute for Dutch Lexicology (INL, Instituut voor Nederlandse Lexicologie)
Address: Matthias de Vrieshof 2-3, 2311 BZ Leiden
E-mail: remco.vanveendendaal@inl.nl
Tel: +31 71-5272495 / +32 32654601
Fax: +31 71 5272115
Role(s): Data Provider

Formatted: English (U.S.)

Formatted: English (U.S.)

Name: to be hired
Function: computational linguist
Organization: UIL-OTS
Role(s): User, Technology Provider

Name: to be hired
Function: (scientific) programmer
Role(s): Technology Provider

4. CLARIN centre

The project results (the curated DUELME resource, and the converters) will be made available on a server of a designated CLARIN centre, more specifically of the Flemish-Dutch HLT Agency or TST-Centrale at INL (http://www.inl.nl). The current version of DUELME, including its GUI, is already available there [2].
INL has a well-established track record in lexicology and lexicography, participated/s in (inter)national projects like DAM-LR, IMPACT, CLARIN and CLARIN-NL and is actively making digital Dutch language resources available through their Flemish-Dutch HLT Agency (TST-Centrale, Centrale voor Taal- en Spraaktechnologie), an initiative of and financed by the Dutch Language Union (NTU, Nederlandse Taalunie). The INL has a strong ambition to become a (type B) Center in the CLARIN infrastructure.

5. Requested Budget: 60,000 Euro

6. Description of the Proposed Project

The goal of the project is develop a converter from DUELME into LMF format and to apply this converter to create a curated DUELME resource fully compliant with standards supported by CLARIN. A second converter, from LMF to DUELME format will allow the use of the DUELME-GUI on DUELME-compliant lexicons in LMF format. A mapping will be defined between DUELME-specific data categories and (possibly newly created) ISOCAT data categories to ensure semantic interoperability of the curated resource with other resources and tools.

1.1 Research Question(s)

The DUELME database makes it possible to address a variety of research questions related to Dutch multi-word expressions, in particular, lexicographic and linguistic research questions on their syntactic structure, composition, and other properties, as well as on their actual occurrence in large Dutch corpora (esp. the Twente Nieuws Corpus). The DUELME database was designed with an eye towards Natural Language Processing Applications, so it naturally facilitates addressing a variety of research questions not only in the area of computational linguistics and natural language processing, but also in the areas of theoretical linguistics and psycholinguistics, esp. concerning an adequate treatment of multiword expressions in grammars and human language processing..

1.2 Research Data

Research Data The research data consist of a lexical database of app. 5000 Dutch multiword expressions. The data were created by UIL-OTS in the STEVIN IRME project [6] in cooperation with the university of Groningen and Van Dale. The database is wellstructured and properly documented [1]. The database was validated by an external organization (Centre for SprakTechnology (CST), University of Copenhagen), and found to be “a skillfully elaborated language resource of a high quality” ([3] p. 3). The errors found by CST (mostly minor ones) have been corrected in the version (1.1) that is currently available via INL. A user interface to view and edit the data has been made available and is in use for on-line consultation of the database via INL.

Current state and format The format of the database is completely idiosyncratic to DUELME. Though the DUELME database was developed only recently, it was decided at the time not to stick to any standardized format yet, for two reasons: (1) the most likely candidate standard (LMF) was still under development and not yet finalized; (2) with the DUELME database an experiment was carried out to create a database in accordance with the Equivalence Class Method (ECM, Odijk 2004), and it was not completely clear at the beginning of the IRME project what form this database was going to take and to what extent LMF would have provisions for properties required by an ECM-based database. At this point in time, however, with the DUELME database finished, and LMF being an official standard, it makes sense to investigate a conversion from DUELME format to LMF format and vice-versa.

State and format of the metadata Though we believe that all relevant information required for a metadata description is available or easily recoverable, no description of the metadata according to any standard or best practice is currently available.

State and format of its documentation The documentation of the resource is elaborate and adequate. The major documents are in English and in PDF format. The documentation includes a description of the encoding protocol, a description of the data formats and the way the data have been obtained and selected, and a document on one of the central components of DUELME, the pattern descriptions. In addition, tab-separated value files are available enumerating all possible values for attributes with a finite number of possible values. Various articles exist describing the design principles behind this database (e.g. [4] and [5]), as well as publicly available internal reports (see [6]).

Annotation schemes used The MWE-patterns are described using a LaTeX-based notation for representing syntactic trees. The labels on nodes and edges in these trees are based on labels used in the Spoken Dutch Corpus (CGN). Morpho-syntactic features are represented in an idiosyncratic set, which to a large extent makes similar distinctions as the CGN-tag set and can probably relatively easily be mapped onto it (though it remains to be seen how perfect the match will be).

IPR and/or ethical issues Ownership of the resource is with the Dutch Language Union (NTU). The resource is maintained and distributed by INL (TST-Centrale) and they and Utrecht University have the rights to modify and enhance the resource, provided that the results will be made available to the Dutch Language Union again. An agreement arranging these matters between NTU and Utrecht University is in place and can be made available upon request.

Familiarity of the participants with the research data Uil-OTS, more specifically a researcher under the supervision of the current project proposal's coordinator (Jan Odijk) designed and created the lexical database. Though the researcher who created the database will not be available for the current project, still a large amount of knowledge and expertise with the DUELME Database is available at UIL-OTS.
INL/TST-Centrale maintains and distributes the resource. UIL-OTS has provided all data, documentation and background documents and it has done extensive knowledge transfer to TST-Centrale employees. INL/TST-Centrale has specific employees dedicated to maintaining and distributing this resource and supporting its users. INL has also expertise with LMF gained in the IMPACT project and in working on LMF versions for the (bilingual) ALVV data. In short, UIL-OTS and INL form the perfect combination to carry out this project.

1.3 Technology

The project aims to develop two software programs that carry out conversions, on the one hand a conversion from the DUELME format into LMF-format, and on the other hand a conversion from LMF into the DUELME format (for LMF-resources for which this is possible). These converters will also use mappings between DUELME-specific tags and labels and ISOCAT-supported data categories, improving interoperability. These converters will be written in such a way that the core conversion functionality is clearly separated from the user interface. A (very likely very simple) API will be defined and implemented for the core conversion module. The interface will be defined in terms of this API. This set-up will make it possible to easily turn the conversion modules later into web-services if this would be desirable.

1.4 Description

Using the converters to be developed, including the mappings between DUELME-specific data categories and ISOCAT-supported data categories will maintain all the opportunities that the research data already provide in its current state for carrying out research in the area of MWEs, but it will in particular improve the (formal and semantic) interoperability between this lexical database and other lexical tools and data, as well as with lexicon-related data (e.g. corpora from which the lexical data have been obtained). The standardization of the data categories using mappings with ISOCAT-supported data categories in particular, will contribute significantly to the semantic interoperability of this resource and other relevant resources.

1.5 Plan

Type: Resource Curation project
We foresee two roles for the execution of this project: a computational linguist (CL), and a (scientific) programmer (SP). In addition, we will require Infrastructure Specialists on specific infrastructure-related aspects (the latter will be provided by CLARIN-NL). We describe here the tasks to be carried out and refer to deliverables and milestones listed in the next section.

Both the CL and the SP have to acquaint themselves with the DUELME database and the LMF standard. We assume that each needs 6 person weeks (PW) for this. After this, both can start working on a specification for a conversion between DUELME and LMF (CL 4 PW, SP 2 PW), resulting in D6. The SP starts preparations for creating a converter and uses the specification D6 as a basis for implementing and testing the converter (8PW) from DUELME to LMF, resulting in D7 (which in this stage can only be tested for formal aspects). The SP goes on to implement the converter from LMF -> DUELME (which will only be possible for a subset of lexical databases in LMF format), resulting in D8 (7.5PW), and creates the necessary documentation (incl. documentation of the API), resulting in D12 (2.5PW). The CL creates new ISOCAT data categories if necessary (milestone D9, 4PW) and creates a mapping between DUELME-specific data categories and ISOCAT data categories (4PW, resulting in deliverable D10. The CL next creates the metadata for the resource (1PW, including persistent identifiers (where we rely on CLARIN-NL to provide a service for issuing them), resulting in D2, which are made available on an accessible INL-server (D3) by the SP, and the CL contributes to carrying out a metadata harvesting test (1 PW, with the help of an IS) resulting in D13. The SP runs the converter, now including the ISOCAT mapping so that semantic operability is improved, and, together with the CL (4PW), tests the result for compatibility (e.g. compatibility with the LMF XML Schema) resulting in the curated resource (D4), which is also made available on an accessible INL server (D5). Finally the CL makes documentation of the curated resource, resulting in D11.

Obviously, in all tasks described above, problems can arise which make a mapping to LMF, a description using the CMDI/IMDI metadata schemata, and mappings to ISOCAT data registries difficult, impose special requirements, or makes certain aspects desiderata of an infrastructure. All of these will be collected and reported, resulting in deliverable D1.

The plan has been represented in a GANTT-chart, which can be made available upon request.

Concerning Infrastructure Specialists (IS), this project requires expertise from Infrastructure Specialists in the areas of LMF, Metadata and Metadata harvesting, and semantic interoperability, in particular ISOCAT.

7. Deliverables and Milestones

Here is the list of deliverables and milestones for this project, including, for each deliverable or milestone a unique identifier (Id), a target date (Wi means i weeks after the project start), its type, a description and the person (role) responsible for it:

Id	Target Date	Type	Description	Who
D1	W26	Document	Document describing requirements and desiderata for the CLARIN infrastructure, justified by the findings of the current project	CL

D2	W23	Metadata	Metadata of the resource (CMDI or IMDI format)	CL
D3	W23	Milestone	Metadata made available on accessible INL server	SP
D4	W26	Data	Curated Resource	SP
D5	W26	Milestone	Curated Resource on accessible INL Server	SP
D6	W10	Document	Specification DUELME – LMF Converters	CL
D7	W16	Software	Converter DUELME -> LMF	SP
D8	W24	Software	Converter LMF -> DUELME	SP
D9	W14	Milestone	ISOCAT extended with new entries in the user space	CL
D10	W18	Data	Mapping DUELME and ISOCAT Data categories	CL
D11	W26	Document	Documentation of the curated resource	CL
D12	W26	Document	Documentation of the converters incl. API description	SP
D13	W24	Milestone	Metadata Successfully Harvested	CL

8. IPR and Ethical Issues: Risks

No IPR or Ethical Issues risks are expected.

9. Expertise of the applicant(s)

Jan Odijk has excellent expertise in linguistic resources in general, and started focusing on lexicons for NLP purposes originally from an MT perspective. He has been involved in a range of projects for the standardization of lexical resources (incl. Multilex, EAGLES and ISLE). He coordinated the project in which the DUELME database was designed and created.

INL/TST-Centrale maintains and distributes the resource. INL/TST-Centrale has specific employees dedicated to maintaining and distributing this resource and supporting its users. INL has also expertise with LMF gained in the IMPACT project and in working on LMF versions for the (bilingual) ALVV data.

10. Project budget details

Salary costs have been assumed to be 4750 per PM (actual costs for an employee in scale 11.0 is 4834.24Euro)

Participant	**Organization**	**Effort (PM)**	**Salary Costs/PM (Euro)**	**Salary Costs (Euro)**	**Travel & subsistence (Euro)**	**Total (Euro)**
To be hired	UIL-OTS	6	4750	28500, 0	1500	**30000, 0**
To be hired	UIL-OTS/INL	6	4750	28500, 0	1500, 0	**30000, 0**
Jan Odijk	UIL-OTS	PM		,0 0	,0 0	**,0 0**
Remco van Veenendaal	INL	PM		,0 0	,0 0	**,0 0**
Total		**12, 0**		**57000, 0**	**3000, 0**	**60000, 0**

11. Literature

[1] http://tst.inl.nl/producten/DuELME/docs/DocumentatieDuELME.zip
[2] http://www.inl.nl/index.php?option=com_content&task=view&id=611&Itemid=667
[3] Bart Jongejan (2007), Validation Report: the IRME database of MWEs - Linguistic Validation, CST, Copenhagen. Obtainable via TST-Centrale and UIL-OTS
[4] Nicole Grégoire (2007) `Design and Implementation of a Lexicon of Dutch Multiword Expressions', in Nicole Grégoire, Stefan Evert and Su Nam Kim (eds.), 2007. 'Proceedings of the Workshop on A Broader Perspective on Multiword Expressions', ACL 2007, Prague, Czech Republic. June 28, 2007, pp. 17-24. http://acl.ldc.upenn.edu/W/W07/W07-1103.pdf
[5] Nicole Grégoire (to appear), *DuELME: A Dutch Electronic Lexicon of Multiword Expressions*, to appear in Language Resources and Evaluation.
[6] http://www-uilots.let.uu.nl/irme/