

## 1 Project Title and Abstract

**Title** Identification and Representation of Multi-word Expressions

**Abstract** The central problems that the project addresses are (i) the lack of large and rich formalized lexicons for multi-word expressions for use in NLP; (ii) the lack of proper methods and tools to extend the lexicon of an NLP-system for multi-word expressions given a text corpus in a maximally automated manner. Therefore, the project aims to develop innovative methods and tools for the automatic identification and lexical representation of multi-word expressions. Concomitantly, a 5000 entry corpus-based multi-word expression lexical database for Dutch will be developed. The database will be externally validated, and its usability will be evaluated in two independent NLP-systems for Dutch.

The project contributes to the development of electronic lexicons, in particular for Dutch. The MWE database to be developed fills a gap in existing lexical resources for Dutch. The project carries out strategic research into generic methods and tools for MWE identification and lexical representation, focusing on Dutch, but these tools will be largely language-independent and can also be used for other languages, new domains, and beyond this project. In this way the project contributes directly to strengthening the digital infrastructure for Dutch.

## 2 Principal Investigator/Co-ordinator

Prof. dr. J. Odijk  
UIL-OTS, Utrecht University  
Trans 10  
3512 JK Utrecht  
The Netherlands  
tel. +31 30 2536076  
fax. +31 30 2536000  
e-mail: jan.odijk@let.uu.nl

## 3 Composition of Research Team

The project will be carried out by a consortium consisting of researchers from UIL-OTS, Utrecht Alpha-informatica, Groningen, and Van Dale Lexicography, Utrecht.

- Applicants
  - Prof. dr. J. Odijk (Co-ordinator), UIL-OTS, Utrecht University
  - dr. G van Noord, Alpha-Informatica, Groningen University
  - dr. G. Bouma, , Alpha-Informatica, Groningen University
  - Dr. A. Schenk, Van Dale Lexicography, Utrecht
- Researchers
  - B. Villada Moirón, Alpha-Informatica, Groningen University
  - post-doc (computational) linguist, UIL-OTS, Utrecht (candidate not known yet)

- programmer, UIL-OTS, Utrecht (candidate not known yet)
- lexicographer, Van Dale (candidate to be determined)

The applicants and researcher from Groningen will focus their research on methods and tools for automatic identification of multi-word expressions in text corpora, and for testing the lexical representation method in the Alpino system.

The applicant and researcher from Utrecht will focus their research on methods and tools for a standard lexical representation of multi-word expressions, on the development of a lexical database for multi-word expressions, and on testing the lexical representation method in the (Dutch part of the) Rosetta system. The programmer in Utrecht will make a running version of the Rosetta system, and will contribute to the development of tools for (semi-)automatic incorporation of multi-word expressions into NLP systems in general and the Rosetta system in particular.

Van Dale will contribute by (1) making available data to the research project for evaluation purposes; (2) by bringing in its excellent expertise in the area of lexicography, esp. with regard to multi-word expressions in the form of advising and providing feedback to the academic participants

The project aims to develop methods and tools for the automatic identification and lexical representation of multi-word expressions and to evaluate these in different NLP-systems for Dutch. The two aspects of identification and lexical representation are both essential ingredients to increase the capabilities of a wide range of NLP-systems to deal properly with multi-word expressions.

The expertise available in Groningen relevant to this project encompasses robust syntactic parsing of Dutch, the Alpino system, and the automatic identification of multi-word expressions and their properties in text corpora. The expertise available in Utrecht relevant to this project encompasses syntactic analysis and parsing (including multi-word expressions) in the Rosetta system, standards for lexicons, and the lexical representation of multi-word expressions. The expertise of Van Dale, one of the most renowned publishers of dictionaries in the Dutch-speaking region, relates to lexicography and computational lexicography and they bring in a substantial amount of multi-word expression data for evaluation purposes

It is clear that the expertise available in Groningen and Utrecht fits in perfectly with the project goals, and is largely complementary. This, complemented by the active participation of Van Dale and their making available a large set of multi-word expressions for evaluation purposes, creates an ideal combination of consortium partners for this project.

Prof. Odijk will be the coordinator of the project. Co-ordination and management of this project will be described in more detail in section 6.5

## 4 Requested Budget

**Duration of the Project** 2 years

**Targeted Start Date** January 1, 2005

**Budget** kEURO 410

## 5 Stevin priorities

The project contributes to the development of electronic lexicons. First, a corpus-based, independently validated lexical database of multi-word expressions for Dutch tested in two NLP-systems for

Dutch will be developed, filling a gap in the existing lexical resources for Dutch, and directly contributing to the digital infrastructure for Dutch. Second, the methods and tools developed can be used to support the development of such lexicons, and for rapidly tuning NLP lexicons to specific domains (adaptation), not only for Dutch but also for other languages. Indirectly, the project also contributes to robust syntactic analysis, in particular of multi-word expressions.

## 6 Description of the Proposed Research project

### 6.1 Scientific Aspects and Innovative Power

The research project aims to stimulate the development of rich electronic lexicons for Multi-Word Expressions (MWEs) suited for NLP systems and rapid tuning of NLP-lexicons to specific domains for MWEs (*adaptation*). This project therefore focuses on the following two problems:

- *Automatic identification* of multi-word expressions and their properties
- *Standard lexical representation* of multi-word expressions and their properties

With respect to automatic identification, we propose to develop novel computational lexicographic methods and associated tools which combine rich linguistic information with statistical techniques to extract MWEs and their properties from richly annotated text corpora. Even though the proper treatment of multi-word units is an unsolved problem for the automatic syntactic analysis of Dutch, it is also true that automatic syntactic analysis (of a somewhat shallower nature) can be extremely useful in order to identify multi-word expressions, as well as their various properties (Villada Moirón, To Appear).

With respect to the lexical representation of MWEs, we intend to develop a standard representation for multi-word expressions building on the *Equivalence Class Method* which was recently proposed (Odiijk, 2004b,a). This innovative method provides (1) a lexical representation which is independent of any particular grammatical framework or NLP system, and (2) a generic method to incorporate these lexical representations into specific NLP-systems. The purpose of this project is to elaborate this method, to thoroughly test its empirical validity and its applicability, and to extend it to other types of MWEs than idioms for which it was originally proposed.

Concrete results that can be expected of the project are:

- Novel methods and associated tools for identifying and extracting MWEs from richly annotated text corpora
- A novel method for representing MWEs lexically in a way that is as independent from specific grammatical frameworks, theories or their implementations as possible. This method therefore qualifies as an excellent candidate for a standard lexical representation of MWEs
- A novel method and associated tools to incorporate the MWE lexical representations into NLP-systems. The method will be tested using two NLP systems, the Alpino system and the (Dutch part of the) Rosetta system
- A lexical database of MWEs of Dutch, in accordance with the lexical representation method mentioned above, consisting of 5000 MWEs.
- New insights into the automatic acquisition of MWEs and their properties, and their lexical representation for NLP purposes, to be published in reports, journal articles and/or conference contributions.

### 6.1.1 Background

**Multi-word expressions** Multi-word expressions are sequences of words that have lexical, syntactic, morphological, semantic, or pragmatic properties which cannot be deduced from the individual words or from general syntactic rules. Therefore, those properties must be stipulated in the lexicon. State-of-the-art natural language processing systems have difficulty with multi-word expressions, and it has been argued (Sag et al., 2001) that multi-word expressions are an important bottleneck for successful NLP applications such as information retrieval, machine translation, question answering and automatic summarization.

Several NLP-systems exist that claim to be able to deal with certain MWE types<sup>1</sup>, also for Dutch (Alpino, Rosetta). However, as stated this usually is true only for specific subtypes, and most systems only have small lexicons for MWEs that are generally not easy to extend due to lack of resources and the rich property set required to deal with MWEs.

Although for Dutch a number of electronic lexicons is available (e.g. Celex, Parole, RBN), the description of multi-word expressions in these resources is either not available at all, or available only for a small subset of expressions. In addition, the available information associated with multi-word expressions is insufficient. Traditional dictionaries contain a lot of multi-word expressions, but usually in an insufficiently formalized way to be directly useful for NLP-applications. Finally, new domains generally come with their own domain-specific multiword terminology, which generally is not covered by lexicons for general language, and only partially by domain-specific lexicons and glossaries. The latter often have a normative function, do not represent actual usage and do not cover newly emerging multi-word terms that continuously arise in rapidly developing domains.

A broad variety of expressions qualify as multi-word expressions, e.g. multi-word terms, idioms, collocations, support verb constructions, phrasal verbs, sayings, etc. Multi-word expressions constitute a non-negligible portion of the expressions existing in a language.<sup>2</sup> Furthermore, frequency figures reveal that these expressions are too frequent to be left untreated in an NLP system (Sag et al., 2001).

MWEs show idiosyncracies at different levels of analysis, i.e. lexical, morphological, syntactic, semantic and pragmatic: components of an MWE cannot be replaced by a synonym; components of an MWE must often occur with specific morphology (e.g., only in singular); insertion of modifiers is highly restricted, there are restrictions on syntactic flexibility, the meaning of the MWE is often not compositional (Nunberg et al., 1994). At the same time, all these restrictions do not imply that they behave as a fixed sequence of words such as e.g. *ad hoc*. In fact, open argument slots and restricted variation may be possible, although this is not uniformly observed across all types of multi-word expressions. As examples, the expression *iemand in het vaarwater zitten* or *in iemand's vaarwater zitten* denote the same meaning 'to work against s.o.' but they are structurally different. *Iemand aan de/zijn tand voelen* 'to grill s.o. (by asking lots of questions)' allows both a definite *de* or a possessive determiner *zijn*. The mentioned realizations are found in corpora of current Dutch, thus the corresponding multi-word expression description in a lexicon ought to take such variants into account (Moon, 1998; Riehemann, 2001; Sag et al., 2001; Villada Moirón, To Appear). These facts have been captured in existing formal systems by treating multi-word expressions as partially lexicalized expressions (Sailer, 2000; Riehemann, 2001; Odijk, 2004b; Schenk, 1994).

**Identification of MWEs** Large corpora are potentially a rich source of MWEs and an invaluable collection of their linguistic behavior. In specific domains within a company or organization, electronic

<sup>1</sup>See e.g. Wehrli (1998); Abeillé and Schabes (1989); Breidt and Segond (1995); Riehemann (2001).

<sup>2</sup>Estimates vary from as many as there are single words to 10 times as many ((Jackendoff, 1995; Mel'čuk, 1995) )

corpora are usually available for tuning a generic NLP-system to these domains for domain-specific MWES, provided appropriate tools for doing this exist.

The task of identification of MWES in corpora has to tackle a few problems: (1) finding word combinations that show the idiosyncratic behavior of MWES –as opposed to the productivity and syntactic regularity of other expressions; in concrete, problematic aspects are the non-uniform external form and internal morpho-syntactic structure, the fact that component lexemes in MWES may be non-adjacent to each other, and the non-compositional meaning; (2) identifying low frequency MWES in corpora; and (3), establishing when an identification model is sufficiently effective (validation).

Developments in NLP technology have provided data that is automatically annotated with linguistic information. Such developments enabled models that extract corpus patterns that satisfy certain morpho-syntactic requirements. So-called hybrid models incorporate linguistic information and statistics.

Villada Moirón (To Appear) has applied hybrid models to automatically identify Dutch support verb constructions in large corpora. Corpora annotated with the Alpino parser (van der Beek et al., 2002a) has been crucial to extract potential support verb constructions. Candidate patterns vary in length thus, a simplified pattern representation is adopted. Patterns longer than 2 words are represented as bigrams (two word combination) so that common probabilistic statistics can be applied to measure the association strength between the component words. This association strength approximates the lexical affinity between the words.

Whereas lexical affinities pose no problem, the restricted syntactic flexibility or the non-compositional meaning of MWES cannot be captured by simple statistical tests. Thus, systematic errors made by the statistical model are discarded by applying linguistic diagnostics. This filtering mechanism exploits the idiosyncratic syntax and/or semantics of support verb constructions thus, producing an error rate decrease of 24.7% and consequently, significantly improving the accuracy of the automatic identification models (Villada Moirón, 2004a,b). The hybrid model provides a list of patterns that consist of the minimum required lexemes in an MWE (e.g. *op gedachten brengen* represents the base form of the expression *iemand op gedachten brengen* ‘give s.o. the idea’).

In a next step, a corpus query tool (Bouma and Kloosterman, 2002) retrieves evidence of morpho-syntactic variation and modification in those instances of the patterns found in syntactically annotated corpora. The extracted evidence shows a distinction between various types of support verb constructions that range from totally fixed to flexible expressions. Such evidence is crucial to improve the description of the mentioned patterns in lexical resources.

**Lexical Representation of MWES** After identifying and extracting MWES and their properties, the MWES and their properties must be represented lexically, so that they can be incorporated into and used in NLP systems.

The problem with the lexical representation of MWES is that NLP-systems that are able to deal with MWES require rich and highly-system specific lexical representations. This has been clearly shown by (Odijk, 2004b) using the Rosetta machine translation system as illustration. The latter system requires, for idiomatic expressions, (1) reference to a highly Rosetta-specific syntactic structure, and (2) a sequence of references to lexical entries of the lexicon of the Rosetta system. In this sequence the presence/absence of these references, the order in the sequence, and the references themselves are all highly specific to the Rosetta system.

Lexical representations for MWES that are highly specific to one particular grammatical framework or even to one specific implementation are undesirable, since it requires effort in making such lexical representations for each new NLP system again and reuse of significant effort is not possible. In

addition, though the identification methods to be developed in this project, if successful, will make it possible to identify MWEs and their properties, the results of this identification will generally not easily fit the system-specific requirements of an independent NLP-system.

In order to overcome these problems, Odijk (2004b) proposes a method of lexically representing idiomatic expressions that is maximally independent of any particular grammatical framework or specific implementation. In this method (called the Equivalence Class Method, ECM) idiomatic expressions are partitioned into equivalence classes in such a way that members of the same equivalence class have identical syntactic structures. Each idiomatic expression is represented by (1) the name of its equivalence class; (2) a list of lemmas for the lexical items making up the idiomatic expression (in the same order within each equivalence class); (3) an example sentence (where the syntactic structures of all example sentences from the same equivalence class are identical modulo lexical items).

Using such representations, Odijk shows that by manually incorporating one instance of the equivalence class into a specific NLP system, all other members of the equivalence class can be incorporated in a fully automatic manner.

### 6.1.2 Research aims

**Identification** Villada Moirón’s (VM) method to extract Dutch support verb constructions still suffers from a few limitations. Although VM’s method applies standard statistical tests to candidate patterns of length three and of any frequency, it is desirable to have an identification model that generalizes to patterns of any length and that is not too sensitive to low-frequency data. The filtering mechanism used to discard systematic errors fails to distinguish actual support verb constructions from combinations of a verb with idiosyncratic phrases (directional, predicative, etc.); these phrases behave in some respects like dependents in support verb constructions but their use is not restricted to a specific predicate. Due to the lack of a large amount of validation data, VM’s method has not been tested at a large scale; in order to verify its applicability to other MWE types, this method requires further testing.

In order to extract reliable evidence of modification and other morpho-syntactic restrictions, VM’s corpus-based method assumes knowledge of the valence patterns of each expression. Furthermore, the output of this corpus exploration method needs to be manually supervised to ensure that the MWE interpretation is present.

After identifying expressions with a strong lexical affinity between component words, state-of-the-art approaches tried to capture the semantic non-compositionality of English phrasal verbs (a subtype of MWEs) by applying semantic clustering techniques such as latent semantic analysis and other models that make use of ontologies such as WordNet or thesauri (Baldwin et al., 2003; McCarthy et al., 2003). These models are built on the idea that the context around a MWE differs from the context around a fully regular and compositional expression. So far, these models have yielded poor results nevertheless, their potential for identification of (other types of) MWEs is largely unexplored. The main drawback of these models is the need of a large amount of data.

Building on previous research on automatic identification of support verb constructions, we will apply hybrid models to identify other multi-word expression types in large corpora. Following Villada Moiron’s (To Appear) work, we choose for a bottom-up approach. In this order, models identify the minimum required lexemes, the valence patterns of each multi-word expression and other morpho-syntactic restrictions of the lexical items or the phrasal constituents that are observed in large corpora. We expect the most difficult part to be the acquisition of the minimum required lexemes.

Villada Moirón’s preliminary studies show that morpho-syntactic irregularities clearly distinguish support verb constructions from regular verb phrases. The focus of the project is to determine which

properties of multi-word expressions can best be captured by identification models leading to a successful performance?

Departing from previous work, we aim at investigating more expressive statistical models (maximum entropy models, latent semantic analysis and support vector machines) that capture the interaction of features across various linguistic levels. What we mean is that a model, ought to infer potential interactions between lexical affinities, defective morphology, rigid syntax (phrase internal and sentential), etc. Furthermore, the models should be applicable to any type of MWE. We will build models that first infer significant partial interactions; next, we investigate higher order interactions that include the significant partial interactions across linguistic levels. We expect that significant interactions show a distinction between regular productive expressions and MWEs.

With the knowledge acquired in the first phase, we seek to acquire the valence pattern of each multi-word expression, the morpho-syntactic irregularities of its lexical items and variants of the expressions observed in large corpora. Techniques successfully applied in acquisition of subcategorization information of English verbs (Briscoe and Carroll, 1997) will be tried. In the next stage, we automatically extract other constraints that affect the lexemes, their morphology, phrase syntax, particulars concerning open slots, agreement relations, etc.

**Lexical representation** The ECM method has originally been developed for idiomatic expressions, one subtype of MWEs, and initial experiments have been carried out to determine the applicability of the method, with promising results. But the method still requires testing against a large set of idiomatic expressions to give it a solid empirical basis. This requires research into idiomatic expressions of Dutch and the various forms in which they can occur, guided as much as possible by the outcome of the identification methods. It also requires a classification of MWEs into equivalence classes, taking into consideration many different grammatical frameworks and specific implementations of grammatical frameworks. Many types of Dutch idiomatic construction require special attention and research, including idioms that are or contain negative polarity items, idioms containing inalienable possession constructions, idioms with optional arguments, idioms containing copular verbs, and idioms containing ‘small clauses’, etc. For all of these (and others), the proper analysis and lexical representation is not obvious and requires research.

Odiijk shows that the method shows promising results for idiomatic expressions assuming so-called *parameterized* equivalence classes. In the project, this parameterized approach will have to be elaborated and formalized, and the equivalence class parameters required for Dutch will have to be established.

In addition, it has to be investigated whether and how this method can be extended to other types of MWEs, in particular to semi-idioms (‘collocations’, in the sense of (Mel’čuk, 1995)) and support verb constructions.

In order to carry out this investigation, a lexical database for Dutch MWEs consisting of 5000 MWEs and their properties will be constructed. It will serve as an empirical basis to conduct the research and will yield an independent result of the project as well.

Though the ECM is a promising method for lexically representing MWEs, there is of course no guarantee that the method will work properly for all types of MWEs. This is of course inherent to research, in which new methods are being explored. However, we would like to ensure that the project yields a useful concrete MWE database as one of its results even if the ECM would turn out not to work in all cases, or at all. In this connection, we can point out that even small-scale experiments already show that the equivalence classes used should be defined in a way that is transparent to computational linguists: just assigning an identifier to each equivalence class makes it very difficult to manage

the data as soon as they start to grow in size. Underspecified syntactic structures can be used to provide a structured yet limited way of creating new identifiers for equivalence classes. In addition, using such underspecified syntactic structures has the advantage that if the equivalence class method would not work properly, still a useful (though not so theory and implementation-independent) lexical database will result, comparable in nature to the SAID database developed for English (Kuiper et al., 2003). The project will investigate the use of such underspecified syntactic structures, the nature of the underspecification required, and the optimal way to use them.

### **6.1.3 Innovative Power**

The project is innovative in many respects. In the area of automatic identification and lexical acquisition (i) methods are extended to deal with complex types of MWEs (discontinuous, with variation in word order, etc.); (ii) novel statistical techniques will be investigated which set no constraint on the size of the candidate expressions; (iii) the acquisition models to be explored are the first ones in attempting to model irregularities at various linguistic levels simultaneously, not only for Dutch but for any other language.

In the area of lexical representation, an innovative method that can truly claim maximal independence of grammatical framework, specific MWE theories and their implementation in specific NLP systems is being investigated. If successful, it has the potential to significantly contribute to better treatment of MWEs in a wide variety of NLP-systems.

Finally, the project combines two essential ingredients for properly dealing with MWEs in NLP-systems that have to operate on real-word texts, viz. identification and lexical representation of MWEs. Though the research will be done on Dutch, the project will yield methods and tools that can also be applied to other languages and to new domains within Dutch not covered by the current project.

## **6.2 Economic Aspects**

State-of-the-art NLP systems work best if they are tuned to a specific domain or sublanguage, and rapid adaptation of a generic NLP-system to a specific domain or sublanguage with its own multi-word terminology is essential for a successful deployment of such NLP-systems. The Dutch NOTaS group (<http://www.stichtingnotas.nl/>), a conglomerate of commercial and academic NLP-developers points out that a proper treatment of MWEs is of great importance, e.g. in the processing industries (Akkermans et al., 2004, 118). The current work in many research institutions is too much focused on work concerning the individual head words only, and on general rather than domain specific language. Similar statements are made by other companies such as LinguaTec (e.g. (Thurmair, 2003, 2004)). Also Van Dale has shown interest in this topic and actively participates in the project by making available its expertise and providing valuable data for evaluation purposes.

We have requested a number of people from academia and industry to participate in a user group (possibly not for this project alone), so that we get regular feedback on the right directions the project should take. A number of these people have expressed their willingness and ability to participate in such a user group. These include Gregor Thurmair (Linguatec, Germany), Ann Copestake (Cambridge), and Nicoletta Calzolari (Pisa). For some others we are awaiting their response.

This clearly shows that the problems being investigated in this project are directly relevant to increasing the successful deployment of NLP systems in industry.

### 6.3 Contribution to the STEVIN-programme

The project contributes to the development of electronic lexicons, in particular for Dutch. It does so by directly contributing to the digital infrastructure for Dutch (MWE lexical database) and by carrying out strategic research in the area of lexical acquisition and representation of MWEs, which, if successful, will lead to innovative methods and tools in this area. The methods and tools to be developed can be used to support the development of MWE lexicons, not only in this project but also outside of this project, e.g. when specific MWE-lexicons have to be made in the process of tuning a generic NLP-application to a specific domain. In this way the project contributes to *adaptation* of NLP-technology. By applying the tools developed in the project, and by exploring the methods of the ECM, a lexical database for Dutch MWEs will result which has wide applicability in a range of NLP modules that may differ in the grammatical framework used, the specific theory of MWEs adopted, and the specific implementation details, filling a gap in existing lexical electronic resources for Dutch. Indirectly, the project also contributes to robust syntactic analysis, in particular of multi-word expressions.

### 6.4 IPR and standards

The methods explored will be made public via the normal publication channels (journal articles, conference contributions, etc.). The tools to apply the extraction models, the tools that are necessary to support the ECM, and the lexical database of MWEs created in this project will be made publicly available via the TST-centrale.

The data that Van Dale will provide will be made available to the research groups of Utrecht and Groningen for evaluation purposes and its usage is restricted to the current project. These data will not be made available to the TST-centrale. Van Dale will have early and free access to the results of the project (reports, publications, tools, MWE database), also for use after the project has ended. Utrecht and Groningen will conclude an agreement with Van Dale (if the project gets approved) to formally and unambiguously arrange these matters.

Concerning the Rosetta system, all the source code is available in Utrecht, and the rights to use this source code for research purposes in Utrecht have been secured.

Several international projects have proposed standards for the lexical representation of (specific subtypes of) MWEs. Example are the ISLE project<sup>3</sup> and the XMELLT<sup>4</sup> project. To our knowledge, these proposed standards have not yet been actually used on a large scale. Both proposed standards also have disadvantages, in that they are limited to specific subtypes of MWEs and in that they are not really neutral with regard to grammatical framework adopted. The proposed ISLE standard allows for a fine granularity, but though that is on the one hand an advantage, it is at the same time a problem since compatibility of independently created representations is not guaranteed, making the use of this proposed standard more limited. The ECM is, to our knowledge the first method that qualifies as a potential standard that overcomes these problems. In that sense, this project directly contributes to the development of a new standard for the lexical representation of MWEs.

### 6.5 Coordination and project management

Coordination and project management will be carried out by

- regular status meetings of the research groups involved (once every 3 months, or linked to specific delivery dates), physically and/or by telephone conference

---

<sup>3</sup>[www.ilc.pi.cnr.it/EAGLES96/isle/ISLE\\_Home\\_Page.htm](http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm)

<sup>4</sup>[www.cs.vassar.edu/~ide/XMELLT.html](http://www.cs.vassar.edu/~ide/XMELLT.html)

- exchange of information on a daily basis between the researchers via a network
- supervision and guidance of the researchers by the applicants

Each responsible for a work package has the duty to make sure there is progress and timely delivery of the deliverables with the required quality, to report regularly and especially to warn the coordinator in case of problems that might lead to delays.

A separate work package has been set up for coordination and project management. The principal investigator, Jan Odijk, is responsible for this work package.

## 6.6 Evaluation, validation and success criteria

**Evaluation** Rigorous full-scale evaluation of the identification models will be carried out as follows.

The accuracy in extracting the minimum required lexemes, valency pattern and (perhaps) morpho-syntactic constraints, will be evaluated against the (small) MWE database created for NLP purposes by Jan Odijk (Odijk, 2004a), while, the accuracy and coverage of the models will be evaluated against the MWE database made available by Van Dale, and any other MWE lists or databases that become available. This will make it possible to establish how successful the identification models are.

Concerning the lexical representation of MWEs, the ECM will be evaluated by testing whether it can be successfully used for the purpose it was designed for: semi-automatic incorporation of lexical representations into NLP systems. We will test this by applying the incorporation method to the Alpino system, and to the Rosetta machine translation system (Rosetta, 1994). In order to test it on the Rosetta system, we plan to make a running version of this system (which is currently not available). All the Rosetta source code is available in Utrecht, and the rights to use this source code for research purposes in Utrecht have been secured.

**Validation** We will have the MWE lexical database resulting from this project validated by an independent organization. The Center for Sprogteknologi (CST), University of Copenhagen, which is ELRA's validation unit for written resources and has extensive knowledge of and expertise with validating lexical resources, has declared it is willing and able to carry out this validation. We intend to involve CST at an early stage in the project so that validation is an integrated aspect of the creation of the lexical database, and we have reserved a budget to remunerate CST for the work involved. A rough characterization of the validation work to be done has been agreed upon, and though no detailed specification of the validation task is available yet, CST confirmed that this budget is realistic.

**Success Criteria** We consider the project a success if (1) new insights into the (semi-)automatic acquisition and lexical representation of MWEs and their properties for NLP-purposes have been obtained and laid down in 4 publications; and (2) the project delivers a high-quality 5000 entry MWE lexical database for Dutch, independently validated, which is as neutral as possible with regard to grammatical framework, theory or specific implementation; and (3) this database has been successfully tested in at least one NLP-system for Dutch.

Of course, we aim higher. We target to obtain not only insights but actually concrete methods and supporting tools to (semi-)automatically acquire and lexically represent MWEs. And we target to reduce the manual part of these methods as much as possible. And we target for testing the database to yield successful results for both NLP-systems mentioned, and to indeed obtain a maximally neutral lexical representation of MWEs, based on the ECM or a modification thereof. But these are all new aspects that require investigation, and though initial experiments show promising results and we are

confident that the approaches we propose are promising directions, there is no guarantee of success in these areas.

## 7 Work Programme

### 7.1 Work Packages

The project has been subdivided into a number of work packages (WP), each with the persons/organizations involved, a clearly identified responsible, and a budgeted effort or cost:

WP	Description	Effort/Cost	Responsible	other participants
WP0	Coordination and Project Management	2.4MM	Jan Odijk	Gertjan van Noord, Gosse Bouma
WP1	MWE Identification	22MM	B. Villada Moirón	
WP2	MWE Lexical Representation	22MM	Linguist Utrecht	
WP3	Rosetta running version	4 MM	Programmer Utrecht	
WP4	ECM Incorporation Tools	2 MM	Programmer Utrecht	
WP5	Evaluation with Alpino	2MM	B. Villada Moirón	
WP6	Evaluation with Rosetta	2 MM	Linguist Utrecht	
WP7	Validation of the MWE database	10kEURO	Linguist Utrecht	CST
WP8	Preparation of Van Dale MWE database	1MM	Van Dale	
WP9	Evaluation by Van Dale	1MM	Van Dale	

### 7.2 Deliverables

The deliverables are specified in table 1.

The table specifies the deliverable number<sup>5</sup>, the delivery date<sup>6</sup>, the deliverable, the responsible, and the dependencies on other deliverables.

In WP1, first a specification of the models to be used for MWE identification will be made (D1.1). The two models being investigated will then be evaluated (D1.2, D1.3) and compared to each other (D1.4). Next, a specification of tools to extract grammatical properties of MWEs will be made (D1.5). The tools will be implemented and applied to text corpora, yielding a set of automatically acquired data (D1.6), in a format to be agreed upon between the participants. The automatically acquired data will be integrated into an ECM-based MWE database (D1.7).

In WP2, first the ECM method will be further elaborated, especially with regard to parameterization for Dutch (D2.1). Next, the ECM will be extended to semi-idioms (D2.2), and to support verb constructions (D2.3). The empirical material for this will come from Odijk's database, the data delivered by Van Dale, and a separate list of MWEs available in Utrecht. An initial version of an ECM-based MWE database will be produced. With the input from the automatically acquired data from WP1, a the initial (unvalidated) ECM-database will result (D2.4).

In WP3, a running version of the Rosetta system will be created.

In WP4, tools to support incorporation of ECM-based databases into NLP-systems will be developed and tested for the Rosetta system (D4.1).

<sup>5</sup>( $D_{i,j}$  means deliverable  $j$  of WP  $i$ )

<sup>6</sup>( $T_i$  means  $i$  months after the start of the project)

In WP5, the integration of the ECM database into Alpino will be evaluated (D5.1), and the same is done in WP6 for Rosetta (D6.1).

In WP7, the ECM-based MWE database will be validated by CST (D7.1).

In WP8, Van Dale will prepare its data for delivery to the project (D8.1).

In WP9, Van Dale will contribute to the evaluation of the automatically acquired data and the ECM database (D9.1).

## 8 International Perspective

Computational lexicographic projects aimed at expanding English and German dictionaries with collocations and other types of MWEs make use of automatically annotated corpora and statistics. Examples are the WASP project by Adam Kilgarrif and David Tugwell, CPA by Patrick Hanks and TFB by a team lead by Ulrich Heid at IMS in Germany. Such projects were carried out under a close collaboration between academic and industrial partners. The academic sites provide large collections of candidate MWEs and the industrial partners perform the validation and decide the adequate lexical representation in the dictionaries. The type of target information has changed from mere word collocates and their co-occurrence frequency in the earlier attempts, to word collocates, valence patterns and morpho-syntactic requirements in the latest efforts. Identification models applied in these projects use shallow annotated corpus data, with the exception of Spranger's (2004) approach that uses a chunk parser. Language specific constraints (s.a. free word order) and the type of MWEs aimed at enforce certain requirements in identification models. This motivates our decision to explore models that set no constraint on the external form or length of the MWEs.

In the last decade several projects have aimed at developing standards for the representation of lexical items, e.g. Multilex, EAGLES, and ISLE. Though these projects were carried out by large consortia of mainly academic but also industrial partners, the results of these projects in establishing a standard have remained limited to the area of single word lexical items. Only the ISLE project dealt with standards for the lexical representation of multi-word expressions, as part of their proposals for multilingual standards, in close collaboration with the XMELLT project. There are several projects, both in the past and ongoing, that are related to the topics dealt with in this project. In particular we mention:

**LINGO** an MWE Research Project (<http://lingo.stanford.edu/mwe/>) carried out at CSLI (Stanford), in Cambridge UK and at NTT in Japan

**SAID database project** in which a large database of syntactically annotated idiomatic expressions for English has been created (Kuiper et al., 2003)

**Kollokationen im Wörterbuch** See <http://www.bbaw.de/forschung/kollokationen/>. One of the activities in this project is a categorization of a large set of German MWEs by syntactic patterns (Geyken, 2004; Stantcheva, 2004).

Finally, we can mention the research activities on MWEs by prof.dr. Martin Everaert, the UIL-OTS scientific director, in his capacity as professor at the Radboud University, Nijmegen (Everaert, 2003). The project will certainly benefit from his knowledge and expertise on MWEs, even though his research activities are not directed towards NLP (but rather to theoretical syntax).

The large number of projects working on MWEs in a variety of languages in the international computational linguistics community shows that MWEs are generally considered an important problem to be tackled. It is therefore highly desirable that existing research activities in this domain focusing on Dutch are continued and new research activities are initiated so that the position of Dutch in the modern information and communication society can be improved or at least secured.

## 9 Short CV Principal Applicant(s)

Prof. Dr. Jan Odijk, is professor of Language and Speech Technology at UIL-OTS, Utrecht. In the Rosetta project he has contributed significantly to the design and development of the syntactic components, to the approach towards the treatment of idioms in this system, and to the design of the Rosetta lexicons. He has been involved in a variety of projects on NLP lexicons and NLP lexicon standards, (LEXIC, Multilex, EAGLES, and ISLE). He is Vice-President of ELRA and member of a number of ELRA committees (on production and validation of linguistic resources). He was a member of the Spoken Dutch Corpus project steering committee and currently chairs the STEVIN programme committee.

## 10 Literature

### 10.1 Selection of Publications

(Odijk, 2004b,a, 1997, 1998; van Deemter and Odijk, 2000; Heuvel et al., 2003; Villada Moirón, 2004b; Bouma and Villada, 2002; Villada Moirón, 2004a; van der Beek et al., 2002b)

### 10.2 International Literature

(de Schryver, 2003; Jackendoff, 1995; Nunberg et al., 1994; Schenk, 1994; Mel'čuk, 1995; Stantcheva, 2004; Riehemann, 2001; Sag et al., 2001; Kilgarrif and Tugwell, 2001)

## 11 Project budget details

We have assumed a cost of kEURO 60 for a FTE senior researcher and a cost of kEURO 40 for a FTE programmer

<b>FTE</b>	<b>Description</b>	<b>Amount</b>
1 fte for 2 years (24MM)	computational linguist Groningen	kEURO 120
1 fte for 2 years (24MM)	computational linguist Utrecht	kEURO 120
.50 fte for 1 year (6MM)	programmer, Utrecht	kEURO 20
1/12 fte for 2 years (2MM)	lexicographer Van Dale	kEURO 10
.2 fte for two years (4.8MM)	Applicants	kEURO 24
	Costs for validating the MWE lexical database	kEURO 10
2.5 FTE	Forfait	kEURO 53.5
	'Opslag'	k EURO 70.5
	<b>Total</b>	<b>k EURO 416</b>

Groningen matches the costs for coordination and management, so that the total requested funding is **kEURO 410**.

## References

- Anne Abeillé and Yves Schabes. Parsing idioms in lexicalized TAGs. In *Proceedings of the European ACL*, pages 1–9, Manchester, 1989.
- Jaap Akkermans, Brigit van Berkel, Chris Frowein, Linda van Groos, and Dirk Van Compernelle. *Technologieverkenning Nederlandse taal- en spraaktechnologie*. Technical report, M&I/Partners, Amersfoort/Leuven, 23 January 2004. Rapport bij project 103185, versie 01.
- T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows. An Empirical Model of Multiword Expressions Decomposability. In *Proc. of the ACL-2003 Workshop on Multiword Expression s: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan, 2003.
- Gosse Bouma and Geert Kloosterman. Querying dependency treebanks in XML. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, volume V, pages 1686–1691, Las Palmas de Gran Canaria, Spain, 2002.
- Gosse Bouma and Begoña Villada. Corpus-based acquisition of collocational prepositional phrases. In M. Theune, A Nijholt, and H. Hondorp, editors, *Computational Linguistics in the Netherlands 2001. Selected Papers from the Twelfth CLIN Meeting*, pages 23–37, Amsterdam, 2002. Rodopi.
- Lisa Breidt and Frédérique Segond. Idarex: formal description of German and French multi-word expressions with finite-state technology. *MLTT*, 022:1036–1040, November 1995.
- Ted Briscoe and John Carroll. Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th ACL conference on applied Natural Language Processing*, pages 356–363, Washington, D.C., 1997.
- Gilles-Maurice de Schryver. Lexicographer’s dreams in the electronic-dictionary age. *International Journal of Lexicography*, 16(2):143–199, 2003.
- Martin Everaert. *Wijzen van Zeggen*. Nijmegen University, Nijmegen, 18 June 2003. ISBN 90-9017667-5. Inaugural Lecture.
- Martin Everaert, Erik-Jan van der Linden, André Schenk, and Rob Schreuder, editors. *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum Associates, Hilldale, New Jersey/Hove, UK, 1995. ISBN 0-8058-1505-8.
- Alexander Geyken. Bootstrapping a database of German multi-word expressions. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, volume III, pages 911–914, Lisbon, May, 26–28, 2004. ELRA.
- Henk van den Heuvel, Khalid Choukri, Harald Höge, Bente Maegaard, Jan Odijk, and Valérie Mapelli. Quality control of language resources at ELRA. In *Proceedings of Eurospeech 2003*, pages 1541–1544, Geneva, September 2003.
- Ray Jackendoff. The boundaries of the lexicon. In Everaert et al. (1995), chapter 7, pages 133–165. ISBN 0-8058-1505-8.

- Adam Kilgarrif and David Tugwell. Word sketch: Extraction & display of significant collocations for lexicography. In *Proceedings of the 39th ACL & 10th EACL -workshop 'Collocation: Computational Extraction, Analysis and Exploitation'*, pages 32–38, Toulouse, 2001.
- Koenraad Kuiper, Heather McCann, Heidi Quinn, Therese Aitchison, and Kees van der Veer. SAID: A syntactically annotated idiom dataset. Linguistic Data Consortium, LDC2003T10, Pennsylvania, 2003. ISBN 1-58563-268-6. URL <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T10>.
- Diana McCarthy, Bill Keller, and John Carroll. Detecting a Continuum of Compositionality in Phrasal V verbs. In *Proc. of the ACL-2003 Workshop on Multiword Expression s: Analysis, Acquisition and Treatment*, Sapporo, Japan, 2003.
- Igor Mel'čuk. Phrasemes in language and phraseology in linguistics. In Everaert et al. (1995), chapter 8, pages 167–232. ISBN 0-8058-1505-8.
- Rosamund Moon. *Fixed expressions and Idioms in English. A corpus-based approach*. Clarendon Press, Oxford, 1998.
- Geoffrey Nunberg, Ivan A. Sag, and Tom Wasow. Idioms. *Language*, 70:491–538, 1994.
- Jan Odijk. C-selection and s-selection. *Linguistic Inquiry*, 28(2), 1997.
- Jan Odijk. Topicalization of non-extraposed complements in Dutch. *Natural Language and Linguistic Theory*, 16(1):191–222, 1998.
- Jan Odijk. A proposed standard for the lexical representation of idioms. In *Proceedings of 11th EURALEX International Congress*, volume III, pages 153–164, Lorient, France, 2004a.
- Jan Odijk. Reusable lexical representation for idioms. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC) 2004*, volume III, pages 903–906, Lisbon, Portugal, 2004b. European Language Resources Association.
- Susanne Riehemann. *A constructional approach to idioms and word formation*. PhD thesis, Stanford University, 2001.
- M.T. Rosetta. *Compositional Translation*, volume 273 of *Kluwer International Series in Engineering and Computer Science (Natural Language Processing and Machine Translation)*. Kluwer Academic Publishers, Dordrecht, 1994.
- Ivan Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. Multiword expressions: a pain in the neck for NLP. LinGO Working Paper No. 2001-03, 2001.
- Manfred Sailer. *Combinatorial Semantics & Idiomatic Expressions in Head-Driven Phrase Structure Grammar*. PhD thesis, University of Tuebingen, 2000.
- André Schenk. *Idioms and Collocations in Compositional Grammars*. PhD thesis, University of Utrecht, 1994.
- Kristina Spranger. Beyond subcategorization acquisition. Multi-parameter extraction from German text corpora. In *Proceedings of 11th EURALEX International Congress*, volume I, pages 171–177, Lorient, France, 2004.

- Diana Stantcheva. Ermittlung des Komponentbestandes von Idiomen – Versuch eines Modells. In Geoffrey Williams and Sandra Vessier, editors, *EURALEX 2004 Proceedings*, Lorient, July, 6-10, 2004. Université de Bretagne Sud.
- Gregor Thurmair. Industrial requirements for cross-language retrieval. Enabler Workshop, Paris, <http://www.enabler-network.org/final-workshop.htm>, 28 August 2003.
- Gregor Thurmair. Multilingual content processing. Invited Talk at LREC-2004, May 2004.
- Kees van Deemter and Jan Odijk. Formal and computational models of context for natural language generation. In P. Bonzon et al., editor, *Formal Aspects of Context*. Kluwer Academic Publishers, Dordrecht, 2000.
- Leonoor van der Beek, Gosse Bouma, Jan Daciuk, Tanja Gaustad, Robert Malouf, Gertjan van Noord, Robbert Prins, and Begoña Villada. Algorithms for Linguistic Processing NWO PIONEER Progress Report. Available electronically at <http://odur.let.rug.nl/~vannoord/alp>., Groningen, 2002a.
- Leonoor van der Beek, Gosse Bouma, and Gertjan van Noord. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde*, 2002b.
- Begoña Villada Moirón. *Computational aspects of fixed expressions: acquisition and modification potential (working title)*. PhD thesis, University of Groningen, To Appear.
- M. Begoña Villada Moirón. Discarding noise in an automatically acquired lexicon of support verb constructions. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC) 2004*, volume V, pages 1859–1862, Lisbon, Portugal, 2004a. European Language Resources Association.
- M. Begoña Villada Moirón. Distinguishing prepositional complements from fixed arguments. In *Proceedings of 11th EURALEX International Congress*, volume III, pages 935–942, Lorient, France, 2004b.
- Eric Wehrli. Translating idioms. In *Proceedings of COLING-ACL '98*, volume 2, pages 1388–1392, Montreal, Canada, 1998.

Del.	Date	Description	Responsible	Depends on
D1.1	T3	Specification of MaxEntropy and LSA/SVM models	Villada Moirón	
D2.1	T3	Report on formalizing and elaborating Parameterized ECM for Dutch	Linguist Utrecht	
D3.1	T8	Running version of Rosetta3 system	Programmer Utrecht	
D4.1	T12	ECM Incorporation Tools	Programmer Utrecht	D3.1
D1.2	T6	Report on performance of MaxEntropy models on acquisition of support verb constructions	Villada Moirón	D1.1
D1.3	T9	Report on performance of LSA/SVM models on acquisition of phrasal verbs	Villada Moirón	D1.1
D6.1	T15	Report on results of incorporating idiomatic expressions into Rosetta	Linguist Utrecht	D4.1
D1.4	T12	Report on model comparison (MaxEntropy and LSA/SVM) and evaluation on other MWEs types	Villada Moirón	D1.2, D1.3
D2.2	T12	Report on extending the ECM to semi-idioms	Linguist Utrecht	D2.1
D1.5	T18	Specifications of tools to acquire valence patterns & morpho-syntactic restrictions	Villada Moirón	
D2.3	T18	Report on extending the ECM to support verb constructions	Linguist Utrecht	D2.1, D2.2
D1.6	T20	Delivery of automatically acquired data	Villada Moirón	D1.2, D1.3, D1.4, D1.5
D1.7	T21	Report on integration of acquired lexical knowledge into ECM-based database	Villada Moirón	D1.6
D2.4	T22	Initial version of MWE Lexical database, with associated documentation	Linguist Utrecht	D1.6, D2.3
D5.1	T24	Report on Evaluation of integrating ECM Lexical database in Alpino system	Villada Moirón	D2.4
D6.2	T24	Report on Evaluation of integrating ECM lexical database in Rosetta	Linguist Utrecht	D2.4
D7.1	T24	Final version of MWE Lexical database, with associated documentation, validated	Linguist Utrecht	D2.4
D8.1	T4	Delivery of Van Dale MWE database	Van Dale	
D9.1	T24	Report on evaluation by Van Dale	Van Dale	D1.6, D2.4

Table 1: Workplan