The DOT project: Data and Datastructures for a Bank of Governmental Terminology

Isa MAKS, Willy MARTIN, Inger GROESZ,
Amsterdam, The Netherlands

Abstract
In this paper we present some of the features of the model used in the DOT-project. The aim of this pilot project is to find out how to deal with official governmental terminological data in an efficient, consistent and multifunctional way, assuring a maximum of accessibility and user-friendliness.

The project has started on January, 1, 1999 and will be finalised at the end of June 2000. Although the project shows both representational and acquisitional aspects (such as term acquisition), this paper will only focus on aspects of the datamodel such as entities and links.

1 Introduction

DOT (acronym for Databank OverheidsTerminologie = Databank of Governmental Terminology) is a pilot study on governmental terminology, which has been started in January, 1999 and has a duration of 18 months. It has been commissioned by both official organisations (the Nederlandse Taalunie, the Commissie Lexicografische Vertaalvoorzieningen) and industrial organisations (Vlaams Economisch Verbond).

The goal of DOT is to develop and implement a system for the description and comparison of Dutch governmental terms used both in Flanders and the Netherlands. The system should act as a role model for a term bank on Dutch official language and consists of the following parts:

- a data model for the input and description of terms;
- a data bank and user interface for entering and consulting terms, with emphasis on term comparison (intra- and interlingual);
- the description of 200 Dutch-Flanders and Dutch-Netherlands terms to be used as an example for the future user;
- a term extractor.

The end-users of DOT will be governmental employees of both Flanders and the Netherlands, translators, terminologists, lawyers specialised in social security, etc. The project is carried out by a consortium consisting of:

- Free University of Amsterdam, Department of Lexicology (I. Groesz, I. Maks and W. Martin);
- University of Stuttgart, Institut für maschinelle Sprachverarbeitung (U. Heid and M. Freese);

- Zeno Software, Antwerp (H. Suykerbuyk and G. François).

The Free University of Amsterdam coordinates the project, with Prof. Willy Martin acting as the project leader.

In this project a broad definition of governmental terminology is used: all terms that are found in governmental texts are considered to be governmental terminology. However, as the government is considered to be a legal organisation, i.e. one that is by law entrusted with the authority over citizens and/or other public institutions, all governmental terms should have a judicial status in order to be considered as such. Next to that, the items in question should be terms, meaning that they are linguistic expressions of concepts which are typically used within a particular knowledge domain and by particular members of the linguistic community (domain experts). In this project the sample data have been restricted to the three following subdomains: health and disability insurance, taxes, and child allowance.

In the rest of the paper the data model will be presented, paying attention to the use of entities and links between them

2 Data and Data modeling

2.1 Different perspectives

The data model is meant for the following tasks:

- description and editing of terms
- comparison of concepts from different (legal/governmental) systems
- comparison of terms from different languages/ language areas
- translation of terms

It consists of entities describing concepts, terms and collocations, and of links relating the entities to each other.

2.2 Entities

2.2.1 Core concepts

The core concept entity corresponds to a semantic unit expressed by one or more terms in one or more languages. It describes that part (and only that part) of the conceptual meaning that the terms which are linked to it, have in common. As a rule this part of the conceptual meaning is the most essential one. A (core) concept entity may correspond to more than one term entity.

2.2.2 Terms

The term entity represents one term and its description. The description consist of a linguistic description, like spelling, language, word category, inflection data, and pragmatic values describing the usage of the term, a full text definition and a frame based definition.

2.3 Illustration of core concept and term entities

The following example gives the core concept entity which is shared by the Dutch-Netherlands term *vakantietoeslag* (i.e. *holiday allowance*) and the Dutch-Flanders term *vakantiegeld*. Both the Dutch and the Belgian governmental systems grant such an allowance and the most essential characteristic of the meaning, namely is *what the allowance is meant for*, is the same in the social security systems of both countries.

Core concept entity *vakantietoeslag (holiday allowance)*

- CORE CONCEPT

DOMAIN	government
SUBDOMAIN	social security
LEGAL SYSTEM	The Netherlands, Belgium
CONCEPT TYPE	allowance

- FRAME-BASED DEFINITIONS (CORE PART)

MOTIVE	extra allowance for the holidays

The term entity presents the differences between the Dutch-Netherlands and the Dutch-Flanders terms. Only the description of *vakantietoeslag* is given.

Term (and specific concept) entity: *vakantietoeslag*

- TERM ENTITY: GENERAL INFORMATION

SPELLING	VAKANTIETOESLAG
LANGUAGE	Dutch
WORDSTRUCTURE	simplex

- INFLECTION

WORDCATEGORY	noun
PLURAL FORM	vakantietoeslagen
ARTICLE	de
GENDER	masculine

- PRAGMATICS

LANGUAGE AREA	The Netherlands
CHRONOLOGY	synchronous
STYLE	neutral
COMM SITUATION	internal
TEXT TYPE	law questions& answer in parliament jurisprudence

- FRAME-BASED DEFINITION (SPECIFIC PART)

GIVER	employer
BENEFICIARY	employee
AMOUNT OF PAYMENT	at least 8% of gross income
TIME OF PAYMENT	not later than the end of June
ON THE BASIS OF	Law …..

- TEXTUAL DEFINITION

aanspraak van de werknemer jegens zijn werkgever op ten minste 8% van zijn ten laste van de werkgever komende loon, alsmede van de uitkeringen waarop hij tijdens de dienstbetrekking aanspraak heeft; bedraagt de som van dit loon en deze uitkeringen méér dan driemaal het minimumloon, dan mag het meerdere daarbij buiten de berekening blijven; de vakantietoeslag moet betaald worden voorafgaande aan de vakantie van de werknemer, maar niet later dan in juni	
SOURCE	Juridisch Woordenboek (1997)

2.3.1 Collocations

The collocation entity represents a collocation, its linguistic description and its usage. The collocation entity also contains examples which are illustrative for the terms the entity is linked to (in that case the value for the feature *collocation type* is *free).*

2.4 Links

The entities are connected with each other by means of links. There are explicit, implicit and overruling links. Explicit links are defined by the user of the system. Implicit links are based upon existing explicit links and are derived by the system. Overruling links are user-defined and overrule implicit links.

2.4.1 Explicit links

Explicit links are links which are defined by the user of the system, when describing new terms. They exist between:

- a concept entity and a term entity,
  to indicate that the meaning of the term corresponds to the given concept;
- a concept entity and another concept entity,
  to indicate hypernymy, hyponymy, related synonymy, near synonymy, etc. ;
- a term entity and a collocation entity,
  to indicate to which term a collocation belongs;
- a collocation entity and another collocation entity,
  to indicate translation equivalence.

2.4.2 Implicit links

The implicit links are the links which represent the unmarked and most frequent relations between terms. They are based upon existing explicit links and will be derived by the system. They exist between:

- a term entity and another term entity (from the same or from different languages),
  The system derives the links according to the following principles:
  - if terms are linked to the same concept and if they belong to the same language, full synonymy or context restricted synonymy is implied;
  - if terms are linked to the same concept and if they belong to different languages, complete translation equivalence is implied;
  - if terms are linked to concepts of different conceptual systems and if they are near synonyms, near translation equivalence is implied;
  - if terms are linked to the same concepts but have different pragmatic values, restricted translation equivalence is implied.

2.4.3 Overruling links

Overruling links are defined by the user. They overrule the implicit links in those cases where the deriving principles fail.

They exist (like the implicit links) between a term entity and another term entity.

2.4.4 Illustration of the data model

The example in figure 1 gives a schematic representation of the relations between four terms:

- *vakantiegeld* (Dutch-Flanders),
- *vakantietoeslag* (Dutch-Netherlands),
- *vakantiegeld* (Dutch-Netherlands),
- *pécule de vacances* (French-Belgium).

All terms have a common core concept, namely the one already described in the example above: "extra allowance meant for the holidays". On the basis of the explicit links to the same core concept, the system derives translation links between the terms of the different languages. The term *vakantiegeld (Dutch-Netherlands*) is restricted in its usage, since it is not used in law texts and, consequently, the derived implicit links between that term and the others indicate "Restricted Translation Equivalence". In this example the overruling links don't apply.

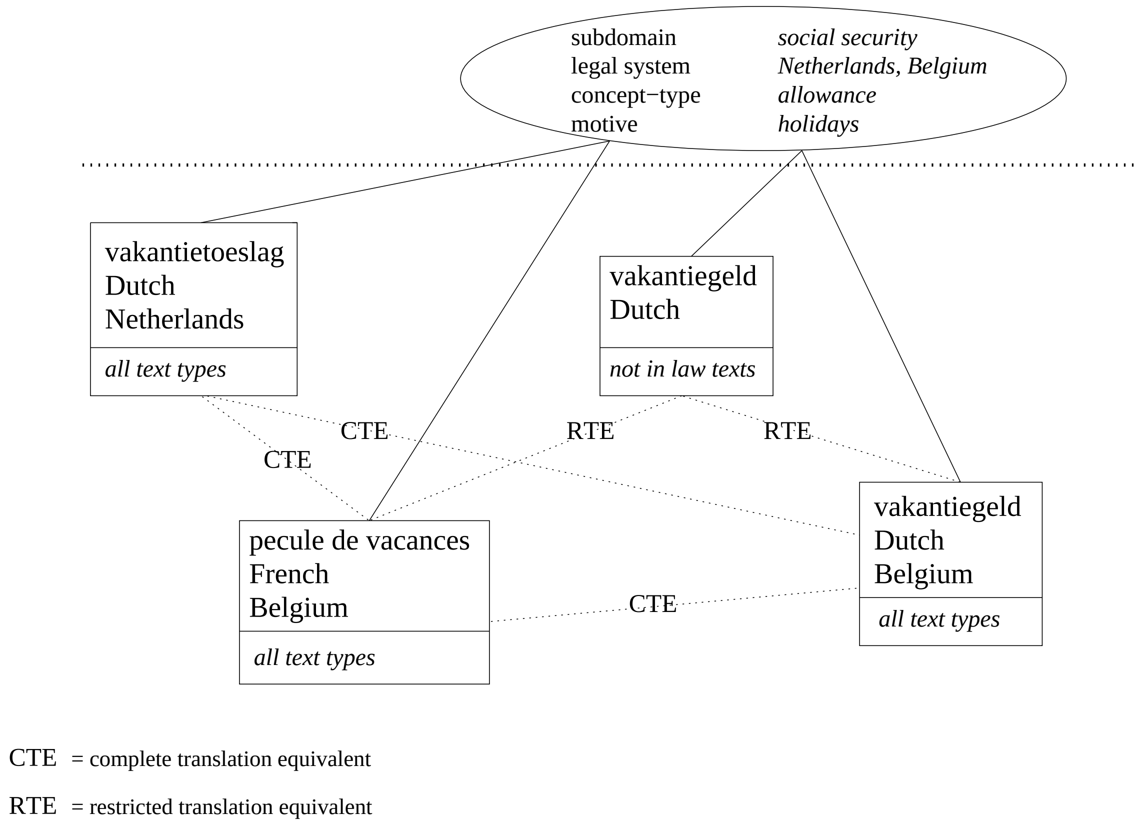

Figure 1: Relations between terms

3 Discussion

The data model represents the different aspects of the description of a term in a systematic and explicit way, which helps to indicate the relevant relations between terms. This ensures that the model can fulfil its task (as mentioned in 2.1): to describe, compare and translate terms.

The following points are crucial:

- Two types of definitions:
  The contextual definition is the traditionial type of definition in terminology and is easy to be read and interpreted by a human user.

  The frame-based definition is a systematic representation of the meaning of the term. The frames provide a view on the data by which terms and concepts can be related between each other. In many cases the fillers are other terms of the domain which are described within the system. In this way a network of terms and concepts will be formed, giving access to all kinds of relevant information. By describing groups of terms using the same set of slots, differences and correspondencies between the fillers are easily noted. Thus the frames are a powerful tool to carry out a comparison between terms. Since the frames are language independent the comparison may be both interlingual and intralingual.

- The notions *core concept* and *specific concept*:
  The concepts of different law systems are never completely equivalent. However, the function of a concept within one law system may be equal to the function of a concept within another law system. In other words, the answer to *what does it regulate* may be the same for the different law systems. The core concept entity represents this part of the concept, its function within a law system, thus expressing that part of the meaning which the terms of different law systems have in common. The answer to *how does it regulate* is never the same. It refers to different rules, laws, institutions, etc. The specific concept entity represents those details of the meaning and contains the information about what the terms do not have in common . By distinguishing the two parts of the meaning, the data model helps to stress both similarities and dissimililarities between the terms.

- Automatically derived links:
  By the use of derived links (see 2.3.2) , the description of the terms can be efficient and complete. The more terms are linked to a concept, the more links will be derived automatically.

References

Algra, N.E. & H. R. W. Gokkel (1997). *Verwijzend en verklarend woordenboek,* Alphen aan de Rijn: Samson.

Groesz, I., Martin, W., Ten Pas, E., Stemmerik, Y. (1996). *Onderwijsterminologie,* Vrije Universiteit, Amsterdam

Mayer, F. (1998). *Eintragsmodelle für terminologische Datenbanken*, Tübingen: Gunter Narr.

Minsky, M. (1977). Frame-system theory. In: *Thinking*, P.N. Johnson-Laird and P.C. Watson

(eds.) pp. 355-376, Cambridge.

Redel, J.V., Stemmerik, Y.M. (1995). *Overheidsterminologie,* Vrije Universiteit, Amsterdam

Martin, W. (1998). Frames as definition models for terms. In: *Proceedings of the 4th Infoterm Symposium. Vol. II*, A. Munteanu (ed.), pp. 189-221, Vienna.

Sandrini, P. (1996). Terminologiearbeit im Recht. *Wien:Termnet.*