Scientific Report Short Term Scientific Mission IS1305-20655

COST STSM Reference Number:	COST-STSM-IS1305-20655
Period:	2014-08-18 00:00:00 to 2014-08-22 00:00:00
COST Action:	IS1305
STSM type:	Regular (from Netherlands to Belgium)
STSM Applicant:	Dr Carole Tiberius Instituut voor Nederlandse Lexicologie, Leiden (NL) carole.tiberius@inl.nl
STSM Topic:	A database of linguistic terms for Dutch
Host:	Prof. Dr Frieda Steurs University Leuven (campus Antwerp), Antwerp (BE) frieda.steurs@arts.kuleuven.be

1 Purpose of the STSM

The goal of this STSM was to further develop a database of linguistic terms for Dutch (to be used within the context of one of the projects at INL). The original goal was to set up a database of linguistic terms for Dutch in accordance with the international debate on the standardisation of linguistic terminology (ISOcat) in the context of CLARIN-EU. To this end, the GOLD data set (ISOcat version) has been used as a starting point for the Dutch database of linguistic terms. Terms that are not used in Dutch have been removed from the list and terms from other lists (available within the project i.e. glossary of terms from the Syntax of Dutch and Grammar of Words etc.) have been added together with their definitions. The result is a relative mishmash of terms and a diversion from the ISOcat standard.

INL lacks expertise on terminology which is available at the University Leuven (campus Antwerp) with Frieda Steurs.

During this STSM, Carole Tiberius will receive an in-depth introduction to term management, concept modelling and standardisation of terminology and a plan will be developed to rectify the problems with the current version of the Dutch database of linguistic terms.

2 Description of the work carried out during the STSM

During the STSM, Carole Tiberius has received an in-depth introduction to term management, concept modelling and standardisation of terminology. The day-to-day planning was as follows:

Day 1: welcome; introductory meeting with Frieda Steurs and Hendrik Kockaert discussing the schedule of the week; background reading (preprint of the Handbook of Terminology which will be published in October 2014); meeting with Hendrik Kockaert about ISO standardisation norms.

Day 2: background reading (Handbook of Terminology); meeting with Ken De Wachter about software for termbases as well as their integration in translation software (Multiterm, Trados).

Day 3: background reading (Handbook of Terminology); meeting with research assistants Leen Boel and Veerle Verschakelen on the Qualetra-project (http://www.eulita.eu/fr/qualetra).

Day 4: background reading (Handbook of Terminology); meeting with Kris Heylen preparing for the COST WG3 meeting of August 2015.

Day 5: background reading (Handbook of Terminology); writing a draft of the scientific report.

3 Description of the main results obtained

The result of the STSM is a set of recommendations for improving the Dutch database of linguistic terms. First some background information is given. Then the current state of the termbase is discussed and finally a number of recommendations are given for improving the termbase and turning it into a sound terminological project.

3.1 Background

The database of linguistic terms which we are considering here, is part of the Taalportaal project. Taalportaal is a large-scale NWO-funded project (2011-2015) carried out by a consortium of four partners: the Meertens Instituut, the Fryske Akademy, The University of Leiden and the Institute of Dutch Lexicology (INL). The latter is responsible for the technical infrastructure.

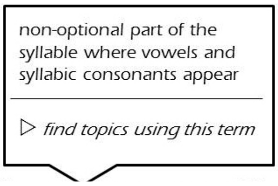
The aim of the project is to create an online portal offering access to a comprehensive grammar of Dutch and Frisian, covering morphology, phonology and syntax of both languages. Besides the grammar module, the portal will contain an **ontology of linguistic terms** and an extensive bibliography. In order to serve the international scientific community, all content is in English.

The final version of the portal will be delivered at the end of 2015. As of February 2014, a beta version of the portal is available online at www.taalportaal.org.

3.2 Taalportaal database of linguistic terms

The Taalportaal database of linguistic terms will serve the functionality of the portal for three reasons.

- First, it will provide users with a list of terms and their definitions. This list will be available, and searchable, on a separate page of the portal. Also, the definitions will be linked to occurrences of the terms in the actual topics. Thus, users will be able to get instant information on a linguistic term (by means of a popup, see figure 1).



non-optional part of the syllable where vowels and syllabic consonants appear

▷ *find topics using this term*

Figure 1 Example of a popup appearing when a user hovers over a linguistic term in a topic text (Landsbergen et al. 2014)

- Second, the list will serve as a register, providing links from each term to a series of topics.
- Third, the hierarchical structure of the database (through relationships between terms) will enable query expansion of user searches, thus enriching retrieval options. For example, a search for the term *article* can be extended to occurrences of the word *determiner* or even the actual Dutch determiners *de* and *het* (table 1).

query	expanded query
article	article, determiner, *de*, *het*, *een, it*
agreement	agreement, concord
lexical morpheme	lexical morpheme, free morpheme

Table 1: Examples of query expansion through the use of the ontology of linguistic terms (Landsbergen et al. 2014)

The Taalportaal termbase is not being constructed from scratch, but is based on existing resources, i.e. GOLD ontology (2010), glossary of Geert Booij - *Grammar of Words* - (2010) and glossary of Hans Broekhuis – *Syntax of Dutch* - (2013). Initially we assumed that the GOLD ontology would form the default, but experience has learned that the material from Geert Booij and Hans Broekhuis (also Taalportaal authors) is preferred within the context of the project.

We are thus dealing with a terminological project in the context of a larger project. Considering the classification of terminological projects discussed by Dobrina (2014: 182), the Taalportaal termbase has the following characteristics.

Needs to be met: A domain project where existing resources fail to meet the needs of the target group

Objective: Enhance the quality of terminological information in an existing resource

Target users: Domain experts

Properties of the resource: (1) scope: terminology belonging to the scope of the Taalportaal project
(2) types of terminological information presented:

- Term information: terms, equivalents, specification of a term status
- Concept information: definitions, references to related concepts

(3) directionality: first phase is source language oriented. In a second phase possibly also target language oriented
(4) prescription level: descriptive

Summarising, the Taalportaal termbase is a descriptive project. The most important goal is to create high-quality terminology to be used within the context of the Taalportaal project.

From the point of view of standardisation, it is important that the Taalportaal termbase takes the ISO rules (ISO/FDIS 704, 2009) for definition writing into account and ensures compatibility with data exchange formats, such as TBX, ISOcat, to allow for a wider use of the termbase in the future. We will discuss these issues further below. First we will consider the format of the definitions in section 3.2.1, then we will briefly look at related lists of linguistic terms (section 3.2.2) and finally we will discuss the structure of the termbase and its compatibility with existing data exchange formats (section 3.2.3).

3.2.1 Definition writing

In terminology work, the following types of definitions are recognised (ISO/FDIS 704, 2009: 19):

- intensional definitions
- extensional definitions
- ostensive definitions

For the Taalportaal termbase, **intensional definitions** are preferred. The role of an intensional definition is to provide the minimum amount of information that forms the basis for abstraction and that allows one to recognise and differentiate the concept from other related concepts, especially coordinate concepts. Unlike an encyclopaedic description or an explanation, a definition's main purpose is not to provide a means for a complete understanding of a given concept but rather to provide enough understanding as to avoid confusing the concept in question with other related concepts. One of the requirements of Taalportaal is that the definitions can be shown in a popup on the screen and thus they need to be concise.

The **format** in which the definition is written in published dictionaries varies from language to language. Every language has its own conventions and definitions should respect them.

> EXAMPLE
> According to convention in English, the dictionary definition is a statement that does not form a complete sentence. It must be combined with the entry term in order to be read as a sentence: the subject is the designation, the copula is understood to be the verb “is“ and the predicate constitutes the definition. The dictionary entry:
>
> **lead pencil**
> a pencil whose graphite core is fixed in a wooden casing that is removed for usage by sharpening
>
> is to be read as: “[A] lead pencil [is] a pencil whose graphite core is fixed in a wooden casing that is removed for usage by sharpening”.

(ISO/FDIS 704, 2009: 19)

An overview of the rules for writing and assessing intensional definitions is given in Löckinger, Kockaert & Budin (2014). They are:

1. Preciseness: it must contain all the delimiting characteristics necessary to describe the concept in question unambiguously
2. Conciseness: it should ideally consist of a single sentence, including subclauses
3. Reference to the immediate superordinate concept
4. Use of terms designating known or defined concepts
5. Objectivity: a definition should not push a particular point of view or cause
6. Source reliability
7. Suitability for the relevant target group
8. Indication of the scope of application
9. Reference to the relevant domain: it must contain those characteristics that mirror the perspective of a given domain
10. Reference to a concept system: it must express the relations of the concept to be defined with other concepts of the given concept system
11. Linguistic correctness

12. Absence of circularity/tautology (i.e. the definiendum should not be repeated in the definition; when the designation/definiendum is a compound term, the nucleus of that compound or complex term may, and usually does, introduce the definition, when the nucleus is the generic term
13. Affirmativeness (avoidance of negative definitions): it must describe what it is, not what it is not
14. Avoidance of translated intensional definitions
15. Avoidance of hidden definitions of other concepts
16. Absence of characteristics of superordinate or subordinate concepts

These guidelines form the starting point for checking and correcting the definitions that are currently in the Taalportaal termbase. If the definitions are well-formed (i.e. they all start with a reference to the immediate superordinate concept), they can be used to automatically infer hierarchical relations between the concepts, which is important for the successful development of the online portal (i.e. requirement 3 for the termbase).

Recommendation: All definitions in the database need to be checked systematically to make them consistent and conform the ISO rules.

Table 2 illustrates a set of complex concepts all related to case together with their definitions in the current version of the termbase marking some of the shortcomings.

Term	Taalportaal	GOLD
abessive_case	case that expresses[1] ‘absence of, distance from’	AbessiveCase expresses the lack or absence of the referent of the noun it marks. It has the meaning of the English preposition 'without' [Pei and Gaynor 1954: 3, 35].
ablative_case	case that expresses ‘movement from’	Ablative case denotes the source, agent, means, and occasionally also time or place of an an act or occurence. [Pei and Gaynor 1954: 3]
absolutive_case	case marking of the single argument of a verb with one argument, and of the object argument for verbs with more than one argument in an absolutive-ergative system	AbsolutiveCase in ergative-absolutive languages marks referents that would generally be the subjects of intransitive verbs or the objects of transitive verbs in the translational equivalents of nominative-accusative languages [Anderson 1985: 181; Crystal 1985: 1; Andrews 1985: 138].
accusative_case	in a nominative-accusative system,[2] case that marks direct objects of verbs	AccusativeCase in nominative-accusative languages marks certain syntactic functions, usually direct objects [Hartmann and Stork 1972: 3, 156; Crystal 1980: 11, 246; Andrews 1985: 75; Anderson 1985: 181].
adessive_case	case that expresses ‘position at’	AdessiveCase expresses that the referent of the noun it marks is the location near/at which another referent exists. It has the meaning of 'at' or 'near' [Crystal 1997: 8].
allative_case	case that expresses ‘movement towards’	AllativeCase expresses motion to or toward the referent of the noun it marks [Pei and Gaynor 1954: 6, 9, 216; Lyons 1968: 299; Crystal 1985: 1213].
benefactive_case	case that expresses that the referent of the noun it marks receives the benefit of the situation expressed by the clause	BenefactiveCase expresses that the referent of the noun it marks receives the benefit of the situation expressed by the clause [Crystal 1980: 43].
case	the marking of words. that encodes their relationship to other elements in the sentence, in particular verbs, prepositions, and other nouns	CaseProperty is the class of properties that concerns the grammatical encoding of a noun's relationship (syntactic or semantic) to some other element in the sentence, such as a verb, noun, pronoun, or adposition [Pei and Gaynor 1954: 35; Crystal 1980: 53-54; Anderson 1985: 179-180; Andrews 1985: 7172; Kuno 1973: 45; Blake 2001].
comitative_case	case that expresses ‘accompaniment’	ComitativeCase expresses accompaniment. It carries the meaning 'with' or 'accompanied by' [Anderson 1985: 186; Pei and Gaynor 1954: 42; Dixon, R. 1972: 12].
dative_case	case that marks the indirect object of a sentence	DativeCase marks 1) Indirect objects (for languages in which they are held to exist) or 2) nouns having the role of recipient (as of things given), beneficiary of an action, or possessor of an item [Crystal 1980: 102].
direct_case	a synonym of ‘structural case’[3]	
ergative_case	case marking of the subject argument for verbs with more than one argument in an absolutive-ergative system	ErgativeCase in ergative-absolutive languages generally identifies the subject of transitive verbs in the translation equivalents of nominative-accusative Languages such as English [Crystal 1980: 134; Hartmann and Stork 1972: 78; Pei and Gaynor 1954: 67; Andrews 1985: 138].
genitive_case	case that marks the dependency of a noun or noun phrase on another noun	GenitiveCase is used to mark the noun whose referent is the possessor of the referent of another noun [Crystal 1980: 161; Hartmann and Stork 1972: 94-95, 180; Pei and Gaynor 1954: 82, 172; Anderson 1985: 185; Fleming 1988: 10].
inherent_case	a synonym of ‘semantic case’	

[1] Ideally, the same phrase is used in all definitions defining similar concepts, i.e. all types of cases are defined as “case that expresses…” or “case that marks …” but not both.

[2] Ideally the generic concept occurs at the beginning of the definition, allowing automatic construction of a concept hierarchy.

[3] If a term is referring to the same concept as another term, the definition of that concept should be given on both occasions and not only for the preferred term. There is one concept which can be designated by more than one term.

instrumental_case	case that indicates that the referent of the noun it marks is the means of the accomplishment of the action expressed by the clause	InstrumentalCase indicates that the referent of the noun it marks is the means of the accomplishment of the action expressed by the clause [Crystal 1980: 187; Hartmann and Stork 1972: 114].
local_case	case that expresses a spatial notion	
nominative_case	case that marks the subject in a nominative-accusative system	NominativeCase identifies clause subjects in nominative-accusative languages. It is usually the unmarked case. Nouns used in isolation often have this case. [Crystal 1980: 242; Pei and Gaynor 1954: 147; Hartmann and Stork 1972: 224]
oblique_case	non-nominative structural case	In a direct/oblique system or in a nominative/oblique system, oblique case is the term for all roles not marked by the direct case or nominative case. In the phrase 'the oblique cases' it is used to refer to a set of cases excluding the nominative (occasionally the nominative or accusative). [Bauer 2004: 27]
partitive_case	case that expresses the notion ‘part of ‘	PartitiveCase expresses the partial nature of the referent of the noun it marks, as opposed to expressing the whole unit or class of which the referent is a part. This case may be found in items such as the following: existential clauses, nouns that are accompanied by numerals or units of measure, or predications of material from which something is made. It often has a meaning similar to the English word 'some'. [Pei and Gaynor 1954: 161; Richards, Platt and Weber 1985: 208; Quirk et al. 1985: 249; Sebeok 1946: 1214]
semantic_case	case used for the expression of a semantic notion	
structural_case	case required by a syntactic context	

Table 2 Set of complex concepts which have case as their generic concept together with their definitions in the current version of the Taalportaal termbase.

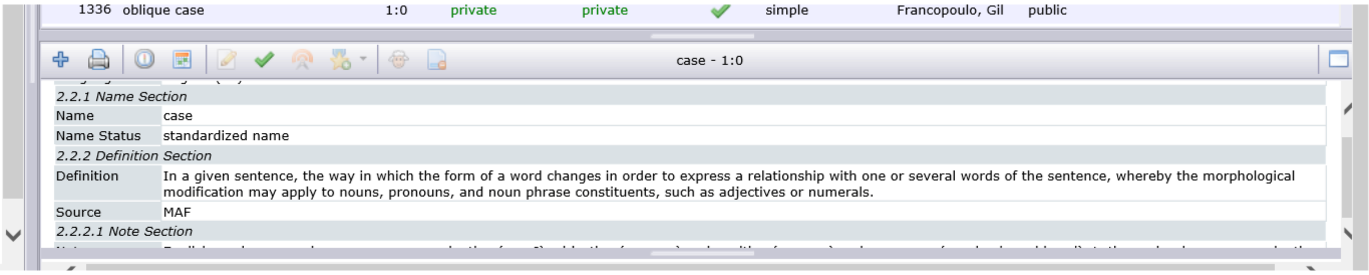
3.2.2 Taalportaal termbase: multiple terminology lists

As stated above, the Taalportaal termbase has not been constructed from scratch, but uses input from different existing resources, mainly the GOLD ontology, the glossary from Geert Booij (2010) and the glossary from Hans Broekhuis (2013). We have also looked at lists that are available in ISOcat[4].
When dealing with multiple lists, it is possible that the same term occurs in more than one list, but is defined differently, as is illustrated below for the concept 'case'.

case:
the marking of words, that encodes their relationship to other elements in the sentence, in particular verbs, prepositions, and other nouns. (Booij 2010)

CaseProperty is the class of properties that concerns the grammatical encoding of a noun's relationship (syntactic or semantic) to some other element in the sentence, such as a verb, noun, pronoun, or adposition [Pei and Gaynor 1954: 35; Crystal 1980: 53-54; Anderson 1985: 179-180; Andrews 1985: 7172; Kuno 1973: 45; Blake 2001]. (GOLD 2010)

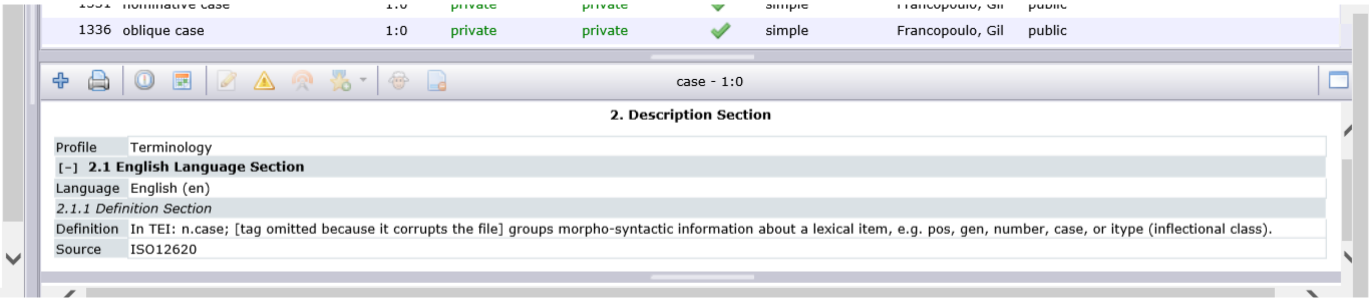
A search for 'case' in ISOcat results in 3 hits, shown in the screenshots below.



1336	oblique case	1:0	private	private	✓	simple	Francopoulo, Gil	public

case - 1:0

2.2.1 Name Section
Name: case
Name Status: standardized name
2.2.2 Definition Section
Definition: In a given sentence, the way in which the form of a word changes in order to express a relationship with one or several words of the sentence, whereby the morphological modification may apply to nouns, pronouns, and noun phrase constituents, such as adjectives or numerals.
Source: MAF
2.2.2.1 Note Section

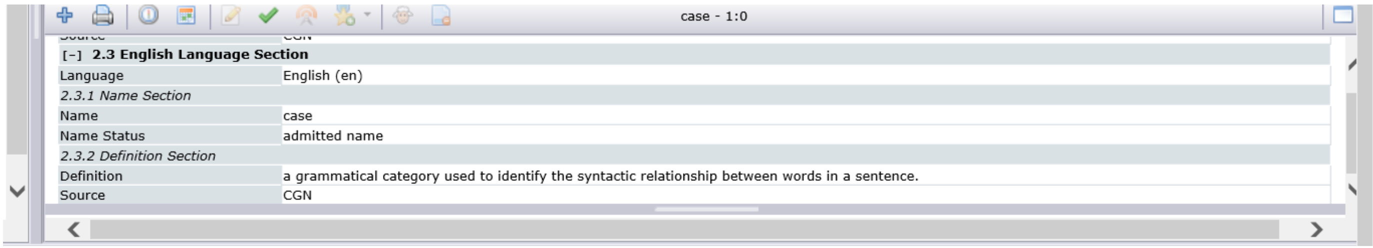


1336	oblique case	1:0	private	private	✓	simple	Francopoulo, Gil	public

case - 1:0

2. Description Section

Profile: Terminology
[-] **2.1 English Language Section**
Language: English (en)
2.1.1 Definition Section
Definition: In TEI: n.case; [tag omitted because it corrupts the file] groups morpho-syntactic information about a lexical item, e.g. pos, gen, number, case, or itype (inflectional class).
Source: ISO12620



case - 1:0

[-] **2.3 English Language Section**
Language: English (en)
2.3.1 Name Section
Name: case
Name Status: admitted name
2.3.2 Definition Section
Definition: a grammatical category used to identify the syntactic relationship between words in a sentence.
Source: CGN

[4] http://www.isocat.org/

Recommendation: As the Taalportaal termbase is a descriptive termbase, we use the terms and definitions that are relevant within the context of the project. At a later stage and time permitting, we will investigate whether the Taalportaal termbase can be merged with other lists. When the same term occurs in more than one lists with different definitions, we will have to decide whether the all refer to the same concept, or whether we are dealing with different concepts for which the same term is used. For instance, the CGN dataset in ISOcat adopts a more functional approach, whereas Taalportaal and GOLD use the formal marking as a criterion in defining concepts. This will result in multiple concepts.

3.3.3 Format and structure of the Taalportaal termbase and its compatibility

In this section, we give an overview of the structure of the Taalportaal termbase in relation to data models in Terminology as described in the chapter by Nistrup Madsen & Erdmann Thomsen (2014) in the Handbook of Terminology.

The current set-up of the Taalportaal termbase is as follows:

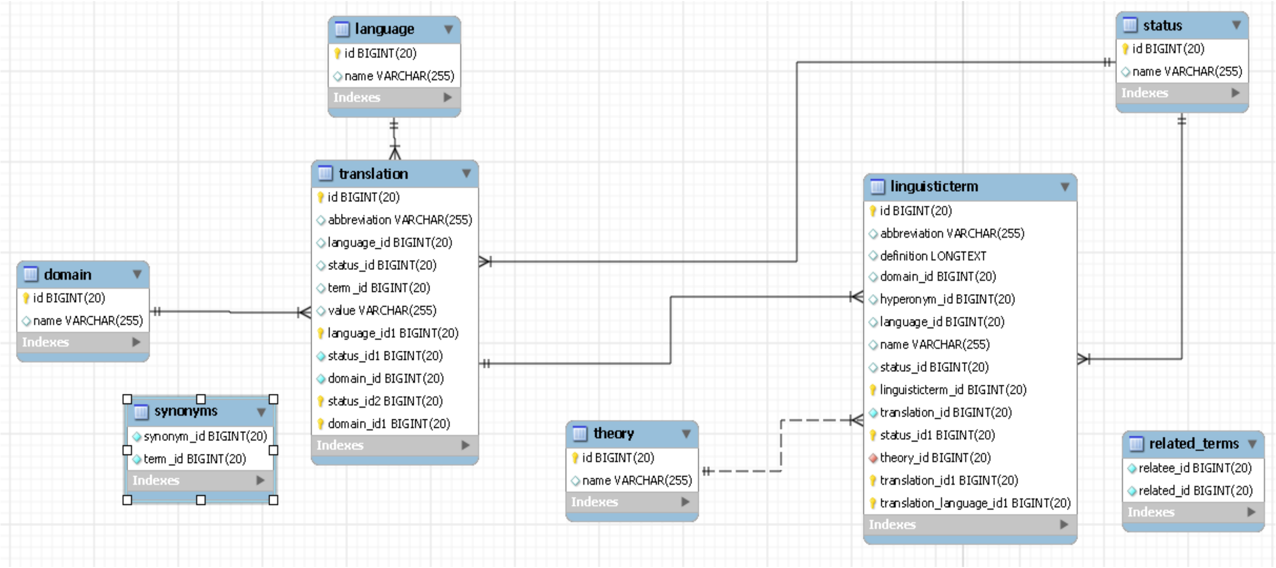


Figure 2 Database structure of the Taalportaal termbase

Each record in the database has a label for the concept (name) and is marked for being a preferred term or not (status). Each concept also has a definition and a source of the definition. This is obligatory information.

Furthermore, the domain (syntax, morphology, phonology) of the concept is specified, its status (checked by Taaportaal expert), its relevance for Taalportaal, and information on its place in the hierarchy (hyperonym / subClass of). If applicable an abbreviation can be specified.[5] Translations can be added and it is possible to specify if a term belongs to a particular theory.

The source of the concept is also given, together with a dcr-id taken from the GOLD dataset (OWL-ISOcat version 2010).

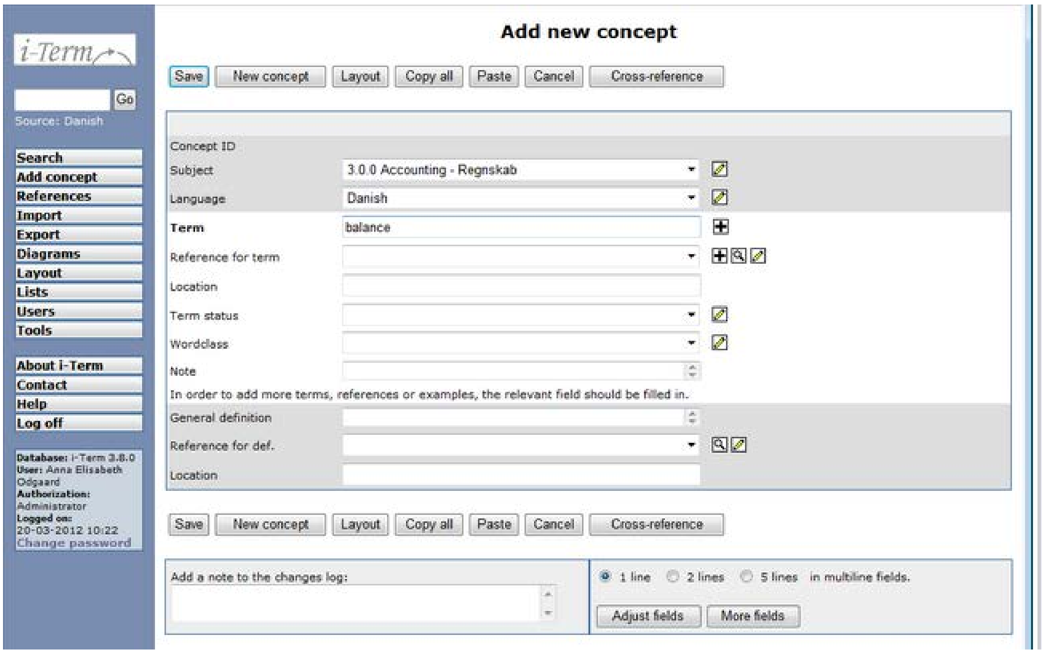
Furthermore there is information on whether the term for this concept also the occurred in the glossaries of Geert Booij and Hans Broekhuis.

Comparing this to, for instance, the i-term[6] default structure (see the add new concept window in figure 3), suggests that the Taalportaal termbase has a rather rich data structure. At the moment, however, only a minimal amount of information is specified in the database, i.e. term, definition and source of definition. The data in the Taalportaal database can be exported as OWL[7].

[5] It would probably be better to use a label for distinguishing between full forms and abbreviations.

[6] http://www.iterm.dk/

[7] The W3C Web Ontology Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things.



Source: Danish

Database: i-Term 3.8.0
User: Anna Elisabeth Odgaard
Authorization: Administrator
Logged on: 20-03-2012 10:22
Change password

Add new concept

Concept ID
Subject: 3.0.0 Accounting - Regnskab
Language: Danish
Term: balance
Reference for term
Location
Term status
Wordclass
Note
In order to add more terms, references or examples, the relevant field should be filled in.
General definition
Reference for def.
Location

Add a note to the changes log:

● 1 line ○ 2 lines ○ 5 lines in multiline fields.

Figure 3 Input screen of i-term

Recommendation: Ensure that the minimal information that is required is completed consistently and conform available standards. Thus, an entry in the database should contain at least the following information:

```
<termEntry id=XXX>
        <definition></definition>
        <definitionSource></definitionSource>
        <languageSection xml:lang=’en’>
                <termSection>
                        <term></term>
                        <termStatus></termStatus>
                        <location></location>
                        </termSection>
                <termSection>
                        …
                </termSection>*
        </languageSection>
```

The tag ‘termStatus’ indicates whether the term is a preferred term or not. The tag ‘location’ specifies where the term is used. In this model, abbreviations are listed as a term in a separate term section (maybe we need a tag ‘termType’ to indicate that a term is either a full form, multiword expression or an abbreviation)
The advantage of this set up is that we do not need a tag ‘relatedTerms’ anymore.
At the moment we only deal with English, but if we wanted to add Dutch and Frisian equivalents this could be done in the same way as listing variant terms in a term section in a term entry, but by adding a new language section.
The proposed structure is very much inline with the TBX format[8] (Melby 2014).

[8] TBX (TermBase eXchange) is a system for the exchange of terminological data. It is an ISO standard (number 30042) and an industry standard (formerly from LISA; now from ETSI). It includes a family of XML markup languages (called TMLs, for Terminological Markup Languages; also called TBX dialects).

References:

Booij, Geert (2010). *Construction morphology.* Oxford University Press.
Broekhuis, Hans (2013). *Syntax of Dutch. Adjectives and adjective phrases.* Amsterdam University Press.
Conference of Translation Services of European States (COTSOES) (2002). *Recommendations for Terminology Work.* Berne: Federal Chancellery.
Dobrina, Claudia (2014). 'Getting to the core of a terminology project'. In: Hendrik J. Kockaert and Frieda Steurs (eds). *Handbook of Terminology*. Volume 1. John Benjamins. 180-199.
Kockaert, Hendrik, J. and Frieda Steurs (eds) (2014). *Handbook of Terminology.* Volume 1. John Benjamins.
Landsbergen, Frank, Carole Tiberius and Roderik Dernison (2014). 'Taalportaal: an Online Grammar of Dutch and Frisian'. In: Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds). Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland.
Melby, Alan K. (2014). 'TBX: A terminology exchange format for the translation and localisation industry'. In: Hendrik J. Kockaert and Frieda Steurs (eds). *Handbook of Terminology*. Volume 1. John Benjamins. 393-424.
Nistrup Madsen, Bodil and Hanne Erdman Thomsen. (2014). 'Concept modelling vs. data modelling in practice'. In: Hendrik J. Kockaert and Frieda Steurs (eds). *Handbook of Terminology*. Volume 1. John Benjamins. 250-275.

Useful resources:
GOLD (2010) http://linguistics-ontology.org/.
http://tnc.se/the-swedish-centre-for-terminology.html
http://www.iso.org/iso/iso_technical_committee.html%3Fcommid%3D48104
https://www.iso.org/obp/ui/#home
http://www.granddictionnaire.com/Resultat.aspx

4 Future collaboration with the host institution (if applicable)

INL and KU Leuven Campus Antwerp will continue to collaborate in the future exchanging knowledge on lexicography and terminology.

5 Foreseen publications resulting from the STSM (if applicable)

The work on the Taalportaal termbase may result in a publication, co-authored with Menzo Windhouwer, on merging the Taalportaal termbase with other lists of linguistic terms, e.g. GOLD.

6 Other comments (if any)

N/A