

BISTRO: the online platform for terminology management. Structuring terminology without entry structures

Oliver Streiter, Natascia Ralli, Isabella Ties & Leonhard Voltmer
EURAC, European Academy of Bolzano

BISTRO is an online platform which supports the translation process in various phases. The phases which can be distinguished are the terminological preparation of the source text, the creation of terminological glossaries and the retrieval of related documents and their terminological elaboration. For this purpose BISTRO hyperlinks a terminology database with bilingual and trilingual corpora. Term tools such as term extraction (TE), term recognition (TR) and keyword-in-context (KWIC) may be applied to the query results, which consist of retrieved terms or corpus segments. BISTRO's architecture is open for new tools and contents, providing at the same time the interface for the management of the underlying data structure and the constant update of the terminological data.

1. Introduction to BISTRO

BISTRO (<http://www.eurac.edu/bistro>), the Juridical Terminological Information System Bolzano, is a free online platform developed by EURAC to assist the compilation of complex terminological data. It integrates a terminological database, bilingual and trilingual corpora and a specialised meta-search engine. With the help of BISTRO in-house collaborators validate terminological knowledge, enter new data into a database, check consistencies and present the data to the public.

For the external users such as lawyers or translators, BISTRO is a specialised terminological database. BISTRO allows feedback to the terminologists and, if need be, guides the user in his or her own terminological research, granting access to essentially the same tools as in-house terminologists may use. These tools support the translation process in various phases, i.e. the terminological preparation of the source text, the creation of terminological glossaries and the retrieval of related documents and their terminological elaboration.

As a general feature, BISTRO hyperlinks a terminology database with bilingual and trilingual corpora via the functions term extraction (TE), term recognition (TR) and keyword-in-context (KWIC). This process creates dynamically hyperlinked virtual documents, from which new queries may be started. For example, a text document can be converted into a glossary of term candidates. These candidates may be validated through an integrated meta-search that discovers additional documents from recommended sites.

What distinguishes BISTRO from more traditional terminology management systems is not only its dynamic and constructive aspects but

also its structuring of terminological data, which requires special efforts in terms of presentation: BISTRO has been designed to overcome the cumbersome and inefficient structuring of terminological data in “entries”. Data are organised in networks, in which terms, definitions, contexts, books, websites, laws, and grammatical specifications are unique nodes which are inter-linked by bidirectional graphs. These graphs represent relations as *translates as*, *is equivalent to*, *is documented by*, etc. This approach, which has been outlined in Streiter and Voltmer (2003) has been argued to allow not only for a more systematic and controlled collection of data, but also for a diversified and user-adapted presentation, grouping and sorting of the data. As we shall see, these improved presentation facilities are not merely nice to have, but they form an essential tool for contemporary terminographers in their production of high quality terminological data.

In order to help the external BISTRO user to understand the representation of terminological data, we suggest the introduction of another organising feature in terms of presentation. Since we abandon the uniform presentation format of terminological entries (easy to read, but possibly uninformative, imprecise and misleading) we elaborate a consistent chain of diversified presentations such that the visualisation of the current communicative perspective emerges as their constant feature. This presentation style marks the theme (the terminological element that fits best to the user query) as dark yellow, the rheme (new information to contextualise the theme) as light yellow, the focus (which ranks or filters the theme) by appropriate buttons, the shift of focus as orange and background information as light blue. In a nutshell, BISTRO provides a systematic rendering of the communicative structure of the query response. Since everybody uses communicative structures actively and passively in everyday conversation, coherent data structuring can be grasped intuitively (Humphreys 1997).

The main aim of this article is to explain the structural properties of BISTRO and to relate them to the requirements of our daily work on legal terminology.

2. Elaborating on legal terminology

South Tyrol is a region in northern Italy with three official languages: Italian, German and Ladin¹. The Autonomy Statute of Trentino-South Tyrol (DPR 31.08.1972, no. 670) grants the same legal status to German as to Italian in South Tyrol (art. 99-101) and requires, in practice, that all public administration (especially documentation from information brochures to laws, jurisdiction, circulars, etc.) can be conducted in both languages. Ladin has to be respected in the Ladin-speaking communes (art. 102) and may therefore be used in local administration. Hence, the usage of correct and coherent terminology is of utmost importance in the South-Tyrolean context. It is precisely in this realm that the activities of the section “Language and Law” of the European Academy of Bolzano have their main focus. The scientific

department is mainly concerned with the drafting and the elaboration of legal South-Tyrolean German and Ladin terms. Since 1994 it has been cooperating with the Commission of Terminology (TERKOM), which is made up of external experts and is committed to the standardisation of legal and administrative South-Tyrolean German terminology to be used in South Tyrolean public life.

Consequently, the current terminological activities of the section “Language and Law” are concentrated around (a) the description and elaboration of Italian legal terms and the German terms used in South Tyrol, while taking into consideration legal terms used in other German-speaking countries (Austria, Germany and Switzerland) and (b) the description and elaboration of Italian and Ladin legal terminology.

The comparative approach with more than one legal system is pushing standard approaches to terminology to its limits and beyond. More than, for example, in biological terminology, we have to account in legal terminology for the fact that each legal system “has its own legal realia, its own conceptual systems and even knowledge structure” (Šarčević 1997: 232). This means that if a specific concept does not occur within the legal system of a another country, it will not be part of the concepts that are relevant to judges, mayors or lawyers of that country, neither at the linguistic level nor at the conceptual level (Rossi, consulted 20.08.2004). Earthquakes and their terminology, for instance, play a paramount role in Italian laws on town planning, but this is not the case, for example, in German legislation. The main reason for this is that Germany, unlike Italy, is not a high-seismic zone. We can thus infer that it is relatively rare for an Italian concept to be matched by a single equivalent in the Austrian, German and Swiss legal systems. Different considerations can be made with regards to Ladin and South-Tyrolean German concepts: since they refer to Italian legislation, they have exact Italian matches. Consequently, they are absolutely equivalent.

However, the complexity of terminological data not only derives from the incongruence of legal systems. The terminology of each legal system in itself may feature homonymies and polysemies across or within its different sub-branches.

3. BISTRO in a nutshell

BISTRO has been created as a means of support to terminography and to store data unambiguously. It provides corpora and meta-search tools because the terminological research of EURAC is essentially based on the study of law texts, jurisdiction and authoritative manuals.

3.1. CATEX²: the Italian/German bilingual corpus

The Italian-German bilingual corpus is “a domain-specific parallel corpus of representative Italian/German texts in machine-readable form which cover the whole area of law and administration and show the use of the terms in various contexts” (Gamper 1998: 10). The most important documents in the corpus are: local laws of the Province of Bolzano and the codes of the Italian legislation, i.e. the Civil Code (*Codice Civile /Italienisches Zivilgesetzbuch*), the Civil Procedure Code (*Codice di Procedura Civile/Italienische Zivilprozessordnung*), the Penal Code (*Codice Penale/Italienisches Strafgesetzbuch*), the Penal Procedure Code (*Codice di Procedura Penale/Italienische Strafprozessordnung*), the Insolvency Code (*Fallimento ed altre procedure concorsuali/Italienisches Konkursrecht und andere Insolvenzverfahren*) and the Consolidated Text On Revenue Taxes (*Testo Unico delle Imposte sui Redditi/Einheitstext der Steuern auf das Einkommen*). CATEX contains “about five million words and 35,898 (66,934) different Italian (German) word forms” (Gamper & Dongilli 1999).

3.2. CLE³: the Italian/German/Ladin corpus

The trilingual corpus is based on about 5,000 official documents such as orders, regulations and records. Most of them originate from the municipalities of the Ladin valleys, some are translations by the *Translation Service 3.2 of the South-Tyrolean Government* of regional legislation. The legal sub-branch covered is administrative law. The corpus also includes non-legal documents such as news reports from the local government and publications provided by the institutes for the development and the conservation of Ladin. The corpus is meant to be balanced as it contains the same bulk of documents for both Ladin variants (Badiot and Gherdëina), different kinds of administrative texts such as orders, regulations and records, and both legal and non-legal documents such as news reports. For more details on this corpus see Streiter et al. (2004).

The corpus search mask provides multiple search criteria including languages (target and source), legal systems, passing and publication dates, abbreviations of the documents, etc. The terms can be retrieved in the corpora using regular expressions in order to refine the search, i.e. ~* ‘kindergarten’ search for school case insensitive.

<p>Premesso che nell'ambito dei lavori di realizzazione dell'area da gioco presso il locale scuola materna è stata riscontrata l'esigenza di dotare lo staccato che si affaccia sulla piazza del comune con utiloni elementi verticali, allo scopo di eliminare situazioni di pericolo dovute allo scavalco dello stesso da parte di bambini;</p>	<p>Vorausgeschickt, dass im Rahmen der Arbeiten zur Verwirklichung des Spielplatzes am örtlichen Kindergarten die Notwendigkeit erkannt wurde, den gegen den Gemeindeplatz errichteten Zaun durch zusätzliche senkrechte Lattenelemente zu vervollständigen, um Gefahrensituationen, die sich durch das Darüberklettern von Seiten der Kinder ergeben, zu beseitigen;</p>	<p>Tigni cunt che intratan i laurs de realisaziun dla plaza dai juc' d'ungia le local dla scola in el stè debojagn da arginè pro la si che c'iaa son plaza di comun stiches verticales laprò, por eliminè le prigo che i mituns vais sura fora;</p>
---	---	---

Figure 1. trilingual parallel legal corpus (Italian, German and Ladin)⁴

In addition to the corpora, BISTRO contains a database of bibliographic references where laws, regulations, text books, websites, etc. are collected and classified according to their subject areas, legal systems, legal hierarchies and legal qualities. For a discussion of these aspects see Streiter and Voltmer (2002). Wherever possible, electronic copies of these documents are stored in a monolingual text repository, which can be understood as pre-stages of corpora.

Corpus segments always link back to the bibliographic data, for example through the link DELJ la_val, 23.05.2002, n. 82 in Fig. 1.

N	id	abbreviation	object	legal_system	legal_quality	subject_area	legal_hierarchy	language	passing_date
1	5708559	DV-STAG	Staatsanwaltschaftsgesetz-DV	AT	law	admin.law	state	de	
2	5708562	RHG	Bundesgesetz über die Harmonisierung der Einwohnerregister und anderer amtlicher Personenregister (Registerharmonisierungsgesetz)	CH	law	admin.law	state	de	
3	5708572	FMedG	Bundesgesetz über die medizinisch unterstützte Fortpflanzung (Fortpflanzungsmedizingesetz)	CH	law	family law	state	de	18.12.1998
4	5708583	URG	Bundesgesetz über das Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz)	CH	law	admin.law	state	de	09.10.1992
5	5708587	SchVG	Bundesgesetz über die Direktversicherung mit Ausnahme der Lebensversicherung (Schadensversicherungsgesetz)	CH	law	insurance law	state	de	20.03.1992
6	5708594	KiG	Bundesgesetz über die Information der Konsumentinnen und Konsumenten (Konsumentinneninformationsgesetz)	CH	law	commercial law	state	de	05.10.1990
7	5708606	BV	Bundesverfassung der Schweizerischen Eidgenossenschaft	CH	law	law	state	de	18.12.1998
8	5708391	NHG	Bundesgesetz über den Natur- und Heimatschutz	CH	law	admin.law	state	de	01.07.2066
9	5708304	DBG	Bundesgesetz über die direkte Bundessteuer	CH	law	tax law	state	de	14.12.1990
10	5708279	UnStG	Bundesgesetz über die Studien an den Universitäten (Universitäts-Studiengesetz)	AT	law	university law	state	de	
11	5708465	StrVo	Strassenverkehrsordnung/Codice della Strada - deutsche Übersetzung für Südtirol	IT	law	traffic law	state	it-de	24.06.1998
12	1778692	WEG	Wohnbau- und Eigentumsförderungsgesetz	CH	law	social welfare	state	de	
13	5708221	Sacra Rota	Sacra Rota	IT	law	ecclesiastical	state	it	24.02.1981

Figure 2: legal text

Term-extraction (TE) tools can process texts from the internet, the text repository and the corpora and produce glossaries of term candidates. TE creates a list of term candidate where each term candidate can be searched against the BISTRO database ('bistro'), in the internal corpora ('CATEX') or on the internet ('Google').

Rank	Term Candidate	KWIC of Term Candidate	
1	plaza dai jüc	[Frame11]	bistro CATEX google KWIC
2	revardènta la condùta	[Frame12]	bistro CATEX google KWIC
3	<u>manutenziun</u> dla plaza	[Frame13]	bistro CATEX google KWIC

Figure 3: Term Extraction (TE) on a corpus segment

Term-recognition (TR) tools can match the text against the existing database and show which terms in the text are covered by the term base. Triggering the term recognition on the corpus search in Fig. 1, we obtain the results in Fig. 4.

*Term abstraction: # KWIC
Tignj cunt che intratan i leürs de realisaziun dla plaza dai jüc dlungia le local dla * scolina di stè debojègn de arjigné pro la si che clara sön plaza di * comun stiches
verticales laprò, por eliminé le gripo che i mituns vals sura fore; ARAC D'OBLIANZA art. 16 L. R. 23. 10. 1998, nr. 10 SOLA REGOLARITÈ TECNICA POSITÍF LE
* RESPONSABIL DL SORVISC sotescrit:

Figure 4: Term Recognition (TR) in the corpus

The terminographer's main goal is to understand and identify the meanings of the terms under investigation, as well as to find homonyms and synonyms and, finally, to check the usage of the terms. The usage of corpora, especially in the KWIC format may support this task.

<u>verschiedene Gegenstände für den Kindergarten von St. Leonhard</u>	* cle * doc
<u>1. Die Lieferung von Einrichtung im Kindergarten von St. Leonhard</u>	* cle * doc
<u>für den Ankauf von Einrichtung im Kindergarten von St. Leonhard</u>	* cle * doc
<u>09.2002 bis 17.06.2003 als Köchin des Kindergartens von St. Kassian</u>	* cle * doc
<u>Erfordernisse der Küchendienste des Kindergartens von St. Kassian</u>	* cle * doc
<u>Ankauf von Teppichen mit Bord für den Kindergarten von St. Kassian</u>	* cle * doc

Figure 5: Kindergarten as Keyword in Context (KWIC)

When a possible corresponding expression in the target language has been found, for example in a parallel corpus, it is necessary to verify its equivalence (partial or total) to the source language term. Distributional features (e.g. the legal hierarchy, legal system and the legal sub-branch) of source and target language terms may help to identify non-equivalences between the terms. A special tool for this purpose has not yet been integrated into BISTRO, but this could easily be done by the terminographer, without the assistance of the system developer, simply by defining a specific VIEW (see our discussion of views below).

Finally, the terminology management system is in charge of storing, retrieving and visualising the available terminological knowledge. Further checks in the terminology management system should support the search of synonyms and false friends, the identification of poly-

semies to group terminological findings per legal system, per legal sub-branch or per language, to discover gaps and inconsistencies. For these latter steps it has become necessary for BISTRO to follow unconventional paths and to overcome the limitations of current entry-based models. An entire section will be dedicated to this subject.

4. From file cards to BISTRO

Traditional terminology has worked with fixed knowledge units condensed in lexical entries. In early times a lexical entry was all information that was filed on a single paper card. This partition arrangement of terminological knowledge was reproduced when data became electronic. Many terminology management tools are still based on this model. Their common credo states that all terminological information pertaining to one concept including all terms in all languages must be handled as one terminological unit (Schmitz 2002). This approach keeps data simple and controllable at the expense of arbitrary cuts in the linguistic continuum and the uncontrollability of the relations between lexical entries. In fact, most relations in a traditional database are not visible, obfuscating terminographers and users.

Let us for a while follow Schmitz (2002), who suggests a standard organisation for terminological data which first divides the ‘concept’⁵ according to ‘languages’, then into ‘terms’ (Fig. 6). A communicative model, as we propose it, does not exclude this structure. In fact, this standard structure represents a valuable view on data. When used as structure for term presentation, our model would identify this standard structure as follows: The ‘concept’ is the theme, the ‘language’ the focus and the ‘terms’ the rheme. However, when this structure is intended to represent the underlying data structure, a fair number of data are lost or must be expressed implicitly, e.g. the relation between the terms within or between legal systems (e.g. abbreviation of, variant of, translates as). The relation to terms outside this structure is lost, which is generally amended by creating named hyperlinks (e.g. antonym). This additional device, however, is hampered by the fact that a link to only one concept is possible. This is insufficient and creates ambiguities (does this link really refer to the concept or a term, if so which one). Consequently, lists and taxonomies of any type have to be constructed outside the traditional entry structure. Still, the problems within the „entry structure“ remain even after supplying external helping constructs.

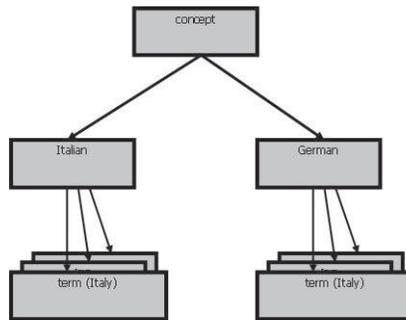


Figure 6: standard data model according to Schmitz

For the management and control of the terminological data, it is certainly of interest to focus on the ‘country’ (Fig. 7a) to check gaps and synonyms within a legal system. Yet, the standard structure can neither store nor display this information. Technically possible transformations would be inconsistent, because they would be in contradiction to the predefined logical structure.

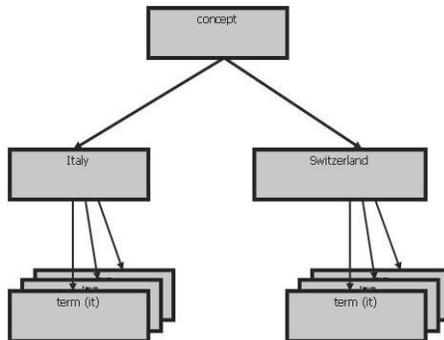


Figure 7a: alternative data model

For verification purposes, it might also be interesting to start with the ‘term’ and not the concept as theme. This would provide an overview over associated ‘concepts’. One might even think of sorting the concepts according to the legal sub-branches or legal systems. This could reveal false friends, differences in legal systems or polysemies.

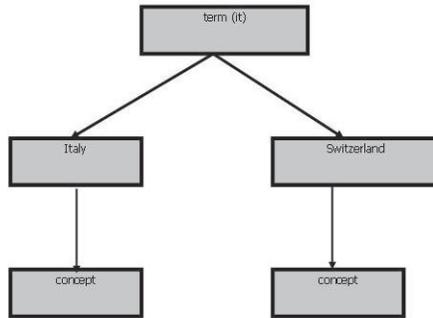


Figure 7b: alternative data model

The advantages of supporting different communicative perspectives go beyond content structuring of electronic data (e.g. in XML) or user-adapted presentation (e.g. in XHTML or pdf). The communicative perspectives can be created only from a fully informed data model. For this reason we have to join the leaves of the tree structure and work with a network as underlying data structure which trespasses the borders of the traditional entries. The structure of underlying data is sketched in Fig. 8, in which all nodes are unique. This corresponds to a normalised relational model.

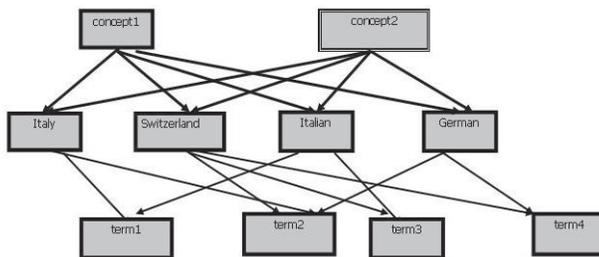


Figure 8: unified dynamic datanet

In this model, nodes are used to represent ‘terms’, ‘concepts’, ‘text segments’ and other elementary data categories. These elementary data categories can easily comply with existing standards such as XCES and Dublin Core, without any need to find an overall standard for the entire system. This architecture guarantees maximal use of the data for a great variety of applications such as monolingual and multilingual dictionaries, specialist and general dictionaries, dictionaries with or without definitions, contexts or other categories. The data structure we are developing for BISTRO along these lines is sketched below.

5. BISTRO's data structure

EURAC's terminological knowledge was originally collected in entries. These entries are currently dissolved into a network implemented in a relational data model (Streiter and Voltmer 2003). Practical reasons favoured the relational model instead of an XML-based implementation: Full-fledged relational databases are freely available; they offer a rich environment for the management of millions of data and have short query times. In addition, multiple users may update and query data simultaneously. In a nutshell, terminology with relational data can exploit all the benefits of modern data management, while XML-based implementation is still in its infancy.

The terminological entries were dissolved into about 25 tables of elementary data categories: denominations (words and expressions), grammatical information, legal system, legal hierarchy, legal quality, subject area, normation status, processing status, language planning qualifier, document contents, document meta-data, alignment data for parallel texts, translation relations between denominations and some more. Tables and relations form a network where tables are nodes and relations are the arcs.

Tables are organised into VIEWS, which show related tables as a larger group, thereby cutting the network into tree structures with a starting point (theme) and end points (rheme). This is illustrated with two tables, the denominations and the grammatical information.

id	object
15	noun masculine
16	noun feminine
17	noun neuter

Table 1: denomination

id	object	gram_id
1	Präsident	15
2	Vorsitzender	15
3	Rektor	15

Table 2: grammar

Unified into a view with <i>theme</i> : Präsident			Unified in the same view with <i>theme</i> : noun masculine		
id	grammar	denomination	id	denomination	grammar
1	noun masculine	Präsident	1	Präsident	noun masculine
2	"	Vorsitzender			
3	"	Rektor			

Table 3: the contents of table denomination and table grammar

VIEWS can be simple or complex. According to the number of arcs followed in the network, the VIEWS span one, two or many relations. The synonymy VIEW, for example, follows the translation relation and then the back-translation of a term into the original language within the same legal sub-branch and legal system. Such information provides insight into the term as well as into the consistency of the data. This additional knowledge drawn from the network helps the user navigate. The query “Präsident”, for example, yields the synonyms “Rektor” and “Vorsitzender”, the former limited to university law, the second in general usage (Table 3).

The same view, fed with a slightly different query, follows different paths and yields ‘related terms’ as contextualisation of the query term and as navigation proposal (Table 4). The less expert a user is, the more such guidance is useful.

(a) Quasi-synonyms of ‘Präsident’ calculated by VIEWS ⁹	(b) Related terms of ‘autostrada’ (5), calculated by VIEWS
[Frame9]	[Frame10]

Table 4

VIEWS can account for extremely complex relations. BISTRO models entire multilingual corpora as views which integrate text segments, alignment information and meta-data (Streiter et al. 2004). The rendering of a VIEW is not different from the rendering of a single TABLE: The data retrieved from an SQL query are uniformly transformed into an internal XML structure, which includes communicative parameters, e.g. the theme, the rheme and the focus. VIEWS and tables are associated with specific XSLT style sheets which in cooperation with CSS style sheets account for the rendering in the Web Interface. VIEWS, XSLT style sheets and the association between them can be created and modified by the terminographer, thus creating complex queries and the adequate presentation of the query results. Unless deleted, such VIEWS become specific tools, such as the suggested tool for checking the legal systems and legal hierarchies of terms through classified corpora. VIEWS are defined in PROLOG-like expressions. The similarity to PROLOG expressions illustrates the deductive power:

```
synonym(source=source,target=target,subject_area=sa,legal_system=ls):  
denomination(object=source,id=source_id,language=l),  
translation(object=source_id,target=mid_id,subject_area=sa,legal_sys-  
tem=-ls),  
denomination(id=mid_id),  
translation(object=mid_id,target=target_id,subject_area=sa,legal_sys-  
tem=-ls),  
denomination(object=target,id=target_id,language=l).
```

6. Conclusions

BISTRO tries to find new ways in terminological data management and presentation. In contrast with the common credo, BISTRO abandons terminographical data models completely. BISTRO stores data in standardised elementary units and avoids arbitrary and non-standard manipulation of data. The resulting network of data creates an unlimited number of combination possibilities. The complexity of such a system can successfully be handled with VIEWS. The VIEWS show data in meaningful combinations and allow effective data management at the same time. BISTRO structures VIEWS and tables from a communicative perspective. The consistent structuring of the user communication guides newcomers as well as expert users through the search process.

The potential output comprises all traditional terminological products such as general and special monolingual and multilingual dictionaries, glossaries and translation memories as well as all imaginable new products such as definition dictionaries or false friend lists. The elementary data structure allows effective data transfer and combination. Recombination, inference and deduction of elementary terminological knowledge open the door towards the spheres of artificial intelligence (AI). BISTRO's network of terminological knowledge can be exploited by all term tools developed in computer linguistics. Terminologists can generate new terminological data and gather new insights even by standard combinations of views and terminology tools, because all output can be put in again in cyclical processing.

BISTRO handles 20,000 external searches per month, a proud number for a specialists' tool. Its showcase approach to terminology has attracted interest in large projects and will spread in the near future under different names. Even if not all problems of the presented entry-less terminology management are resolved yet, we expect it to become a serious alternative to commercial terminology management in the near future.

Bibliography

- Ciola, B. & R. Natascia (2003). “Web-based term mining tra terminologie e memorie di traduzione”. Paper presented at *Seminario sulle memorie di traduzione* (Libera Università S. Pio V – Rome, 30 September – 01 October 2003).
- Gamper, J. (1998). “CATEX– A Project Proposal”. *Academia* 14, 10-12.
- Gamper, J. & P. Dongilli (1999). *Primary Data Encoding of a Bilingual Corpus*. <http://titus.uni-frankfurt.de> (consulted 01.08.2004)
- Humphreys, K. (1997). *Formalising Pragmatic Information for Natural Language Processing*. PhD thesis. University of Oxford.
- Rossi, P. *Ontologie applicate e comparazione giuridica: alcune premesse*. <http://www.dsg.unito.it> (consulted 20.08.2004)
- Sager, J. C. (1990). *A practical course in terminology processing*. Amsterdam/Philadelphia: John Benjamins.
- Šarčević, S. (1997). *New approach to legal translation*. The Hague: Kluwer Law International.
- Schmitz, K.-D. (2002). “Towards a uniform environment for representing terminologies within ISO”. Paper presented at *TKE* (Nancy, 28-30 August 2002).
- Streiter, O. & L. Voltmer (2002). “Document Classification for Corpus-based legal Terminology”. Paper presented at *The 8th International Conference of the International Academy of Linguistic Law* (Iasi - Romania, 24-26 May 2002).
- Streiter, O. & L. Voltmer (2003). “A Model for Dynamic Term Presentation”.. *Proceedings of the TIA-2003 Conference*. Strasbourg: LIIA – ENSAIS & Université Marc Bloch, 201-204
- Streiter, O., M. Stuflesser & I. Ties (2004). “CLE, an aligned Tri-lingual Ladin-Italian-German Corpus. Corpus Design and Interface”. *First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation*. LREC 2004. Paris: ELRA, 84-87
- Wright, S. E. & G. Budin (eds) (1997). *Handbook of Terminology Management*, vol. 1/2, Amsterdam/Philadelphia: John Benjamins.

-
- ¹. Ladin is a Rhaeto-Romance language spoken by 30,000 people in northern Italy.
- ². **C**omputer **A**ssisted **T**erminology **E**xtraction
- ³. **C**orpus **L**adin **d**l’**E**URAC (Ladin Corpus of the EURAC)
- ⁴. The paragraph displayed is part of one of the documents belonging to the trilingual corpus. All documents are originals and have not been edited, translated or reviewed by EURAC employees.
- ⁵. As a typographic convention, we will enclose data categories in single quotation marks and reproduce communicative labels in italics.

