Managing Language Varieties: Examples From Legal Terminology Work

Natascia Natascia Ralli

Eurac Research, Institute for Applied Linguistics, Bolzano/Bozen, Italy

Abstract

Handling pluricentric languages requires addressing their language varieties. This paper explores strategies to represent these varieties in terminology databases, considering factors such as the quantity of terminological data and the availability or absence of language identifiers. Using the legal domain as a reference point, the analysis examines the associated challenges, as well as the advantages and disadvantages of various approaches to representation.

Keywords

language variety, terminology database, legal terminology

1. Introduction

Dealing with multilingual terminology work involves navigating different cultures and languages. Some languages are pluricentric [1], meaning they are used in at least two countries where they have "an official status as state language, co-state language, or regional language" [2]. These languages have multiple standard varieties tied to distinct national or regional contexts [3], referred to here as 'language varieties'. Examples include English, with British, American and Australian varieties, and German, which has standards such as Austrian German and Swiss German.

Well-known differences between language varieties can concern spelling (e.g., British *colour* vs. American *color*) and grammar (e.g., perfect tense usage with *sein* or *haben* for some verbs in German: *ich bin* vs. *ich habe gesessen/gestanden*). However, the most significant challenges in terminology work stem from terminological differences. Explicitly addressing such differences in terminology resources is essential to ensure effective communication and business interactions.

Reflections on language varieties and their representation in terminology databases have been central to the terminology work carried out by the Institute for Applied Linguistics (IAL) of Eurac Research since the mid-1990s. Terminological data in the law domain are compiled in Italian, German, and Ladin[1] and published online in the Information System for Legal Terminology *bistro* [4]. Legal terminology is inherently system-bound [5] [6]: each legal system is shaped by its own historical, economic, social and ethical context, which leads to the development of "its own legal realia and thus its own conceptual system and even knowledge structure" [7]. The resulting legal terminology is unique to the specific legal framework in which it operates, reflecting the values and practices of that framework, which might be different or irrelevant to other legal systems. Such instances can occur even when the same language is used in different legal systems. Consequently, those legal systems may use distinct designations for the same legal concept. For example, the concept of 'stalking' has distinct designations in German depending on the legal system: *beharrliche Verfolgung* in Austria, *Nachstellung* in Germany and *Verfolgungshandlungen* in South Tyrol. Similarly, the same designation may refer to different concepts. For example,

Natascia 4th International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT) 2025, June 19-20, 2025, Thessaloniki, Greece.

natascia.ralli@eurac.edu

0000-0002-9663-8169

[1] Ladin is a Rhaeto-Romance language mainly spoken in some Dolomite valleys in Northern Italy, notably in the Trentino-Alto Adige and Veneto regions.

Befreiungsschein denotes exemption from medical fees in Germany but a work permit for foreigners in Austria [8]. Conversely, the same designation may have a similar meaning across various legal systems [9], as seen with *lockdown*, a COVID-19 containment measure, albeit with variations in regulatory implementation, even across legal systems that share the same language [10].

For this reason, in IAL's daily terminology work, legal comparisons are made between Italian and German-speaking legal systems (Austria, Germany and Switzerland), as well as EU and international law. For every Italian term, we provide equivalents for each German-speaking legal system. For Ladin, we provide terms in the language varieties spoken in the South Tyrolean valleys, namely Val Gardena and Val Badia.

This paper proposes how to represent language varieties in terminology databases, with a particular focus on Trados MultiTerm[2], a commercial terminology management system (TMS) used by the IAL since the mid-1990s. Section 2 describes the terminological metamodel for structuring terminology databases. Section 3 provides some possible ways of representation by considering the presence or absence of terminological data and language identifiers. The analysis also encompasses the related challenges as well as the advantages and disadvantages of each representation. Section 4 concludes the discussion.

2. The terminological metamodel

Terminological data should be organized and managed according to the terminological principles [11, 12, 13]:

- each concept entry should contain information about a single concept (*concept orientation*);
- all terms (e.g., synonyms) in a concept entry are treated as independent sub-units. As such, they are described using the same set of data categories (*term autonomy*);
- data categories should be finely defined (*data granularity*);
- data categories should contain only one data element (*data elementarity*).

Their representation is defined in ISO 16642:2017 [12]. This standard provides a terminological metamodel consisting of two levels of abstraction. The first is the metamodel level, which supports analysis, design and exchange at a broad level [12]. The second is the data model level, which adds the necessary data categories for representing a specific terminological data collection.

This paper focuses on the structure of a concept entry (Figure 1), which is organized into three levels: the concept level (concept entry), the language level (language section) and the term level (term section) [14].

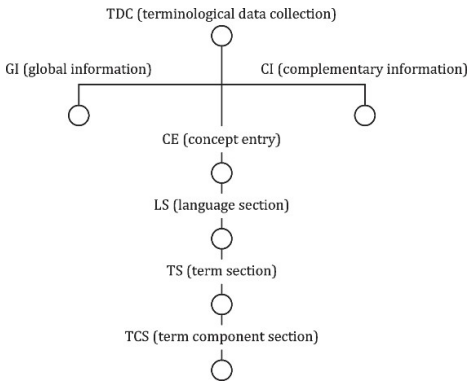

Figure 1: Terminology metamodel – Simplified schematic view [12].

[2] https://www.trados.com/product/multiterm

The first level contains administrative data and language-independent terminological information relevant to the entire concept entry (e.g., /creation date/, /domain/,) [13]; the second level is used to instantiate information about the concept that needs to be available in the respective language (e.g., terms, /definition/) [11]; the third level contains all term-related information (e.g., /context/, /usage note/) [11, 13]. According to cardinalities [12], a single concept can be expressed in *n* languages. A language section can incorporate one or more term sections.

In compliance with terminological principles, a language variety should be treated as a language and, thus, stored at the language level. Nowadays, most TMS support language varieties. According to ISO 639 [15], these are usually assigned to a language identifier (Table 1).

Table 1
Language identifiers for German language varieties[3]

Language variety	**Language identifier**
German (Austria)	de-AT
German (Liechtenstein)	de-LI
German (Luxembourg)	de-LU
German (Switzerland)	de-CH

Language identifiers ensure data interoperability and exchange with existing applications. However, to enable smooth data interchange and interoperability, two conditions must be met:

- Terminological data must have the same structure;
- Language varieties must have language identifiers.

In daily terminology work, the following scenarios can arise:

- The terminology database is empty, containing no terminological data. In this case, it must be structured from scratch, considering language varieties.
- The terminology database contains terminological data and is organized by languages. However, the inclusion of language varieties now requires an *ex-post* intervention. The question is the extent to which modifications can be made to the existing database structure.
- The terminology database includes language varieties that have not yet been codified.

These scenarios are not mutually exclusive; instead, they are interrelated and may co-exist. In the section that follows, we describe the available options to represent them. To ensure clarity and consistency throughout the paper, we will use the term ‘uncodified language variety’ to refer to a language variety without an ISO language identifier, as opposed to ‘codified language variety’.

3. Three ways of representation

3.1. Language varieties as an attributive data category

Before the early 2000s, TMS providers did not support language varieties. As a result, terminology databases created in those years were typically organized by languages rather than language varieties. Suppose

1) there is no possibility of creating a new terminology database, perhaps, due to a large volume of terminological data or limitations in human and financial resources, and/or

[3] https://www.andiamo.co.uk/resources/iso-language-codes

2) the language variety lacks a language identifier. This is the case, for example, of minority languages like South Tyrolean German or Ladin varieties like Gherdëina, Badiot, Fascian, and others.

In both cases, we can treat the language variety as an attribute of a term. To this end, we can add a specific data category (e.g., /geographical usage/, /legal system/) at the term level and define it as a picklist. The values of the picklist can be the language or country codes based on ISO 639 [15] or 3166[4] [16] [13], respectively. Furthermore, fields such as /definition/, /context/ or /usage note/ should be distinguished by, for example, inserting a language or country code within the data category. This approach facilitates filtering and exporting data while clearly indicating which fields belong to a specific language variety (Figure 2).

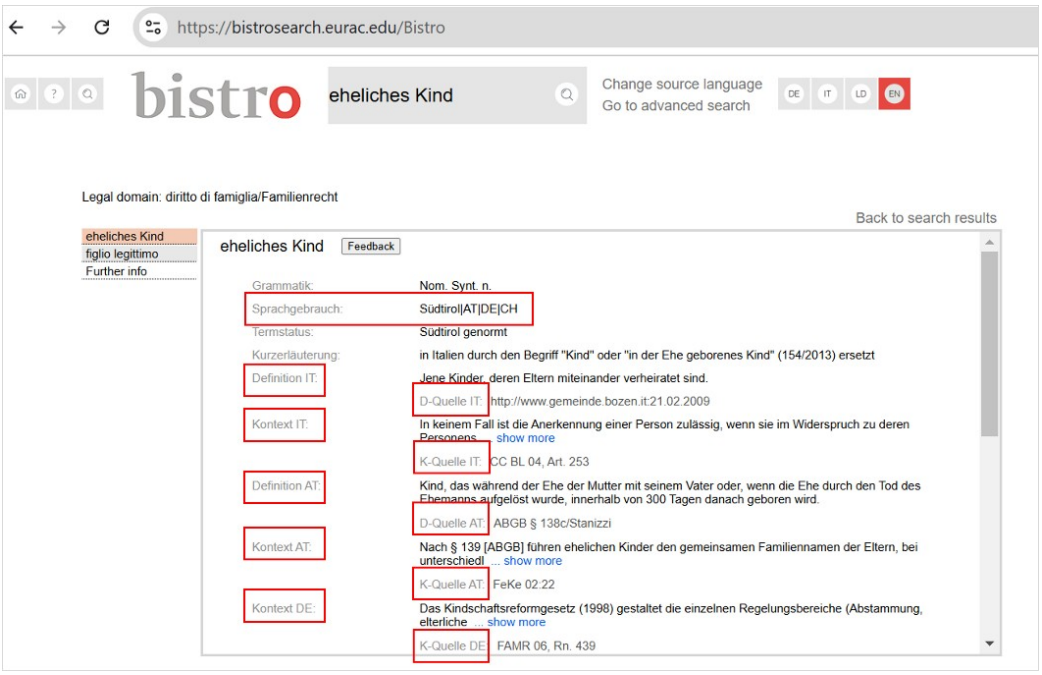

Figure 2: Data category /Sprachgebrauch/ (geographical usage) and related fields with country codes in *bistro* [4].

This solution can serve as a viable compromise when significant modifications to the database structure are not feasible or when working with uncodified language varieties. However, it violates the terminological principle of term autonomy (see Section 2). Consequently, labelling a preferred term or indicating its status for each language variety becomes difficult. Other strategies are needed to give this information. For example, we can explicitly indicate the preference or the status in a dedicated open or closed data category[5] at the term level. Figure 3 shows an example of how to convey such information. Preference is expressed by the closed data category /Termstatus/ (term status) and its picklist value *Südtirol genormt*, indicating that the term has been standardized for use in South Tyrol by a Terminology Commission. Obsolescence is conveyed through the open data category /Kurzerläuterung/ (short note), which specifies a terminological change. For instance, the use of *eheliches Kind* (legitimate child) in the Italian legal system, expressed in German for South Tyrol, is documented in this manner.

[4] See http://www.lingoes.net/en/translator/langcode.htm and https://www.iban.com/country-codes.
[5] According to [11], open data categories are free-text categories like /definition/, /context/ or /note/, while closed data categories contain a finite set of predefined values. Examples of closed data categories are /domain/, /status/ or /geographical usage/ (see also [13]).

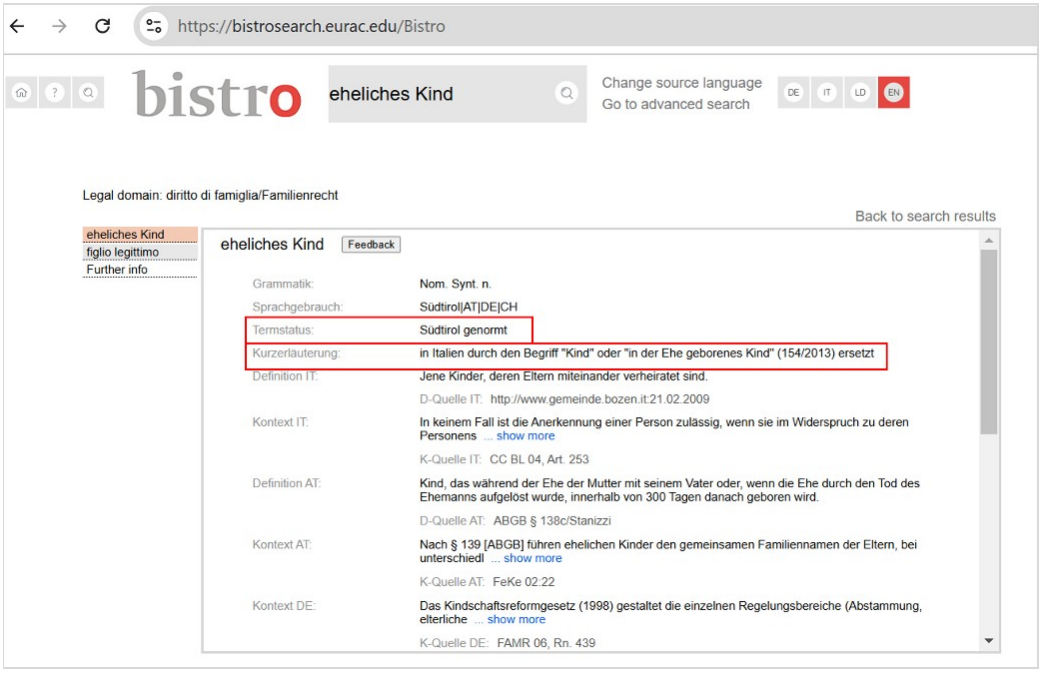

Figure 3: Data categories /Termstatus/ and /Kurzerläuterung/ in *bistro* [4].

This type of representation and the creation of ad hoc data categories complicate data exchange and interoperability, as it does not conform to the terminological metamodel: Being an attributive data category, the language varieties are not organized within a dedicated language section. Indeed, they fall under the language 'German', expressed by the *xml:lang* attribute <language lang= "DE" type= "Deutsch"></language>. The XML excerpt from the terminological entry *eheliches Kind* (Figure 4) illustrates this approach.

```
<languageGrp>
  <language lang="DE" type="Deutsch"/>
 <termGrp>
   <term>eheliches Kind</term>
  <descripGrp>
    <descrip type="Grammatik">Nom. Synt. n.</descrip>
  </descripGrp>
  <descripGrp>
    <descrip type="Sprachgebrauch">Südtirol|AT|DE|CH</descrip>
  </descripGrp>
  <descripGrp>
    <descrip type="Termstatus">Südtirol genormt</descrip>
  </descripGrp>
  <descripGrp>
    <descrip type="Kurzerläuterung">in Italien durch den Begriff "Kind" oder "in der Ehe
    geborenes Kind" (154/2013) ersetzt</descrip>
  </descripGrp>
  <descripGrp>
    <descrip type="Definition IT">Jene Kinder, deren Eltern miteinander verheiratet sind.
    </descrip>
  </descripGrp>
  <descripGrp>
    <descrip type="D-Quelle IT">http://www.gemeinde.bozen.it:21.02.2009</descrip>
  </descripGrp>
  <descripGrp>
    <descrip type="Kontext IT">In keinem Fall ist die Anerkennung einer Person zulässig,
    wenn sie im Widerspruch zu deren Personenstand eines ehelichen oder legitimierten
    Kindes steht.</descrip>
  </descripGrp>
  <descripGrp>
    <descrip type="K-Quelle IT">CC BL 04, Art. 253</descrip>
  </descripGrp>
  <descripGrp>
    <descrip type="Definition AT">Kind, das während der Ehe der Mutter mit seinem Vater
    oder, wenn die Ehe durch den Tod des Ehemanns aufgelöst wurde, innerhalb von 300 Tagen
    danach geboren wird.</descrip>
  </descripGrp>
```

Figure 4: Trados MultiTerm XML excerpt of the terminological entry *eheliches Kind.*

As is evident from the XML, this language section includes a term section, which contains term-related information concerning different legal systems and, hence, distinct German language varieties. This can also make interaction with MT tools more challenging.

3.2. Language varieties at the language level

Language varieties can be stored at the language level when the following conditions are satisfied:

a) The terminology database contains terminological data organized by languages, but there is the possibility of reorganizing it by adding language varieties.
b) The terminology database is empty. Thus, it must be structured from scratch with consideration for language varieties.
c) The language varieties present in the database all have language identifiers.

These ideal scenarios enable the representation of language varieties in a methodologically and technically accurate manner. In full compliance with terminological principles and the terminological metamodel, we can structure terminological data into the concept, language, and term levels, whereby the language level separates terms in one language variety from terms in other language varieties (Figure 5).

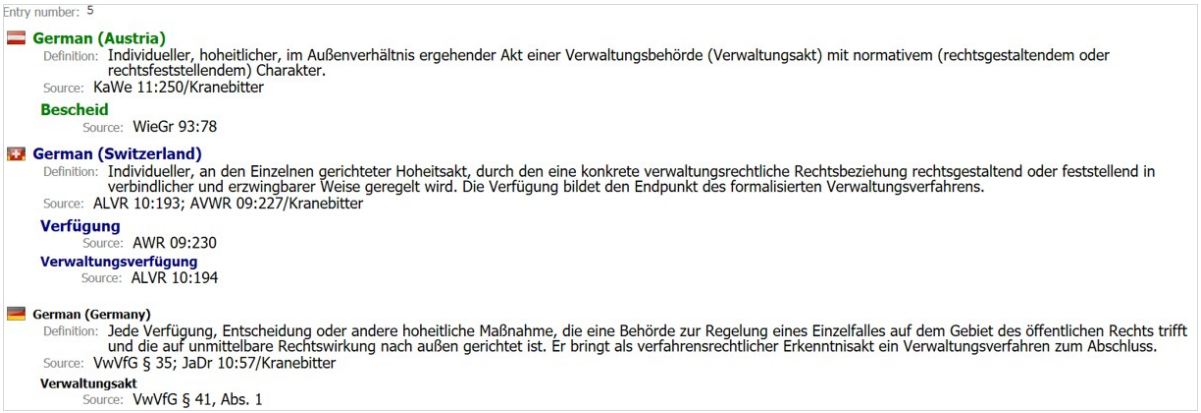

Figure 5: Representation of three German language varieties stored at the language level in Trados MultiTerm.

With this representation, each language variety has its own language section containing one or more term sections. Every language section is identified by a specific *xml:lang* attribute. In the case of Figure 5, these attributes are:

- <language type="German (Austria)" lang="DE-AT" />
- <language type="German (Germany)" lang="DE-DE" />
- <language type="German (Switzerland)" lang="DE-CH" />

This representation avoids creating ad hoc data categories and allows the use of harmonized data categories in compliance with ISO 12620-1:2022 [17] and ISO 12620-2:2022 [18][6]. Furthermore, it enables anchoring the definition at the language level, which is essential for the legal domain. Given the system-bound nature of legal terminology (see Section 1), distinct definitions are required for each legal system. However, this approach is similarly relevant for other domains, like religion, which lack the same cognitive background or internationalization [19]. Additionally, this type of representation simplifies the labelling of a preferred term or the indication of its status, as

[6] See the data category repository DatCatInfo (www.datcatinfo.net).

compared to the approach discussed in Section 3.1. It also enhances the smoothness of exporting or filtering terminological data, interactions with MT tools, data exchange, and interoperability.

However, this representation can generate data redundancy if multiple language varieties from the same language are involved. For instance, the same designation may occur with a very similar meaning in several legal systems. This is the case of *lockdown* (see Section 1) and *Vertrag*. The latter is commonly used in German-speaking legal systems to designate a 'contract' or 'agreement' (Figure 6).

3.3. Language varieties at the language level without a language identifier

The third way of representation is unconventional: artificially assigning an uncodified language variety to a codified one that is otherwise unused in the terminology database (see [20]). This extreme solution enables the storage of uncodified language varieties at the language level. It can be used when dealing with language varieties that the TMS does not support and/or for which there are still no language identifiers.

Figure 6 illustrates this method with two language varieties of Ladin. The concept entry's front end displays the language Ladin alongside its language varieties. However, in the back end, "TA-IN" (Tamil India) and "TA-MY" (Tamil Malaysia) are used as language identifiers. Naturally, this solution precludes data interoperability unless adaptation work follows.

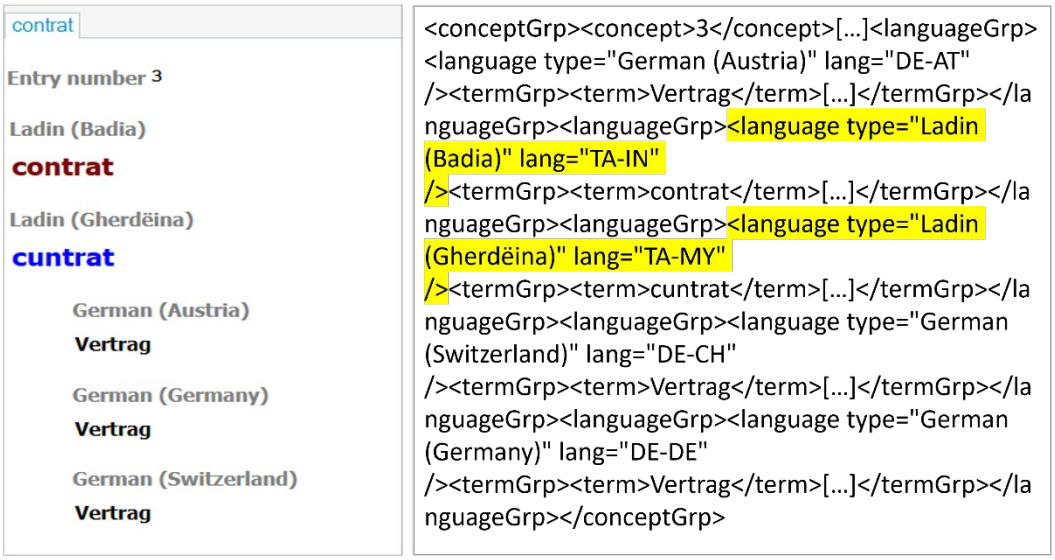

Figure 6: Front and back end of a terminological entry with two Ladin language varieties in Trados MultiTerm.

The third representation makes the complexity of accommodating language varieties in terminology databases even more evident.

4. Conclusions

This paper presents three ways to representing language varieties: the first (Section 3.1) offers a viable compromise, the second (Section 3.2) represents the ideal solution and the third (Section 3.3) employs a workaround. Many factors influence the choice of approach, such as the presence or absence of language identifiers, the amount of terminological data and the availability of human and financial resources to modify a database, particularly for retroactive adjustments. In this regard, it would be desirable to establish guidelines —potentially at the ISO level— to handle attributive categories and related fields (see Section 3.1). Such guidelines could ensure smooth data exchange and interoperability, especially for terminology databases unable to alter their structure due to the large volume of data and the number of working languages involved.

In the case of uncodified language varieties (e.g., South Tyrolean German), one solution might be the development of a generic language identifier to serve as a wildcard. This step would also benefit minority language varieties that currently lack a language identifier.

The discussion on this topic is far from complete. A future comparative analysis of existing tools and widely used TMS could assess their effectiveness in accommodating language varieties. Such an analysis would reveal whether and how the described approaches are implemented in practice, or if there are other methods of representation.

Declaration on Generative AI

In preparing this work, the author used Grammarly for grammar and spelling checks. The content was then reviewed and edited with assistance from a native English speaker. The author takes full responsibility for the content of this publication.

References

[1] M. Clyne, Pluricentric Languages – Introduction, in: M. Clyne (Ed.), Pluricentric Languages: Differing Norms in Different Nations, De Gruyter Mouton, Berlin, Boston, 1991, pp. 1–10. URL: https://doi.org/10.1515/9783110888140.1.

[2] R. Muhr, The state of the art of research on pluricentric languages: Where we were and where we are now, in: R. Muhr, K. E. Fonyuy, I. Zeinab, M. Coreyr (Eds.), Pluricentric Languages and non-dominant Varieties worldwide, volume 1, Peter Lang Verlag, Wien et. al., 2016, pp. 9-32.

[3] U. Ammon, U., H. Bickel, A.N. Lenz (Eds), Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen, 2nd. ed., de Gruyter, Berlin, 2016.

[4] *bistro*: Information System for Legal Terminology, URL: https://bistro.eurac.edu.

[5] G.R. de Groot (1999), Das Übersetzen juristischer Texte". Recht und Übersetzen, in: G.R. de Groot, R. Schulze (Eds.), Recht und Übersetzen, Nomos, Baden-Baden, 1999, pp. 11-46.

[6] D. Cao, Translating Law. Multilingual Matters, Clevedon, 2007.

[7] S. Šarčević, New approach to legal translation, Kluwer Law International, The Hague, 1997.

[8] R. Muhr, Österreichische und deutsche Rechtsterminologie – Typische Unterschiede und Probleme der Beschreibung plurizentrischer Rechtstermini', Schriftenreihe der Deutschsprachigen Gemeinschaft, 13, 2019, pp. 109–133.

[9] N. Ralli, Habe ich nun Vorfahrt, Vorrang, Vortritt oder soll ich doch lieber warten?, Ask a Linguist, Eurac Research Blog. URL: https://www.eurac.edu/en/blogs/connecting-the-dots/habe-ich-nun-vorfahrt-vorrang-vortritt-oder-soll-ich-doch-lieber-warten.

[10] N. Ralli, Natascia, I. Stanizzi, M. Alber, COVID-19 e lavoro terminologico: riflessioni a posteriori, AIDAinformazioni: Rivista di Scienze dell'Informazione, vol. 1-2 (2023), 2023, 91-114. doi: https://doi.org/10.57574/596529285.

[11] ISO 26162-1, Management of terminology resources – Terminology Databases – Part 1: Design, ISO, Genève, 2019.

[12] ISO 16642, Computer applications in terminology — Terminological markup framework, ISO, Genève, 2017.

[13] P. Drewer, K-D. Schmitz, Terminologiemanagement. Grundlagen – Methoden – Werkzeuge, Springer, Berlin, 2017

[14] ISO 30042, Management of terminology resources — TermBase eXchange (TBX), ISO, Genève, 2019.

[15] ISO 639, Code for individual languages and language groups, ISO, Genève, 2023.

[16] ISO 3166, Country codes, URL: https://www.iso.org/iso-3166-country-codes.html.

[17] ISO 12620-1, Management of terminology resources — Data categories — Part 1: Specifications, ISO, Genève, 2022.
[18] ISO 12620-2, Management of terminology resources — Part 2: Repositories, ISO, Genève, 2022.
[19] P. Sandrini, Terminologiearbeit im Recht. Deskriptiver begriffsorientierter Ansatz vom Standpunkt des Übersetzers. Braumüller, Wien, 1996
[20] N. Ralli, A. Norbert, bistro – ein Tool für mehrsprachige Rechtsterminologie, trans-kom 11 (2018) 7–44, URL: http://www.trans-kom.eu/bd11nr01/trans-kom_11_01_02_Ralli_Andreatta_bistro.20180712.pdf.