Open Terminology Management and Sharing Toolkit for Federation of Terminology Databases

Andis Lagzdiņš[†], Uldis Siliņš[†], Mārcis Pinnis[†‡], Toms Bergmanis[†‡], Artūrs Vasiļevskis[†] and Andrejs Vasiļjevs[†‡]
[†]Tilde / Vienības gatve 75A, Riga, Latvia
[‡]Faculty of Computing, University of Latvia / Raiņa bulv. 19, Riga, Latvia
{name.surname}@tilde.lv

Abstract

Consolidated access to current and reliable terms from different subject fields and languages is necessary for content creators and translators. Terminology is also needed in AI applications such as machine translation, speech recognition, information extraction, and other natural language processing tools. In this work, we facilitate standards-based sharing and management of terminology resources by providing an open terminology management solution – the EuroTermBank Toolkit. It allows organisations to manage and search their terms, create term collections, and share them within and outside the organisation by participating in the network of federated databases. The data curated in the federated databases are automatically shared with EuroTermBank, the largest multilingual terminology resource in Europe, allowing translators and language service providers as well as researchers and students to access terminology resources in their most current version.

Keywords: terminology, terminology management, termbank, terminology sharing, terminology database

1. Introduction

Language evolves: new words are coined, existing words change their meaning, and some even become unused. New concepts and terms that denote them are created every day, but many older concepts and their denotations rapidly become obsolete. Consequently, terminological data become obsolete over time if not regularly updated. Individual term collections are usually maintained by the respective institution, such as an industrial company, an academic centre, or a public administration. Still, many institutions lack a proper terminology management system and struggle to maintain their terms current. This has practical and financial consequences as consolidated access to current and reliable terms from different sources is necessary not only for content creators and translators but also for artificial intelligence (AI) applications.

Terminology management is even more challenging for termbanks that provide access to term collections aggregated from different institutions. Although terminology work benefits from a rigorous standardisation process and essential standards developed by ISO TC37 (Vasiljevs and Borzovs, 2006), insufficient supporting tools and infrastructure as well as different terminology management practices (including what data in what format is being stored) that are in place across Europe are factors that hinder terminology data sharing in a timely fashion (Gornostay, 2010).

In this paper, we describe a solution to these challenges that was created for the largest aggregation of European terminology resources, namely EuroTermBank[1] (Vasiljevs et al., 2008) and institutions participating in the EuroTermBank Federated Network. We present the EuroTermBank Toolkit (ETBT), an open terminology management toolkit for the EuroTermBank Federated Network that allows organisations to manage their term collections and share them within and outside the organisation.

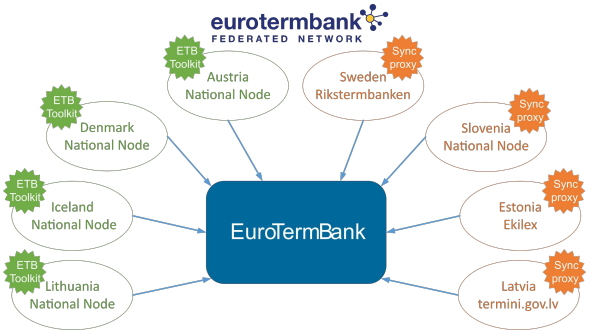

Figure 1: Federated nodes linked to the EuroTermBank Federated Network.

The motivation of this work was to support other initiatives of natural language processing (NLP) like automated text and speech translation with reliable terminology. In the following subsections, we briefly describe the application of terminology in these fields, then provide a short overview of EuroTermBank and the federated approach to terminology consolidation. Then, we continue with a description of the EuroTermBank Toolkit, its functionality and architecture, and the current state of the EuroTermBank Federated Network by providing statistics of terminology resources available within the network and institutions hosting federated nodes.

[1] https://www.eurotermbank.com/

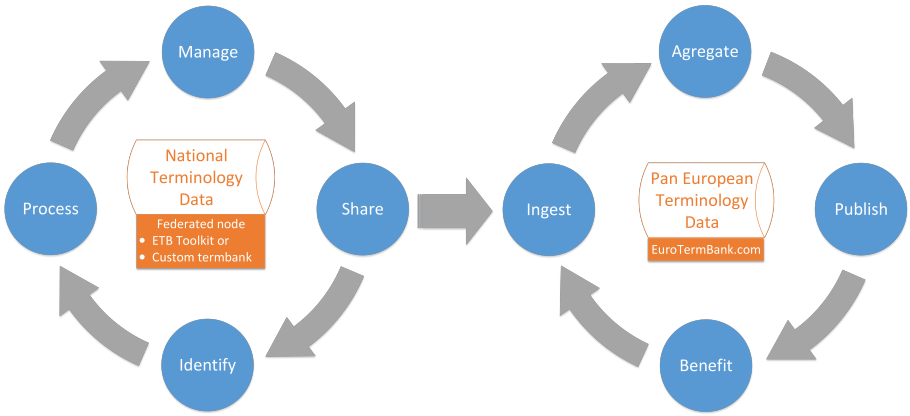

Figure 2: A conceptual depiction of EuroTermBank Federated Network.

1.1. Applications of Terminology in Natural Language Processing

While terminology in NLP is sometimes considered in a monolingual setting, most of its applications are related to multilingualism and translation. Terminology data has been proven to boost the quality of machine translation in the past (Pinnis, 2015) and has been helpful in the work of professional translators via computer-aided translation software (Arcan et al., 2014; Arcan et al., 2017; Verplaetse and Lambrechts, 2019). As of relatively recently, there has been a plethora of research on terminology integration in modern machine translation systems based on artificial neural networks (de Gspert et al., 2018; Dinu et al., 2019; Jon et al., 2021; Bergmanis and Pinnis, 2021b; Exel et al., 2020; Exel et al., 2020; Wang et al., 2022). The sheer volume of research on terminology integration in modern machine translation systems indicates a great interest from the industry of language service providers. Similar trends can be observed with the development of machine translation of speech (Cross Vila et al., 2018; Di Gangi et al., 2019; Vydana et al., 2021) and video subtitles (Matusov et al., 2019; Siekmeier et al., 2021; Schioppa et al., 2021), for which there is also a growing need for correct translation of terminology (Gaido et al., 2021).

Currently, however, the use of terminology in machine translation is not hindered by the lack of technology but rather by the lack of high-quality terminology data. Unlike statistical machine translation systems, which were robust to noise that is present in training data, the modern generation machine translation systems are susceptible to poor quality training data (Belinkov and Bisk, 2018; Khayrallah and Koehn, 2018). The same also applies to the quality of terminology data, which is often created by humans for humans only. Data created for human consumption is often unsuitable for machines as it contains irregularities that render it machine-unreadable (Bergmanis and Pinnis, 2021a)—findings which yet again emphasise the importance of standardised practices for the curation of machine-readable terminology data.

1.2. Overview of EuroTermBank

The objective of the work on EuroTermBank is to contribute to the advancement of the terminology infrastructure in all member countries of the European Union (EU) (Henriksen et al., 2006). The difference between EuroTermBank and other European terminology databases, such as the Interactive Terminology for Europe[2][3] (IATE) (Johnson and Macphail, 2000), is in their primary objectives. Although widely used by translators across Europe, the primary goal of the IATE database, for example, is to serve agencies and institutions of the EU by creating a centralised terminology platform for their translation needs. Thus, while IATE consolidates term collections of EU institutions, EuroTermBank is a collection of term collections of EU and many national and other institutions. As a result, the terminological data assembled in EuroTermBank is not created and managed by a single community but rather in a distributed fashion, often even by geographically focused working groups. The main stakeholders in maintaining the content of EuroTermBank are public institutions dealing with national or international terminology work. Examples are the State Language Centre of Latvia and the Institute of the Estonian Language, which coordinate terminology work in Latvia and Estonia. Other examples include the Institute of the Lithuanian Language, the University of Copenhagen, the Culture Information Systems Centre of Latvia, the Árni Magnússon Institute for Icelandic Studies, the Jožef Stefan Institute, the International Network for Terminology – TermNet, the Swedish Institute of Standards and the Institute for Language and Folklore.

These institutions continuously maintain their terminology resources, meaning that the terminology may be added, altered, and even discarded from the local term collections at any moment, and thus the terminology may change constantly. This poses a challenge for EuroTermBank that aggregates the terminology resources

[2] https://iate.europa.eu/

[3] IATE was originally named as the Inter-Agency Terminology Exchange.

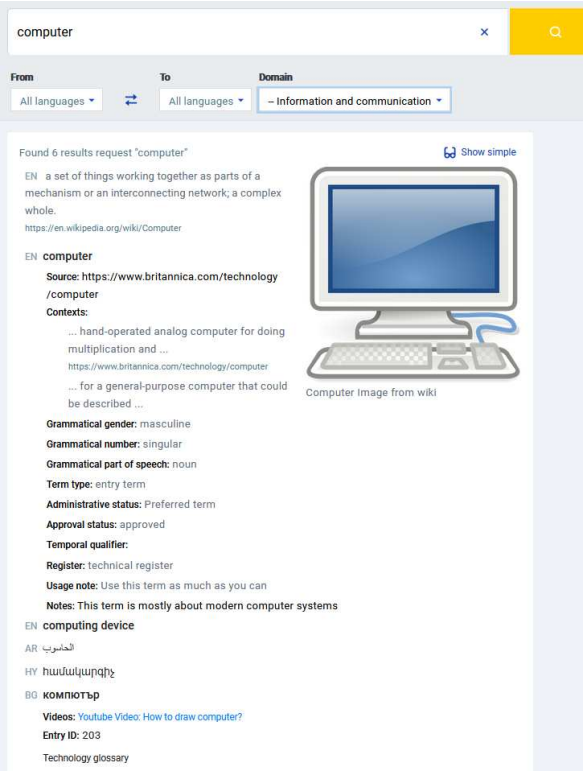

Figure 3: Term search view of ETBT.

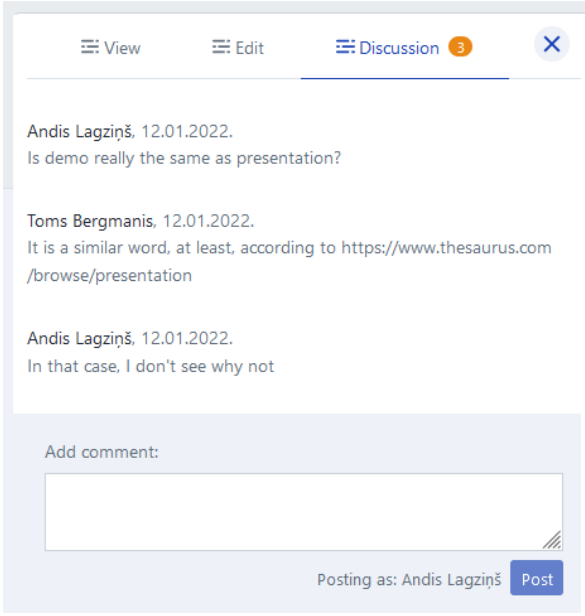

Figure 4: Term discussion view of ETBT.

of these institutions. If terminology keeps changing, there needs to be an automated process that ensures the currentness of terminological data in the global terminological databases.

1.3. Federated Approach in Terminology Consolidation

The necessity to move away from a single, isolated data bank towards a multi-bank environment was suggested by Cabré (1999), who proposed simultaneously accessing several data banks that are all integrated into an overall working structure that includes not only the databases but also other computerised tools and resources. The notion of the collection of cooperating database systems that are autonomous and possibly heterogeneous has been proposed before (Sheth and Larson, 1990). However, it is Galinski (2007) who foresees the federation of term banks as a new concept in linking portals and data repositories that will go far beyond the establishment of pointers or links towards the level of exchangeability and semantic interoperability of data and data structures.

A federated approach to consolidate distributed terminology resources was foreseen from the very beginning of the development of EuroTermBank (Vasiljevs and Rirdance, 2007). The first implementation used distributed search queries over interlinked external termbases and aggregated returned results in a consolidated search results view. This implementation was eventually phased out due to serious practical drawbacks. External bases provided their results in proprietary formats that tended to change over time. Consolidation of different results into a unified structure for representation was complicated because of data format incompatibilities. There were significant delays in providing consolidated output to users due to frequent performance issues in some of the interlinked termbases. For this reason, the federated approach presented in this paper consists of a homogeneous network of participating institutions that use unified data exchange mechanisms based on the latest versions of the TBX standard. Networked institutions either adapt the API of their existing databases to comply with the requirements of EuroTermBank Federated Network or migrate to the open EuroTermBank toolkit. Shareable data is dynamically synchronised with the central EuroTermBank database, where it is consolidated with resources coming from all participating institutions.

1.4. Aims of EuroTermBank Toolkit

The ETBT aims to guarantee the **currentness of the terminological data** available at EuroTermBank by synchronising it with EuroTermBank federated nodes of organisations and institutions throughout Europe. The ETBT also aims to facilitate the **streamlining and standardisation of the terminology curation and sharing** practices throughout Europe, thus lowering the cost and effort required to share terminological data for both the data owners and data users. Last but not least, the open nature of the terminology management toolkit intends to **eliminate the need for non-standard processes** in terminological data sharing.

2. Concept of EuroTermBank Toolkit

The EuroTermBank Federated Network consists of independent Federated Nodes of national, regional, or even organisational scope. These nodes are comprised

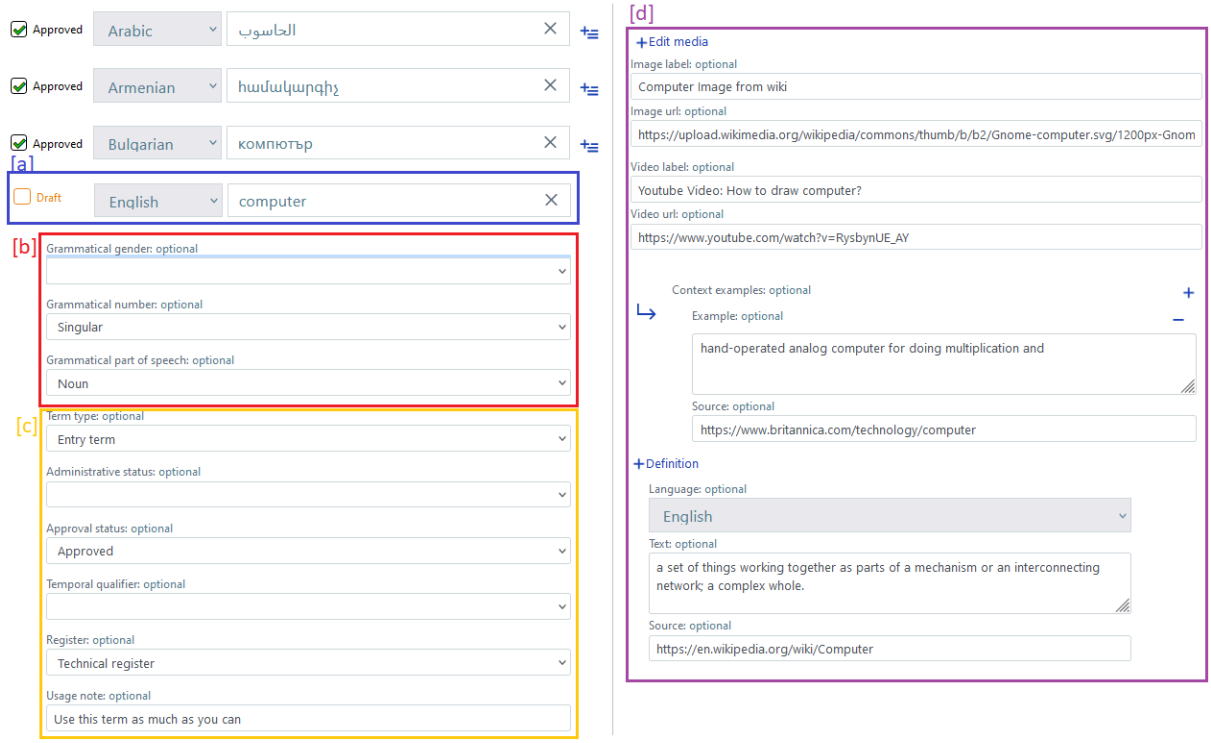

Figure 5: Term entry edit view of the ETBT. The example shows the English language side of a multilingual term.

of institutions that independently identify and coin terminology and administer it to share the resulting data with the pan-European terminology repository—EuroTermBank. EuroTermBank aggregates and publishes the terminology data to make it accessible for stakeholders in Europe and beyond. Figure 2 gives a conceptual view of the EuroTermBank Federated Network.

The ETBT plays a vital role in terminology data sharing both locally and globally because most terminology work is carried out predominantly in a local setting. The ETBT facilitates standardisation and streamlining of terminology curation by offering readily available tools and infrastructure. For example, the ETBT is based on common standards in terminology management and sharing, such as ISO 12620 on data categories (ISO, 2019a), ISO 26162 on terminology databases (ISO, 2019b), and the TermBase eXchange (TBX) 2 standard (ISO, 2008). The application of standards-based tools reduces the cost and effort of terminology curation and guarantees that the resulting terminology collections are mutually compatible, thus ensuring ease of sharing. Compatibility with the same shared standards as assured by the ETBT also enables conformity with a machine-readable data structure—an often overlooked quality for terminology, which nevertheless is paramount for terminology integration in machine translation (Bergmanis and Pinnis, 2021a).

Likewise, the EuroTermBank Toolkit ensures the currentness of the terminological data available at EuroTermBank by synchronising it with EuroTermBank federated nodes of organisations and institutions throughout Europe.

3. Functionality

Search Figure 3 demonstrates the term search view of the ETBT. Terms can be searched for in the entire local database or a specific term collection or set of collections. Likewise, collections and search results can be filtered by domain and language.

Terminology management Terminology data can be added in two principal ways: 1) by creating a new term entry (as well as editing an existing one) and 2) by importing an existing collection. Terminological information for new term entries is added by following the TBX 2 format. Besides basic data categories, such as subject field, term equivalents in different languages, definitions, and examples of how the term is used in context, information about the term's morphological properties (e.g., grammatical part of speech, number, and gender) (Figure 5 b), various administrative information and usage metadata (e.g., register, type, currentness) (Figure 5 c), media – images and videos (Figure 5 d) – and other categories can be added to provide extensive information about the term. Likewise, the same information can be added for the corresponding terms in other languages, thus making the terminology collection multilingual.

Unless *approved*, the term is saved as a *draft* (Figure 5 a), in which case it is visible only to the members of the current group and is not published. The import functionality supports CSV, TBX, and Excel file for-

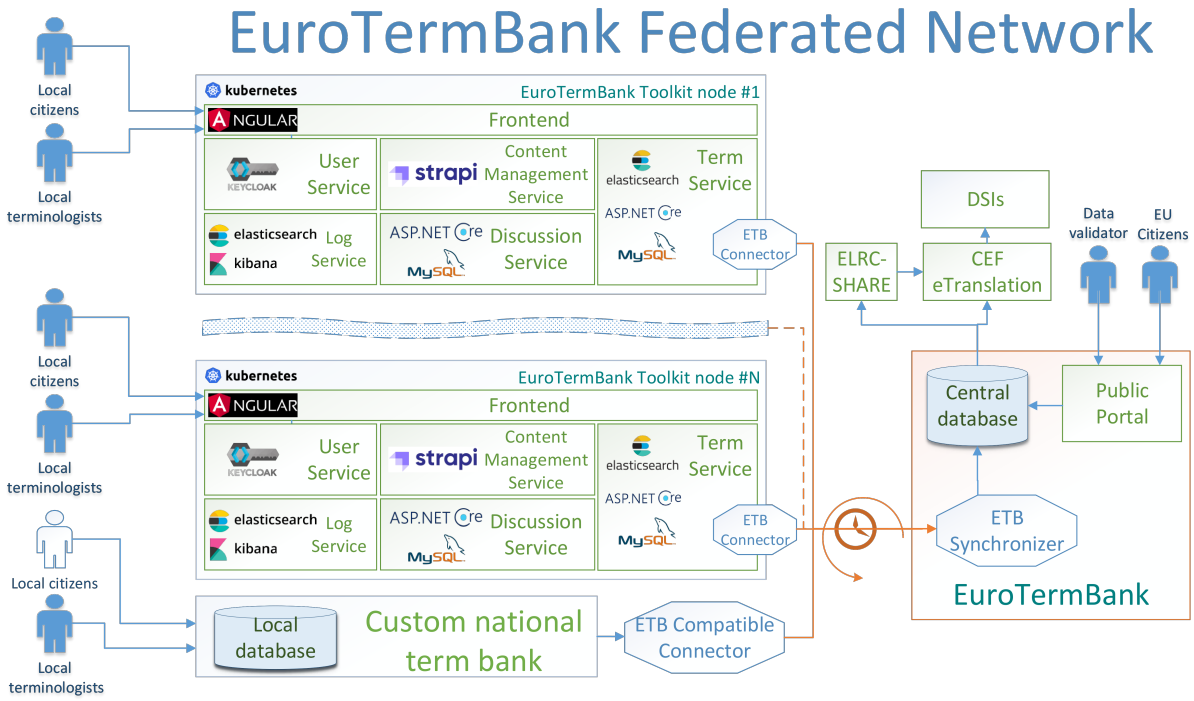

Figure 6: The architecture of the ETBT within the framework of the EuroTermBank Federated Network.

mats allowing to reuse already pre-existing terminology data.

Terminology sharing Term collections can be shared within a user group by adding new collaborators, or they can be exported to CSV, TBX and Excel file formats. If a term collection is made public, it is made accessible to the members of the general public through EuroTermBank.

Collaboration Users can share the term candidates with collaborators, participate in discussions about the concepts and term candidates (see Figure 4), and approve term candidates and new entries.

4. Architecture

The ETBT is designed using the microservices architecture where each service can be deployed as a container using, e.g., the Kubernetes[4] container orchestration system. The architecture of the ETBT within the framework of the EuroTermBank Federated Network is depicted in Figure 6. The ETBT consists of six components:

- A **frontend application**, which provides a graphical user interface for end-users and is developed as a single page application using Angular[5].
- A headless (i.e., without a graphical user interface, but with an application programming interface (API)) **content management system** (CMS) that stores static content for the frontend application. For the CMS, we use the Strapi[6] headless CMS.
- A **user service**, which handles user management, authentication and authorisation. For the user service, we use the Keycloak[7] identity and access management solution.
- A **discussion service**, which provides functionality for terminologists to discuss individual term entries and to enable involvement in terminology work. The discussion service is built as an ASP.NET Core[8] web service with an underlying MySQL[9] database.
- A **log service** that allows to store and visualise log data. The log service utilises the Elastic Search[10] engine for data storage and retrieval and Kibana[11] for visualisation of data that is stored by Elastic Search.
- A **term service** that provides all functionality necessary for terminology management (i.e., creation, editing, import, export, etc.), retrieval, and sharing. The term service is built as an ASP.NET Core web application with Elastic Search and an underlying MySQL database.

[4] https://kubernetes.io
[5] https://angular.io
[6] https://strapi.io
[7] https://www.keycloak.org
[8] https://docs.microsoft.com/en-us/aspnet/core/?view=aspnetcore-6.0
[9] https://www.mysql.com
[10] https://www.elastic.co
[11] https://www.elastic.co/kibana

All terminology specified as public and thus sharable is automatically synchronised with EuroTermBank. The synchronisation is performed by each federated node individually. Federated nodes push changes in public term collections to EuroTermBank's Central synchronisation API. All terminological data exchange is performed using the TBX 2 data format.

5. Current State of the EuroTermBank Federated Network

EuroTermBank is currently the largest centralised online terminology bank in Europe, providing access to more than 14.5 million terms from 463 collections. The EuroTermBank Federated network consortium currently consists of eight members – four of which use a customised EuroTermBank Toolkit solution, while the other four have established a synchronisation proxy with the EuroTermBank database exchanging information with the network. These eight members represent a total of eight countries (Austria, Denmark, Estonia, Iceland, Latvia, Lithuania, Slovenia, and Sweden), which comprise academia and industry leaders in terminology and language technologies. The network's future goals are to have at least one network member in each member state of the European Union.

6. Conclusion

We presented the EuroTermBank Toolkit, an open terminology management toolkit for the EuroTermBank Federated Network. The toolkit addresses the problem of outdated terminology data in shared terminology repositories by providing a standards-based infrastructure for terminology management and sharing for organisations across Europe and beyond. The ETBT facilitates standardisation and streamlining of terminology curation by offering readily available tools and infrastructure for collaboration and data sharing. The common approach enabled by ETBT provides an easy to implement solution for any institution needing a standards-based tool for terminology management and data sharing. It also enables management of machine-readable data for machine translation systems and other NLP tools and facilitates data synchronisation with EuroTermBank – the largest multilingual terminology resource in Europe. The instructions for the deployment of the ETBT are publicly available at: `https://github.com/Eurotermbank/Federated-Network-Toolkit-deployment`.

Acknowledgements

The research leading to these results has received funding from the research project "Competence Centre of Information and Communication Technologies" of EU Structural funds, contract No. 1.2.1.1/18/A/003 signed between IT Competence Centre and Central Finance and Contracting Agency, Research No. 2.9. "Automated multilingual subtitling".

This work was partly done within the scope of eTranslation TermBank Project (Action: 2019-EU-IA-0049) which is co-financed by the European Union's Connecting Europe Facility.

7. Bibliographical References

Arcan, M., Turchi, M., Tonelli, S., and Buitelaar, P. (2014). Enhancing statistical machine translation with bilingual terminology in a cat environment. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, pages 54–68. Association for Machine Translation in the Americas.

Arcan, M., Turchi, M., Tonelli, S., and Buitelaar, P. (2017). Leveraging bilingual terminology to improve machine translation in a cat environment. *Natural Language Engineering*, 23(5):763–788.

Belinkov, Y. and Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Bergmanis, T. and Pinnis, M. (2021a). Dynamic terminology integration for COVID-19 and other emerging domains. In *Proceedings of the Sixth Conference on Machine Translation*, pages 821–827.

Bergmanis, T. and Pinnis, M. (2021b). Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111.

Cabré, M. T. (1999). *Terminology: Theory, methods, and applications*, volume 1. John Benjamins Publishing.

Cross Vila, L., Escolano Peinado, C., Rodríguez Fonollosa, J. A., and Ruiz Costa-Jussà, M. (2018). End-to-end speech translation with the transformer. In *IberSPEECH 2018, Barcelona, November 21-23: program and proceedings*, pages 60–63. Antonio Bonafonte, Jordi Luque and Francesc Alías Pujol.

de Gspert, A., Iglesias, G., Byrne, W., et al. (2018). Neural machine translation decoding with terminology constraints. Association for Computational Linguistics.

Di Gangi, M. A., Negri, M., and Turchi, M. (2019). Adapting transformer to end-to-end spoken language translation. In *INTERSPEECH 2019*, pages 1133–1137. International Speech Communication Association (ISCA).

Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068.

Exel, M., Buschbeck, B., Brandt, L., and Doneva, S. (2020). Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Por-

tugal, November. European Association for Machine Translation.

Gaido, M., Rodríguez, S., Negri, M., Bentivogli, L., and Turchi, M. (2021). Is "moby dick" a whale or a bird? named entities and terminology in speech translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1707–1716.

Galinski, C. (2007). New ideas on how to support terminology standardization projects. *eDITion*, 1:2007.

Gornostay, T. (2010). Terminology management in real use. In *Proceedings of the 5th International Conference Applied Linguistics in Science and Education*, pages 25–26.

Henriksen, L., Povlsen, C., and Vasiļjevs, A. (2006). EuroTermBank-a terminology resource based on best practice. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

ISO, (2008). *Systems to Manage Terminology, Knowledge and Content–TermBase eXchange (TBX)*.

ISO, (2019a). *ISO 12620, Management of terminology resources — Data category specifications*.

ISO, (2019b). *ISO 26162-2 Management of terminology resources — Terminology databases*.

Johnson, I. and Macphail, A. (2000). IATE-interagency terminology exchange: development of a single central terminology database for the institutions and agencies of the european union. In *Workshop on Terminology resources and computation*.

Jon, J., Aires, J. P., Varis, D., and Bojar, O. (2021). End-to-end lexically constrained machine translation for morphologically rich languages. *CoRR*, abs/2106.12398.

Khayrallah, H. and Koehn, P. (2018). On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83.

Matusov, E., Wilken, P., and Georgakopoulou, Y. (2019). Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93.

Pinnis, M. (2015). *Terminoloģijas integrācija statistiskajā mašīntulkošanā*. Ph.D. thesis, University of Latvia.

Schioppa, A., Vilar, D., Sokolov, A., and Filippova, K. (2021). Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Sheth, A. P. and Larson, J. A. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys (CSUR)*, 22(3):183–236.

Siekmeier, A., Lee, W., Kwon, H., and Lee, J.-H. (2021). Tag assisted neural machine translation of film subtitles. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 255–262, Bangkok, Thailand (online), August. Association for Computational Linguistics.

Vasiljevs, A. and Borzovs, J. (2006). Terminology standards in the aspect of harmonization for international term database. *Terminologija*, 13:17.

Vasiljevs, A. and Rirdance, S. (2007). Consolidation and unification of dispersed multilingual terminology data. In *International Conference RANLP 2007 (Recent Advances in Natural Language Processing)*, pages 614–618.

Vasiljevs, A., Rirdance, S., and Liedskalnins, A. (2008). EuroTermBank: Towards greater interoperability of dispersed multilingual terminology data. In *Proceedings of the First International Conference on Global Interoperability for Language Resources ICGL*, pages 213–220.

Verplaetse, H. and Lambrechts, A. (2019). Surveying the use of cat tools, terminology management systems and corpora among professional translators: general state of the art and adoption of corpus support by translator profile. *Parallèles*, 31(2):3–31.

Vydana, H. K., Karafiát, M., Zmolikova, K., Burget, L., and Černockỳ, H. (2021). Jointly trained transformers models for spoken language translation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7513–7517. IEEE.

Wang, S., Tan, Z., and Liu, Y. (2022). Integrating vectorized lexical constraints for neural machine translation.