Proceedings of the First workshop on Language Technologies for Government and Public Admnistration (LT4Gov), pages 28–31
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

The Austrian Language Resource Portal for the Use and Provision of Language Resources in a Language Variety by Public Administration – a Showcase for Collaboration between Public Administration and a University

Barbara Heinisch, Vesna Lušicky
Centre for Translation Studies, University of Vienna, Austria
Gymnasiumstraße 50, 1190 Vienna
{barbara.heinisch, vesna.lusicky}@univie.ac.at

Abstract

The Austrian Language Resource Portal (*Sprachressourcenportal Österreichs*) is Austria's central platform for language resources in the area of public administration. It focuses on language resources in the Austrian variety of the German language. As a product of the cooperation between a public administration body and a university, the Portal contains various language resources (terminological resources in the public administration domain, a language guide, named entities based on open public data, translation memories, etc.). German is a pluricentric language that considerably varies in the domain of public administration due to different public administration systems. Therefore, the Austrian Language Resource Portal stresses the importance of language resources specific to a language variety, thus paving the way for the re-use of variety-specific language data for human language technology, such as machine translation training, for the Austrian standard variety.

Keywords: Language varieties, language resource collection, public administration

1. Introduction

The Austrian Language Resource Portal (*Sprachressourcenportal Österreichs*, sprachressourcen.at) is Austria's central platform for language resources in the area of public administration. It focuses on language resources in the Austrian variety of the German language and provides language aids to enable communication about Austrian public administration on a national and European level in English. The Austrian Language Resource Portal aims at offering a sustainable and extendable platform for the provision of both human-readable and machine-readable language resources for multilingual communication on public administration, and language technologies tailored to the Austrian variety of the German language, such as machine translation engines. The main target groups consist of public administration staff, translators, interpreters and the public. The Portal is the tangible result of a long-term cooperation between the Austrian Armed Forces Language Institute as part of the Austrian public administration and the Centre for Translation Studies of the University of Vienna. The language resources that are made available via the Portal are the product of a cooperation between governmental translators and terminologists and scholars from the field of translation and terminology studies.

Before elaborating on the contents of the Austrian Language Resource Portal, it is important to illustrate the significance of an individual language resource portal for the Austrian variety of the German language.

2. The Austrian Variety of the German Language

German is considered to be a pluricentric language, "i.e. a language with several interacting centers, each providing a national variety with at least some of its own (codified) norms" (Clyne, 1995: 20). In the localization industry, there is a similar notion of a locale, i.e. a group of characteristics, information or rules related to linguistic, cultural, domain-specific and geographic conventions in a target group (DIN ISO 18587:2018). These conventions differ between the locales.

2.1 German as a Pluricentric Language

German as a pluricentric language has three standard varieties (Schmidlin, 2011), with Austrian German being one of the three codified standard forms (German, Austrian, Swiss) and, therefore, a diatopic variety.

German is the (co-)official language in seven countries or parts of European countries (Austria, Belgium, Germany, Italy, Liechtenstein, Luxembourg, and Switzerland). The Austrian variety of the German language differs in several respects from other varieties of German (Wiesinger, 1988), whereas lexical differences are the most pronounced and obvious ones. In several cases, words have the same meaning in Germany and Austria, but in Austria these words have an additional meaning as well. If lexical items are unique to Austria, they are called Austriacisms. Other differences between the German and Austrian standard varieties, in addition to the lexical ones, include pronunciation, the grammatical gender of nouns (e.g. 'the yoghurt', which has a masculine gender in the German variety "*der Joghurt*" and a neuter gender in the Austrian (and Swiss) variety "*das Joghurt*") or the use of tenses or prepositions (Wiesinger, 1996).

Before Austria joined the European Union, language and language identity were at the center of a public debate. This resulted in an additional document to Austria's accession treaty (Protocol no.10) that lists 23 Austrian expressions for food (e.g. *Topfen*, *Marille*) that must be used in the EU legislation. Although the goal of this list seems to be mainly to address the concerns of the population before joining the EU (de Cillia & Wodak, 2002), it is still significant as it is the only EU contract document granting a special status to a language variety in the EU.

2.2 Language-variety-specific Terminology

Terminology may also be different between language varieties. This may cause misunderstandings due to diverging concepts of a term. Examples for the German language are, among others, food terminology (Schmidlin 2011) and legal and administrative terminology (Wissik

2013; Lohaus 2000). The diversity in administrative and legal terminology arises from different legal systems (de Groot, 1999). Therefore, the (terminological) difference between the Austrian and German language variety of German is more than a word list, i.e. the list of Austriacisms, which was demanded by Austria during its accession to the European Union (Schreiber, 2002; Markhardt 2002; EU 1995). Major misunderstandings may, of course, originate from terminology, but these two standard varieties also differ with regard to syntax, grammar, morphology, etc. (Wiesinger, 1996).

In the domain of public administration, misunderstandings between speakers of different German varieties may arise since the related terms refer to different administration systems. These include terms such as *Magistrat*, *Bezirk*, *Landeshauptmann* or *Landeshauptfrau*, *Bezirksinspektor* (Heinisch, 2020). The terms *Landeshauptmann* or *Landeshauptfrau* are used in Austria (and some parts of Italy), but are not used in Germany or Switzerland to refer to the head of a *Bundesland* (provincial government). Another example is the German translation of the term *district*. If it is translated as *Kreis*, it rather refers to the German administrative subdivision. Misunderstandings may arise if (Austrian) readers are not aware of the meaning of *Kreis* since they may be rather used to *Bezirk*.

3. The Portal

The Austrian Language Resource Portal understands itself as a user-oriented catalog for the Austrian variety of the German language in the public administration domain. It does not replace existing language resource repositories but presents language resources (LRs) and language technologies (LTs) specific to the Austrian variety of the German language. Here, the usability for the target group of public administration staff, translators, interpreters and terminologists and the aim of increasing the visibility of the Austrian language variety are key. Moreover, it presents the results of various LR and LT projects related to the Austrian German variety on one portal.

The primary users are specialized translators and staff working with the Austrian public administration who communicate in German, and (occasionally) also in English. In addition, the Portal caters to LT developers and natural language processing (NLP) researchers in need for terminological datasets in this domain.

The Austrian Language Resource Portal contains the following language resources and technologies in the area of Austrian public administration and related information:

3.1 Public Administration Terminology

Terminology is an important language resource. Therefore, a crucial component of the Austrian Language Resource Portal is a bilingual terminological database containing terminology used in public administration in Austria.

The terminological resource entitled *Fachglossar Österreichische Verwaltung. Deutsch – Englisch* (glossary of public administration) covers terminology in this domain in German and English. It contains terminology from the areas of Austrian public law, legislation and executive authorities. It is aimed at providing a terminological resource targeted at language professionals, such as translators and interpreters. Since it is tailored to the peculiarities of the Austrian public administration system, the bilingual terminological resource is aimed at offering internationally comprehensible and transparent English terms since there is hardly equivalence of concepts.

The terminology is standardized by an informal working group of translators and terminologists employed with the Austrian federal ministries (*Arbeitsgruppe Gouvernementaler Übersetzungs- und Terminologiedienste*, ARG GUT). These governmental translators and terminologists joined forces to exchange experiences and create language resources under the aegis of the Austrian Armed Forces Language Institute, such as the glossary of public administration, which is available in different human-readable and machine-readable formats, such as .pdf, .csv or .tbx. The work on the terminological resource also revealed inter-ministerial terminological differences such as those related to internal divisions and subdivisions. Therefore, a prescriptive approach was adopted. This bilingual glossary contains 696 entries covering administrative bodies and institutions in Austria, as well as administrative procedures and processes.

3.1.1 Collection of Variety-specific Language Resources for Machine Translation Training

A collection of language resources in the public administration domain for the Austrian variety of German on the Portal stems from the EU Council Presidency Translator project. The EU Council Presidency Translator is a neural machine translation (NMT) system developed, among others, for the trio presidency in 2017 and 2018, i.e. the EU Council Presidencies of Estonia (translate2017.eu), Bulgaria and Austria (translate2018.eu).

For the Austrian Council Presidency in 2018, the system was geared towards texts related to the Presidency domains, thereby specializing in the language directions German-English and English-German. The neural machine translation system was targeted at EU Council Presidency staff, journalists, translators, delegates and visitors. For the Austrian Presidency, the objective was the creation of customized machine translation (MT) engines for the English and Austrian German language pair with a focus on domains and text types related to the work program of the EU Council Presidency.

For the training of the EU Council Presidency Translator for Austria, the following categories of data were collected and are available through the Portal:

1. Austria-related named entities (names of municipalities, names of politicians, common first names and last names of people, etc.) were collected and compiled from Open Data (common names and geographical names), Wikimedia (names of stock companies) and manually compiled (names of politicians, Austrian newspapers, etc.) (15,000 named entities).

The Austrian Open Data Portal (www.data.gv.at) proved to be a useful resource containing data such as named entities (municipalities, regions, common first and last names, etc.). Although the Austrian Open Data Portal listed a rather large amount of language resources, several of these language resources required further processing due to the file formats used, e.g. PDF.

2. German-English parallel data containing news and statements (press releases, interviews and Common Foreign and Security Policy statements) in German and English by the Presidency of the Council of the EU held by Austria in 2006, aligned with HunAlign, a language-independent sentence aligner (Varga et al, 2005) and

manually evaluated by two evaluators (4,973 translation units in .tmx format).
3. Austria-related named entities and terminology related to the topics of the trio presidency was created by the University of Vienna by crawling, extracting and compiling content from Wikipedia (71,000 terms, .txt format).
4. Additionally, a German-English terminological database of the core Austrian administrative terminology outlined in 3.1 was used for the purpose of MT training (1,400 terms).
When collecting these Austrian-German-specific LRs, the major obstacles were not the lack of relevant data, but the rather restrictive usage rights and legal uncertainties related to crawling, collecting, sharing and using the language data. For this reason, the following data are not part of the Portal, although collected and deemed useful for MT training and other NLP applications:
For the news domain covering Austria, two monolingual corpora were compiled. The Austrian German news corpus (2.3 M segments) was compiled from news gathered by focused crawling of all major Austrian daily broadsheets, and press releases from major news and media outlets produced in Austria. For the English news, several sub-corpora were compiled (2.3 M segments) by focused crawling of news and media platforms in Austria, in Germany and at the European level on the topic "Austria".
For the EU Council Presidency domain and its main topics, monolingual and parallel corpora were compiled. A monolingual Austrian German corpus was compiled by focused crawling of websites of relevant public entities. The corpus for training the EU Council Presidency Translator contained a large amount of texts produced in Austria. This is in contrast to eTranslation, the European Commission's machine translation system (for German), which is mainly trained with LRs from Germany. This is also illustrated by the large amount of LRs provided by Germany in ELRC-SHARE (elrc-share.eu/repository), which makes accessible openly licensed LRs that are used for the training of the eTranslation engines.
Our data collection efforts showed that there is a certain number of language resources, which are available to be implemented in language technology applications available for the Austrian German variety in the public administration domain. These resources had to be pre-processed in order to make them suitable for further use.

3.2 Language Guide

The Austrian Language Resource Portal also offers a language guide that provides basic information on communication in English. It contains tips, e.g. for having small talk, chairing a meeting, writing an e-mail, ordering food in the restaurant, explaining culinary specialties or identifying false friends. Moreover, it provides information on avoiding pitfalls in intercultural communication. It is targeted at people who are not used to speak English. It proved to be a useful aid in language learning contexts.

3.3 Compilation of other Language Resources

Finally, the Austrian Language Resource Portal contains a compilation of other relevant language resources and repositories that were not created by the cooperation partners as part of the Portal. Thus, links to external terminological databases that are especially relevant for translators and interpreters are provided.
An overview of all LRs on the Austrian Language Resource Portal is available on the website: sprachressourcen.at.

4. Discussion and Conclusion

The Austrian Language Resource Portal reflects the need for a central platform to access and exchange language resources that are specific to the Austrian variety of the German language. It provides a first attempt to not only highlight the value of language resources and make them freely available, but also show the impact of language resource sharing (Heinisch, 2018). Nevertheless, the amount of language resources on this Portal is not a comprehensive and exhaustive one. This may be due to the fact that the Austrian Language Resource Portal is primarily aimed at the target group of language professionals, such as translators. This is also the reason why the language resources are made available in different formats with giving preference to human-readable formats over machine-readable ones.
Additionally, major obstacles for delivering language resources to the Portal are confidentiality and security issues, legal uncertainty, i.e. the question of whether translators are allowed to share data due to IPR or copyright issues as well as the organizational framework that hinders the delivery of language resources. This shows that in Austria, similar to the situation all over Europe, there is still a lack of awareness for the value of language data (Heinisch and Kotzian, 2018; European Language Resource Coordination, 2019). Nevertheless, these language resources would be especially important for training human language technology, including machine translation systems, with a lower-resourced language variety to achieve quality improvements in language technology output.
The availability of and access to language resources in a language variety, such as Austrian German, may improve NLP and language technology applications (e.g. NMT training and thus increase the quality of NMT output), to avoid the deprivation of language diversity (within a language). This is particularly important in the light of the European Parliament resolution on language equality in the digital age, which recognizes that some (smaller) languages are threatened by digital extinction (EP, 2018). The resolution also states that language technologies can overcome language barriers and facilitate communication in a multilingual European Digital Single Market. Furthermore, the language equality resolution recommends that the member states of the EU define a minimum number of language resources for each European language to counteract digital extinction. These language resources may include lexicons, annotated corpora, speech records, translation memories and encyclopedic content (ibid.). The Austrian Language Resource Portal aims at contributing to this objective by increasing the availability of and access to language-variety-specific language resources.
In this respect, the Portal stresses the significance of differentiating between varieties of German and thus primarily caters for the Austrian German variety. This demonstrates that especially languages for specific purposes may differ significantly between the different standard varieties of the German language, as exemplified by the terminology used in public administration.
Although the Austrian Language Resource Portal focuses on Austrian German it does not mention other Austrian

language varieties, such as dialects, which play a crucial role in Austria. Thus, areas such as non-standard language, e.g. dialects (in machine translation) (Neubarth & Trost, 2017) would require further investigation.

To sum up, the Austrian Language Resource Portal stresses the importance of language resources specific to a language variety, thus paving the way for the re-use of variety-specific language data for human language technology, such as MT training, for the Austrian standard variety.

5. Acknowledgements

This work has been partly funded by the European Union's Connecting Europe Facility under grant agreement no. INEA/CEF/ICT/A2016/1297953. We would also like to thank Jürgen Kotzian from the Austrian Armed Forces Language Institute, who was spearheading the creation of the Austrian Language Resource Portal and Bartholomäus Wloka from the Centre for Translation Studies for his support in LR collection.

6. Bibliographical References

Clyne, M. G. (1995). The German language in a changing Europe. Cambridge: Cambridge University Press.

de Cilia, R. and Wodak, R. (2002). Zwischen Monolingualität und Mehrsprachigkeit. Zur Geschichte der österreichischen Sprachenpolitik. In H. Barkowski & R. Faistauer (Eds.), *in Sachen Deutsch als Fremdsprache. Sprachenpolitik und Mehrsprachigkeit, Unterricht, Interkulturelle Begegnung*. Hohengehren: Schneider Verlag, pp. 12-27.

de Groot, G.-R. (1999). Zweisprachige juristische Wörterbücher. In P. Sandrini (Ed.), *Übersetzen von Rechtstexten. Fachkommunikation im Spannungsfeld zwischen Rechtsordnung und Sprache*. Tübingen: Narr, pp. 203–227.

DIN ISO 18587:2018-02 Übersetzungsdienstleistungen_-Posteditieren maschinell erstellter Übersetzungen - Anforderungen. Berlin: Beuth Verlag GmbH.

European Language Resource Coordination (2019). ELRC White Paper. Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe. Why Language Data Matters. http://www.lr-coordination.eu/sites/default/files/ELRC_Conference/ELRCWhitePaper.pdf.

EP (2018). European Parliament resolution of 11 September 2018 on language equality in the digital age. http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.html.

Heinisch, B. (2018). Dissemination of administrative terminology on Austria's language resource portal as a means of quality assurance. Poster presentation at the EAFT Terminology Summit 2019, San Sebastian, Spain.

Heinisch, B. (2020). Sprachvarietätenabhängige Terminologie in der neuronalen maschinellen Übersetzung: Eine Analyse in der Sprachrichtung Englisch-Deutsch mit Schwerpunkt auf der österreichischen Varietät der deutschen Sprache. In C. Schöch (Ed.), DHd 2020 Spielräume: *Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*, pp. 211–214.

Heinisch, B. and Kotzian, J. (2018). ELRC Workshop Report for Austria: Deliverable D3.2.10. http://lr-coordination.eu/sites/default/files/Austria/2018/ELRC_Workshop_Austria_Report_public_v1_FINAL.PDF.

Lohaus, M. (2000). Recht und Sprache in Österreich und Deutschland: Gemeinsamkeiten und Verschiedenheiten als Folge geschichtlicher Entwicklungen; Untersuchung zur juristischen Fachterminologie in Österreich und Deutschland. Giessen, Fachverl. Köhler.

Markhardt, H. (2002). Das österreichische Deutsch im Rahmen der Europäischen Union : das "Protokoll Nr. 10 über die Verwendung österreichischer Ausdrücke der deutschen Sprache" zum österreichischen EU-Beitrittsvertrag und die Folgen: eine empirische Studie zum österreichischen Deutsch in der EU, Univ. Wien.

Neubarth, F. and Trost, H. (2017). Statistische maschinelle Übersetzung vom Standarddeutschen in den Wiener Dialekt. In C. Resch & W.U. Dressler (Eds.), *Digitale Methoden der Korpusforschung.* Wien, Verlag der österreichischen Akademie der Wissenschaften, pp. 179–203.

Schmidlin, R. (2011). Die Vielfalt des Deutschen: Standard und Variation: Gebrauch, Einschätzung und Kodifizierung einer plurizentrischen Sprache. Studia linguistica Germanica. Berlin: De Gruyter.

Schreiber, M. (2002). Austriazismen in der EU: (k)ein Übersetzungsproblem? *Lebende Sprachen*, 47(4). https://doi.org/10.1515/les.2002.47.4.150

Varga Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh & Viktor Trón (2005). Parallel corpora for medium density languages. Proceedings of RANLP'2005. Borovets, Bulgaria, pp. 590-596.

Wiesinger, P. (1996). Das österreichische Deutsch als eine Varietät der deutschen Sprache. *Die Unterrichtspraxis / Teaching German*, 29, 1996: 10.2307/3531825.

Wiesinger, P., Ed. (1988). Das österreichische Deutsch. Wien, Köln, Graz, Böhlau.

Wissik, T. (2013). Terminologische Variation in der Rechts- und Verwaltungssprache: eine korpusbasierte Analyse der Hochschulterminologie in den Standardvarietäten des Deutschen in Deutschland, Österreich und der Schweiz. Dissertation, Universität Wien.

7. Language Resource References

ARG GUT (2018). Fachglossar Österreichische Verwaltung. Deutsch – Englisch, distributed via Sprachressourcenportal Österreichs, https://www.sprachressourcen.at/verwaltungsglossar/

University of Vienna (2018). Austrian named entities, distributed via ELRC-SHARE, https://www.elrc-share.eu/repository/browse/austrian-named-entities/b0998b12ab9611e8b7d400155d02670612bad73492934202887a45e227312e0e/

University of Vienna (2018). Glossary terms in German related to Austria and the topics of the trio presidency, distributed via ELRC-SHARE, https://www.elrc-share.eu/repository/browse/terms-in-german-related-to-austria-and-the-topics-of-the-trio-presidency/b82781c4ab9e11e8b7d400155d026706f61ef02809fb4748944b1af1b434f0a9/

University of Vienna (2018). German-English parallel data by the Presidency of the Council of the EU held by Austria in 2006, distributed via ELRC-SHARE, https://www.elrc-share.eu/repository/browse/german-english-parallel-data-by-the-presidency-of-the-council-of-the-eu-held-by-austria-in-2006/e38b283eac3e11e8b7d400155d0267062180d233a0fd4e84b8dffb9b25cc1775/