ON THE DEVELOPMENT OF AN INTERDISCIPLINARY ANNOTATION AND CLASSIFICATION SYSTEM FOR LANGUAGE VARIETIES – CHALLENGES AND SOLUTIONS

AGNES KIM[1] – LUDWIG M. BREUER[2]
[1] Department of Slavonic studies, University of Vienna, Austria
[2] Department of German studies, University of Vienna, Austria

KIM, Agnes – BREUER, Ludwig M.: On the Development of an Interdisciplinary Annotation and Classification System for Language Varieties – Challenges and Solutions. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 191 – 207.

Abstract: The Special Research Programme (SFB) 'German in Austria: Variation – Contact – Perception' is a project financed by the Austrian Science Fund (FWF F60). Its nine project parts are collaboratively conducting research on the variation and change of the German language in Austria. The SFB explores the use and the subjective perception of the German language in Austria as well as its contact with other languages. Methodologically and theoretically, most SFB project parts are situated within variationist linguistics, others in contact linguistics and perceptionist linguistics. This paper gives an insight into the conception of a framework for the annotation and ultimately also classification of language varieties, which is being developed within the SFB. It outlines the requirements of the various project parts and reviews, whether and how standardised language codes (ISO 639) and language tags (following BCP 47) can be utilised for the annotation of language varieties in variationist linguistic projects.

Keywords: language varieties, dialects, language tags

1 INTRODUCTION AND CONTENT

The Special Research Programme (SFB) "German in Austria: Variation – Contact – Perception", funded by the Austrian Science Fund (FWF) is an interdisciplinary collaborative project, which conducts research on the variation and change of the German language in Austria. It consists of the three thematic pillars represented in its title, and thus explores the entire spectrum of language variation in German in Austria, the perception of German in Austria, and contact of German in Austria with other languages.

Six of the nine project parts are located at different departments of German linguistics at the Universities of Vienna, Salzburg, and Graz, as well as at the Austrian Academy of Sciences. Of the remaining, two project parts—those focusing on aspects of language contact—are situated at the Department of Slavonic studies at the University of Vienna. Additionally, the Centre of Translations Studies at the University of Vienna hosts the project part responsible for developing and implementing the Collaborative Online Research Platform "German in Austria". This platform will support the working process in the whole research cycle ranging from data querying, input, annotation and analysis to interactive online tools, which allow accessing the data in multiple ways. Furthermore, it will guarantee sustainable preservation of research both data and outcomes.

DE GRUYTER OPEN
DOI 10.1515/jazcas-2017-0029

Hence, the SFB aims at using and (if needed) enhancing existing (and standardised) annotation systems. In some cases (e.g., the tagging of specific syntactic phenomena), a completely new annotation scheme is necessary. Considering all possible annotation levels, the interdisciplinary orientation and the various (theoretical and empirical) approaches of the different project parts, a multidimensional and highly flexible annotation system is crucial to reconcile all these demands. Therefore, the SFB builds on a highly flexible annotation syntax with an underlying, equally strict description scheme.

In this paper, we focus on an annotation and classification framework of language varieties, which is supposed to serve as a basic description level for the language data gathered. Nonetheless, it should be kept in mind that it is a part of a larger annotation scheme, which also describes specific parts of the object language (e.g. part-of-speech-tagging).

In section 2 of this paper, we discuss several aspects of languages and their varieties that have to be considered in creating a custom-made annotation and classification framework of language varieties within the SFB. We discuss the different theoretical and methodological approaches of several project parts in order to outline their requirements of such a framework.

Section 3 discusses language codes and language tags with regard to their standardisation. We evaluate the benefits and difficulties of applying these standards within the SFB. Finally, section 4 proposes an according solution.

This paper gives a glimpse into the considerations of the working group responsible for designing and implementing a variety annotation and classification framework within the SFB.

2 REQUIREMENTS WITHIN THE SFB[1]

2.1 Task Cluster B: Variation

The three project parts within Task Cluster B focus on the dynamics of varieties of German in Austria in their linguistic and social structures. Methodologically and theoretically, they are situated within variationist linguistics. In order to answer their research questions, they collect language data of a large number of informants from all over Austria, from rural as well as urban areas. The elicited corpus will ultimately not only cover dialects from all over Austria but also other colloquial registers between (intended) dialects and the (intended) standard language (for the terminology see [1]). In addition to the data collected by the project parts themselves, comparative language

[1] A large-scale project like the SFB combines many aspects, which may be interesting for different kinds of linguistic projects. Thus, we would preferably describe the whole SFB in detail, considering theoretical and empirical approaches of its nine different project parts as well as the Collaborative Research Platform, which combines the project-internally orientated working infrastructure as well as – externally-orientated – means of dissemination and even elements of citizen science. However, this would go beyond the scope of this paper. Considering the main topic of this paper – the annotation and classification of languages and their varieties –, we focus on Task Cluster C. Since this Task Cluster does not only investigate German varieties, but conducts research into the contact of German in Austria with Slavic languages, too, it requires the whole scale of the presented annotation system. For closer information on the SFB in general, its goals and structure please refer to its homepage: `http://www.dioe.at/en`.

data such as linguistic descriptions are considered in order to trace the development of dialects in Austria over the course of the 20th century until the present. Among others, one outcome of this Task Cluster will be an online "speaking linguistic atlas", in which audio samples are provided within an interactive geographic information system. Within the whole research platform, all data will be linked and interconnected to other data. In terms of (automatic) linking or filtering of these data, a standardised metadata set including normalised variety classifications is necessary [2].

2.2 Task Cluster C: Contact

Task Cluster C is concerned with the contact of German in Austria with other, particularly Slavic languages. It is orientated towards contact linguistics and research into multilingualism and links German with Slavic linguistics in an interdisciplinary fashion. In the first four years, both project parts within that Task Cluster employ a diachronic approach. Therefore, they deal with data types that are clearly distinct from the data elicited by the synchronically orientated project parts [3].

Project Part 5 analyses the context of language policies in several fields of action and tension, i.e., in administration, law, and especially in education. For this purpose, existing data from legislative texts, newspapers, archive materials, and other contemporary documents are being collected, connected, and analysed by methods from the critical discourse analysis. The Project Part's central aim is to reconstruct the functional and metalinguistic dimensions of German in the multilingual Habsburg state and relate them to the conditions of language (and multilingualism) policies and planning in the Second Republic of Austria [4].

On the other hand, Project Part 6 focuses on linguistic contact phenomena, e.g. on all linguistic levels of German in Austria that have been explained by language contact with the Slavic languages. Its main goal is to give a comprehensive overview of alleged contact phenomena, provide a basic assessment of these and thus identify language myths associated with the contact with Slavic languages. The agglomeration of Vienna and its urbanolect will be of special interest. The Project Part particularly focuses on the exhaustive number of Slavic loanwords in German in Austria. For this purpose, linguistic and popular literature on the language contact phenomena in German in Austria [5] will be collected and processed.

Ultimately, Task Cluster C aims at establishing an *Information System on (historical) Multilingualism in Austria* (MiÖ) within the Collaborative Research Platform. This module will link and present quantitative data such as historical census data to qualitative data collected within the two project parts. It will further include a bibliography. The database shall make historical – and in a later stage – also present multilingualism in Austria and its linguistic, societal and historical conditions visible.

In order to ensure public searchability, we need to model and map historical and contemporary names and labels for languages and their varieties. These names' political and contextual restrictions, connotations, as well as their change over time have to be considered in the model. Some of these names have already been well described, such as the German *Tschechisch* 'Czech' and *Böhmisch* 'Bohemian', their relation and development [6].

2.3 Task Cluster D: Perception

Task Cluster D deals with language attitudes and language perception with special regard to German in Austria. Project Part 10, for example, investigates language attitudes and perception within schools, i.e., of pupils and teachers. Of course, in that context not only the so called internal multilingualism, i.e., competence in both dialect and standard language, but also external multilingualism has to be considered.

Project Part 8, on the other hand, compares how standard varieties as well as other registers of German are conceptualised by adults living all over Austria. Therefore, laymen's names for varieties and registers play a prominent role. Again, given the need to classify the gathered qualitative data, the described annotation framework is crucial for this Task Cluster, too. In the semi-standardised interviews, laymen are, e.g., asked how they would call the varieties they use. To ensure comparable quantitative analyses of the variety names expressed, a standardised categorisation is vital [7].

2.4 Task Cluster E: Collaborative Online Research Platform

Task Cluster E develops and implements the Collaborative Online Research Platform of the whole SFB. This is supposed to be the main communication and research hub, as well as the platform for the dissemination of data and results. Thus, the platform does not only play a role amongst and within the various Project Parts, but also connects and presents the SFB and its results to the outside world.

All implemented tools shall provide means of machine-to-machine communication and thus interfaces to share the data with other tools (from other projects etc.). Therefore, Task Cluster E aims at a high interoperability of annotation schemes, corpora and the annotated data itself. Apart from this, the emphasis on addressing the non-academic public as well makes a lucidly comprehensible and explicitly described annotation framework indispensable [8].

2.5 Summary: Summary and Outline of a Model

As shown above, the project parts focus on various aspects of varieties of German in Austria and their variation. Therefore, they take differing angles of view onto them: Task Cluster B mainly considers varieties as its objects of interest, i.e., as its object languages. Task Cluster D, on the other hand, focuses on names of varieties, i.e., *glottonyms* (or: *glossonyms*) and the concepts that speakers connect to them. Within Task Cluster C, both dimensions are relevant depending on the project part. In addition, due to the focus on language contact these projects require a model of genetic language and variety affiliation to ensure the searchability of the data during the working process and on the Research Platform. As Task Cluster E does not focus on the object level, but rather on the more technical and standardisation level it will be neglected in the following discussion.

Resulting from the requirements named above, we propose three modules within our variety annotation and classification system (see Fig. 1).

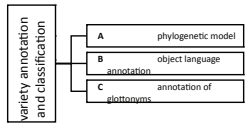


Fig. 1. Outline of a system for the annotation and classification of varieties

The modules have different functions within the working process and therefore differing statuses within the system: Module A provides a model for the phylogenetic affiliation of languages and their varieties. It serves as an auxiliary construction to ensure the searchability of the data. We therefore prefer a pragmatically orientated model within module A and consciously accept the simplifications that will have to be made in order to be able to model various language families and groups. Besides that, module A will be designed and provided as a closed model by the responsible working group. Of course, changes may be requested, but generally the module should not be changed on a regular basis during the working process.

Module B and C, on the other hand, are regarded to be working instruments, which can and should be adapted by the project parts according to their needs. The responsible working group provides the technical and content framework, as well as documentation and adaption guidelines. These modules serve as a comprehensive annotation framework for varieties and their names within the SFB. Possibly, we shall be able to develop classification systems for varieties of German in Austria, as well as for glottonyms based on the detailed analysis provided by several project parts at the end of the SFB.

Once more we will focus on the proposed modules, their content and technical modelling in section 4. However, we first will revise international systems and standards for language coding and language tagging. They are relevant, because one explicit goal of Task Cluster E in building the Online Research Platform is to develop best practise examples for handling variationist linguistic data. Especially this research area has recently led to a big amount of data. Consequently, the need for presenting these data online and connecting data from different sources has risen (cf. [9], [10]). Therefore, the development of best practise examples and new standardised annotation systems is vital for various variationist linguistic projects. Mutual interests lie in the connection of the different datasets gathered in different regions (*horizontal variation*), situations (*vertical variation*) and periods (*diachronic variation*). This requires a flexible, multidimensional, and thus complex but, given the large data sets, easy to use annotation system.

3 LANGUAGE CODES AND LANGUAGE TAGS

When it comes to identifiers of languages or language varieties, language codes need to be distinguished from language tags, even though the latter often refer and make use of the first. *Language codes* are alphabetical, numerical or alphanumerical identifiers, which uniquely refer to a certain language or language variety. The entities that the codes refer to are seen as rather well-defined and comparable according to the underlying language definition.

Language tags, on the other hand, allow for specifying deviations from default values of a given language in a certain text written or spoken in the according language. Therefore, they account for certain degrees of variation in language and are used like annotations rather than identifiers.

Below, we assess whether and how the most common standards and/or systems of language codes and tags may be used for linguistic projects in general and the SFB in special. In order to accomplish that task, we exemplarily compare whether and how

varieties of German in Austria as well as the genetic affiliation of Czech could be identified and modelled by utilising the according code sets. As shown above, both perspectives are needed for our system.

When it comes to language codes, we focus on the ISO 639 standards, because only language tags, which make use of them, can be used in XML annotations, such as those following the TEI standards (after the `xml:lang` attribute) [11]. Therefore, we leave other coding systems such as the *Glottocodes* [12], [13] and the *Linguasphere codes* [14] aside.

3.1 Language Codes: The ISO 639 Standard Family

As of 2017, the ISO 639 standard family comprises five sub-standards (see Tab. 1). Four of them define alpha-2 or alpha-3 codes for "the representation of names of languages". Part 4 sets the principles of coding and provides application guidelines. The ISO 639 standard family is under the responsibility of ISO/TC 37 (ISO Technical Committee 37), which generally facilitates the standardization "of principles, methods and applications relating to terminology and other language and content resources in the contexts of multilingual communication and cultural diversity" [15]. The various ISO 639 sub-standards are rooted in quite divergent disciplines and projects, as we show below.

		FIRST RELEASE	VALID VERSION	NUMBER OF SINGLE CODES[2]
ISO 639-1	*Alpha-2 code*	1967	2002	204
ISO 639-2	*Alpha-3 code*	1998	1998[3]	506
ISO 639-3	*Alpha-3 code for comprehensive coverage of languages*	2007	2007[4]	7459
ISO 639-4	*General principles of coding of the representation of names of languages and related entities, and application guidelines*	2010	2010	
ISO 639-5	*Alpha-3 code for language families and groups*	2008	2008	115[5]

Tab. 1. ISO 639 standard family

In 2009, the ISO published a proposal for ISO 639-6, an 'Alpha-4 code for comprehensive coverage of language variants'. This standard was withdrawn in November 2014 and is not available anymore [16]. According to various sources [17], [18], it was to be based on the *Linguasphere Register of the World's Languages and Speech Communitie*s [14].

[2] The given numbers represent the authors' count based on code of lists that were retrieved from the respective registration authorities' websites in March 2017: [20] (ISO 639-1 and ISO 639-2), [21] (ISO 639-3) and [22] (ISO 639-5).

[3] According to [20], the code list itself has been updated on 18 March 2014 for the last time. The last change is dated to 21 Nov 2012.

[4] The ISO 639-3 codes have had their latest update on 17 Feb 2017 [21], i.e., four days before the 20th edition of the *Ethnologue* was published on 21 Feb 2017.

[5] The last update of an ISO 639-5 element took place on 2 Nov 2013 [22].

Interestingly, in 2016 the same subcommittee that bears responsibility for ISO 639 (ISO/TC 37/SC 2 – *Terminographical and lexicographical working methods*) initiated a new project for the standardisation of the "Identification and description of language varieties" (ISO/AWI 21636) [19]. Unfortunately, there is no further information on this project publically available.

3.1.1 ISO 639-1 and ISO 639-2

Kamusella [17] provides an embedding of the emergence of the ISO 639-1 and ISO 639-2 standards into socio-cultural developments of the 20th century [17, p. 62ff.]. Generally, he associates the processes of standardisation or uniformisation with "modernity", i.e., "the [international] spread of various technologies and cultural practices" [17, p. 59], which also require shared terminologies.

The ISO 639-1 list of alpha-2 codes was compiled for a primary use in terminology, too. It is maintained by the *International Centre for Terminology* (Infoterm) in Austria[6] and includes identifiers for "the most developed languages of the world, having specialized vocabulary and terminology" [23]. Therefore, languages need to fulfil a list of detailed criteria in order to be assigned an ISO 639-1 code.

ISO 639-2, on the other hand, is primarily rooted in bibliography: As the alpha-2 codes proved insufficient to identify a large number of publication languages, the ISO developed an alpha-3 code set based on the *MARC Code List for Languages*, a standard created by the US Library of Congress [24]. This institution was also made the registration authority for the ISO 639-2 standard.

Both the ISO 639-1 and the ISO 639-2 substandard require languages to meet detailed criteria in order to be assigned an own code (see Table 2[7]). As ISO 639-1 is seen as a subset of ISO 639-2, a language needs to fulfil both the criteria for ISO 639-2 and the more specific ones for ISO 639-1 in order to be assigned an alpha-2 code. Table 2 can be read this way, as we first present the requirements for ISO 639-2 and only then the ones for ISO 639-1. Generally, a single language code is provided for languages which are written in multiple orthographies and scripts. Dialects should be represented by the "same language code as that used for the language". According to [25], the difference between a dialect and language is to be decided "on a case-by-case basis".[8][9]

[6] `http://infoterm.info`, retrieved 2017-03-30.

[7] The information presented in Table 2 were retrieved from [25]. Kamusella [17] presents the related list, too. In comparison to the list, we slightly regroup the information in order to enlarge comparability.

[8] Documents such as "specialized texts, such as college or university textbooks, technical documentation manuals, specialized journals, subject-field related books, etc." [25].

[9] "E.g. technical dictionaries, specialized glossaries, vocabularies, etc. in printed or electronic form" [25].

	ISO 639-2	ISO 639-1
DOCUMENTATION	• one agency holds **50 different documents** (not limited to text) in the language or • five agencies hold a total of 50 different documents in the language	• a significant body of existing documents[8] **written** in **specialized languages** • a number of existing **terminologies** in various subject fields[9]
RECOMMENDATION		• recommendation of a specialized authority[10] • support by one or more official bodies
NUMBER OF SPEAKERS		is considered
STATUS		recognized in one or more countries

Tab. 2. Requirements for ISO 639-1 and ISO 639-2

As can be seen from these requirements, both code sets have in common that the underlying language definition is a sociological one. ISO 639-1 basically provides codes for standard languages like `de` for *German* or `cs` for *Czech*. Neither aspects of language variation nor of their affiliation can be covered.

On the other hand, ISO 639-2 roughly lists what could be called *Ausbau* languages according to Kloss [26]. Therefore, in addition to `ger/deu`[11] for *(Standard) German*, there is an own code `gsw` for *Swiss German, Alemannic, Alsatic* (in German only: *Schweitzerdeutsch*).

In contrast to ISO 639-1, ISO 639-2 also provides the possibility to assign a collective alpha-3 code, if the requirements concerning the documentation of a language is not fulfilled [25]. Such collective codes thus identify groups of languages that could be used to model the genetic affiliation of languages, e.g., `ine` for *Indo-European languages* or `sla` for *Slavic languages*. However, the all-together 55 collective codes[12] are of course not sufficient to model linguistic affiliation across several Central European languages as will be required within the SFB.

Next to the individual language level and the language group level, there also exists a diachronic level in the ISO 639-2 code set: Diachronic varieties such as `gmh` *Middle High German* or `goh` *Old High German* can be identified by these alpha-2 codes. These multiple layers and the fact, that they are not clearly distinguished from each other clearly, points to the roots of the ISO 639-2 codes in bibliography.

3.1.2 ISO 639-3

What if an individual language did not meet the criteria for ISO 639-1 and ISO 639-2 and somebody still wanted it to be registered in the ISO 639 standard family? These

[10] Such as "a standards organization, governmental body, linguistic institution, or cultural organization" [25].

[11] 21 languages have alternative codes either for bibliographic use or for use in terminology. In these cases, the bibliographic code is listed first [20].

[12] These numbers are provided by Kamusella [17], who counted 484 codes within ISO 639-2 in 2011. In 2017, we counted 506 items, which means that the number of collective codes might be slightly higher, too.

languages may be "candidates for inclusion in ISO 639-3", as the Library of Congress suggests [27].

The history of the third part of ISO 639 is thoroughly and critically analysed by Kamusella [17]. Generally, it is closely associated with the Summer Institute of Linguistics, now: SIL International, a missionary linguistic organization, which is quite well known for its main publication, the *Ethnologue* [28]. It aims at giving a comprehensive overview of all languages spoken worldwide. Basically, the ISO adapted the identification codes for individual languages that were introduced in the 10th edition of the *Ethnologue* in 1984 [28] as the third part of ISO 639.

In his presentations [30], [31], Gary Simons, currently Chief Research Officer at SIL International[13], frequently cites Einar Haugen [32], who distinguishes a structural and a functional view with regard to the distinction of languages from dialects. The structural view describes "the language itself", whereas the functional view focuses on "its social uses in communication". Similar distinctions can be found in other seminar works of early sociolinguistics, such as Kloss [26], who distinguishes a sociological and a philological view as early as 1929 [33].

Simons associates the structural use of the terms *language* and *dialect* with the one "most commonly held by linguists" [30], thereby legitimating the approach supported by SIL International. Furthermore, he states that, following the premises of variationist linguistics, that "languages are not static objects", a language identifier in ISO 639-3 "denotes some range of language varieties" [34]. The main criterion for the distinction of different languages is their intelligibility (comp. Klosses *Abstand* languages [26]); the ethnolinguistic identity of a group of speakers is only considered in the second place [34].

This leads to a "Bible translation-based overcounting of languages imposed from outside", as Kamusella puts it [17, p. 76]. He assumes that SIL would count up to 40 different languages within the area in Central Europe, where German is spoken [17]. This estimation is quite realistic, if we consider that the *Ethnologue* [28] lists five Germanic languages within Austria (see below).

For modelling language variation of German in Austria, ISO 639-3 provides the code `bar` for *Bavarian* but no equivalent code for Alemannic varieties, which are spoken in Vorarlberg. The code `gsw`, which identifies "*Swiss German, Alemannic, Alsatic*" in ISO 639-2, does only refer to "*Swiss German*" in ISO 639-3. In 2011, a request[14] was made to register a code, preferably `aeg`, for "*Alemannic*". The code was supposed to have macrolanguage status and cover the individual languages `gct` *Colonia Tovar*, `gsw` *Swiss German*, `swg` *Swabian*, and `wae` *Walser*, which were already registered in ISO 639-3. The change request was rejected, because *Alemannic* would not meet the requirements for macrolanguages, as the individual languages listed above would not collectively be referred to as *Alemannic* in any contexts [35].

[13] `https://www.sil.org/biography/gary-f-simons`, retrieved 2017-03-16.

[14] The code request was made by Clemens-Valentin Kientzle, who seems to have been a student at the University of Freiburg, Switzerland at that time and had a leading position in the development of an Alemannic Wikipedia in 2011 (`http://www.freiburger-nachrichten.ch/kanton/sie-schreiben-wie-ihnen-der-schnabel-gewachsen-ist-aus-freude-am-dialekt`, retrieved 2017-03-29). The latter fact may have been a motiviation for the code request.

The existence of a code for Bavarian with the status of an individual language, a status, which could also be questioned, but no equivalent one for Alemannic of course makes it impossible to model, at least the dialectal groups of German in Austria.

As the ISO 639-3 standards and the *Ethnologue* [28] are closely related, it is interesting to take a look at its latest edition. The *Ethnologue* lists several varieties of German in Austria, which would be treated as such in a framework of German variationist linguistics, as individual languages (see Table 3)[15]. Interestingly, the *Ethnologue* uses `gsw` in order to refer to *Alemannic* in general.

It is obvious that from a variationist linguistic point of view, this list and its mapping to certain regions is simplistic, incoherent and based on questionable facts. The underlying *Abstand* paradigm implies distinct languages where a variationist perspective would be more appropriate. Thereby, it renders phenomena of vertical variation within the standard-dialect spectrum invisible.

ISO 639-3 does not assign any collective codes. Therefore, it is not possible to model linguistic affiliation with this code set.

code	***name***	***region***
gsw	Alemannic	Vorarlberg
bar	Bavarian	Lower Austria, Salzburg, Burgenland, Carinthia, Styria
deu	Standard German	Vorarlberg[16]
swg	Swabian	Tyrol, around the town of Ruette
wae	Walser	Tyrol, Paznauntal area

Tab. 3. Varieties of German in Austria according to the *Ethnologue*

3.1.3 ISO 639-5

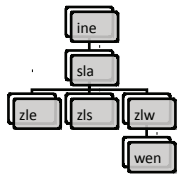
In 2009 the ISO published a fifth part of the ISO 639 standard family. It provides 'codes for language families and groups', some of which were already included in ISO 639-2. According to the Library of Congress, which maintains ISO 639-5 as well, these codes are intended to "support the overall language coding" and do not "provide a scientific classification of the languages of the world" [36].

For modelling the linguistic affiliation of Slavic languages in general, ISO 639-5 currently provides the codes listed in Table 4. As can be seen, for a basic model of linguistic affiliation within the Slavic languages the codes can be used quite accurately. However, if there exists a code for the Sorbian languages `wen`, it would be favourable to also have codes for the Czech-Slovak languages, Lechitic languages, and so forth.

[15] See `https://www.ethnologue.com/country/AT/languages`, retrieved 2017-03-15.

[16] On the relevant map, Standard German is linked to the cities of Vienna, Graz and Linz. On the other hand, it does not assign Standard German to Vorarlberg (see `https://www.ethnologue.com/map/AT`, retrieved 2017-03-15). Thus, the *Ethnologue* even contradicts itself.

[17] The codes that are novel in ISO 639-5 and had not already been part of ISO 639-2 are marked with an asterisk (*).

code[17]	name	ISO 639:5 hierarchy
`ine`	Indo-European languages	
`sla`	Slavic languages	
`zle*`	East Slavic languages	
`zls*`	South Slavic languages	
`zlw*`	West Slavic languages	
`wen`	Sorbian languages	

Tab. 4. Modelling the Slavic languages with ISO 639-5 codes

Generally, we conclude that static lists of language codes, no matter whether designed in library contexts or linguistic enterprises, do hardly account for aspects of language variation. Still, if language data shall be annotated according to machine-readable standards (such as XML) in order to be processed by several applications, ISO 639 codes or language tags according to BCP 47 [37] have to be used.

3.2 Language Tags According to BCP 47

As already emphasised above, in comparison to language codes, language tags allow for annotating certain degrees of variation. In their context language variations can best be described as deviations from default settings. Language tags that can be used in XML annotations, e.g., following the `xml:lang` attribute, have to be designed according to BCP (Best Current Practise) 47 [37], a document issued by the IETF (Internet Engineering Task Force). This organisation's ambitious mission is to "make the Internet work better by producing high quality, relevant technical documents that influence the way people design, use, and manage the Internet" [38]. In contrast to the ISO, IETF relies on free community participation and is organised by the non-profit ISOC (Internet Society).

BCP 47 was issued in September 2009. According to this document, a language tag has the following structure, in which the individual subtags (e.g., `language` or `script`) need to be used in the given order.

```
language-extlang-script-region-variant-extension-privateuse
```

Except for the `language` subtag, the positions do not need to be specified and can be left empty. Some values even have to be suppressed with a certain `language` subtag, e.g., the script must not be specified, if a German text is written in Latin (see Fig. 2). If, on the other hand, it was written in Cyrillic, the script would have to be specified (`de-cyrl`). The W3-Consortium also advises to "keep the tag as short as possible" and thus encourages to leave out redundant subtags [39].

The individual subtags may only have certain values. Valid subtags are registered in the *IANA language subtag registry*[18]; the registration process for new subtags is described in BCP 47. Some subtag values are generally associated with ISO standards (see Table 5).

[18] `http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry`, retrieved 2017-03-21, see Fig. 2 for an example. A subtag search tool is provided on `https://r12a.github.io/app-subtags/`, retrieved 2017-03-21.

```
%%
Type: language
Subtag: de
Description: German
Added: 2005-10-16
Suppress-Script: Latn
%%
```

Fig. 2. Entry for the language subtag "German" in the IANA *language subtag registry*

subtag	ISO standards
language	ISO 639 *Codes for the representation of names of languages*, preferably ISO 639:1
extlang[19]	ISO 639 *Codes for the representation of names of languages*, especially ISO 639:3
script	ISO 15924 *Codes for the representation of names of scripts*
region	ISO 3166 *Country codes*

Tab. 5. ISO-Standards in language tags according to BCP 47

The `variant` subtag may only carry values registered in the IANA *language subtag registry*. Each of these values is tied to a specific language and can therefore only be used in combination with a certain primary `language` subtag. There are two values for the `variant` subtag registered, which can be combined with the `language` subtag `de` (German, see Fig 3.) and none for `cs` (Czech).

```
%%
Type: variant
Subtag: 1901
Description: Traditional German orthography
Added: 2005-10-16
Prefix: de
%%
Type: variant
Subtag: 1996
Description: German orthography of 1996
Added: 2005-10-16
Prefix: de
%%
```

Fig. 3. Variant subtags for German in the IANA *language subtag registry*

If more specifications are needed, there are two possibilities: `extension` subtags are singletons that can be registered with IANA by organisations. Following these singletons, the according organisations themselves may define more subtags and their values.

[19] Extended language subtags, i.e., extlang subtags, are used to identify languages that are closely linked or seen as a variant of another language due to some reasons. Some variants of pluricentric languages such as Arabic can be described in that way, if there are ISO 639-3 codes for their single variants. A language tag consisting of a language subtag and an extlang subtag for Gulf Arabic would thus be ar-afb. But, it could also and should be referred to with a primary language subtag only (afb) [39].

`Private-use` sequences work similarly. They always begin with the singleton `-x-`, which is followed by subtags that are privately agreed on within a certain community. The W3-Consortium advises to use them "with great care", as they are "only meaningful within private agreements and cannot be used interoperability across the Web" [39]. Unfortunately, they seem to be the only solution for scientific projects with a variationist linguistic focus such as the SFB "German in Austria", because variation in language cannot be sufficiently annotated in XML-documents with the basic language tag syntax.

Still, we propose a language tag consisting of a primary language subtag and a region subtag to generically refer to the object languages that the SFB "German in Austria" is interested in, i.e., the whole spectrum of varieties of German used in Austria. This tag will serve as the basis for further specifications as long as our system has the status of a working annotation.

`de-AT`

In the long run, we could, of course, register an `extension` subtag, but such a system should be agreed upon as a community standard within at least German variationist linguistics and needs to be well designed and pretested.

4 SOLUTIONS FOR THE SFB: A SYSTEM IN PROGRESS

4.1 Module A: Modelling the Genetic Affiliation of Languages

As mentioned above, we consider module A as an auxiliary construction and therefore prefer pragmatic solutions and accept simplifications. Hence, we transfer a phylogenetic tree model into a relational database (see Table 6 and 7). Thereby, we make groups and categorisation levels more explicit than in the graphic representation of the tree model, in which the generations of different branches can only be 'seen' implicitly. Concerning the content, we first of all need to agree on a harmonisation of different tree models for several languages used in Central Europe. Secondly, we will have to define, what the name of each language family or group designates in an underlying ontology.

Currently, we have agreed on using and adapting the *Composite model* for the Indo-European languages developed by the MultiTree project [40] for the levels above individual languages, i.e., for language families and groups. Within module A, names for individual languages such as "German" or "Czech" do not refer to codified standard languages but to variety bundles, which are commonly addressed (and/or constructed) as "German" or "Czech". We also model levels of dialectal groups, such as Upper German with its subnodes Bavarian and Alemannic, because these levels are not considered research objects within the SFB.

Tables 6 and 7 exemplarily show, how the linguistic affiliation of Czech and Slovak would be modelled. In the *belongs_to* column, Table 6 refers to its own *ID* column; in the *type* column, it refers to Table 7. Note that type 3 is not assigned to any variety in Table 6. That level is needed to model the affiliation of German based on a simplified model adapted from MultiTree [40], which is depicted in Fig. 4. This figure also represents the type of visualisation underlying Tables 6 and 7, with the greyish bars corresponding to the variety types in Table 7 and the single boxes to the varieties in Table 6.

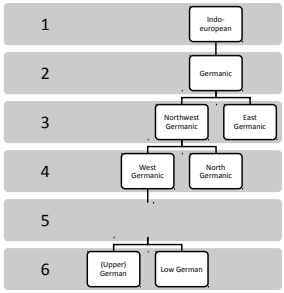


Fig. 4. Tree model for the Germanic languages with focus on German

ID	***variety_name***	***type***	***belongs_to***
1	Indo-European	1	
2	Slavic	2	1
3	East Slavic	4	2
4	West Slavic	4	2
5	South Slavic	4	2
6	Lechitic languages	5	4
7	Sorbian languages	5	4
8	Czech-Slovak languages	5	4
9	Czech	6	8
10	Slovak	6	8

Tab. 6. Variety table from module A for the Slavic languages with focus on Czech and Slovak

ID	**type_name**
1	language family
2	language group
3	subgroup 1
4	subgroup 2
5	subgroup 3
6	individual languages

Tab. 7. Variety type table from module A

4.2 Module B: Object Language Annotation System

The objective of module B is to provide an annotation framework for several dimensions of language variation of German in Austria. It shall enable corpus linguistic analyses, but should not impose a pre-defined classification upon the data. Table 8 shows the dimensions that it will have to account for. Furthermore, we indicate, which factors might specify these dimensions in the corpus of data collected within the SFB, as well as in other, external linguistic sources such as linguistic literature or other language resources.

Dimension	**factors within the corpus**	**factors in other linguistic sources**
vertical variation *on the standard-dialect axis*	intended register	according classification
	code switch or style shift	
horizontal *(diatopic)* ***variation***	place of recording	place of reference
diachronic *variation*	time of recording	time of reference
idiolectal *dimension of variation*	informant	author

Tab. 8. Dimensions of language variation to be considered in module B.

These factors could be transferred into a `private-use` language tag sequence that would have to be specified as belonging to module B by the singleton `-b-`.

```
de-AT-x-b-place-time-intend_register-register_shift-person
```

The `place` and `person` subtags will take their values from the place and person specific parts of the SFB database. Especially the values for subtags, which carry information on vertical variation, will be defined during the working process. The developing working group provides the technical framework, as well as the guidelines for adding and documenting language subtags and their values.

4.3 Module C: Annotation System for the Names of Varieties

The names of varieties and the connotative and/or evaluative meaning they develop depending on their *context*, their *reference* and the *kind of reference*, will be important for some SFB project parts, too. We understand *context* as the kind of text the glottonym appears in. Whether it is an interview conducted by the SFB or a 19th century legal text will clearly make a difference in its meaning. *Reference* is the language or variety, the glottonym refers to. It may specify this language or variety by attaching it to a certain place (e.g., *Viennese*), a certain person (i.e., an idiolect), or by embedding it in time. Furthermore, a glottonym may carry information on the register, which the reference language belongs to. The *kind of reference* expresses, whether the glottonym refers to the person using it and his/her way of speaking (*self-classification*) or whether he/she uses it to describe somebody else's speech (*hetero-classification*).

In XML, a relevant language tag would not follow the `xml:lang` attribute, because this attribute may only specify the object language, i.e., the language a source is written or spoken in. On the other hand, it would follow a `lang` attribute, which allows for the specification of language names, e.g., according to the TEI P5-guidelines [11]. Such a language tag would have the form:

```
language-x-c-place-time-register-person-context-reference_kind
```

The `language`, `place`, `time`, `register` and `person` subtag annotate the reference. Again, the singleton `-c-` indicates the module, in the context of which the tag needs to be interpreted.

5 CONCLUSIONS

This article has provided a glimpse into the development of a custom-made annotation framework for language varieties, which evolves from and shall be used within a collaborative linguistic project. It will serve as a working annotation and ultimately also enable querying the corpus of German in Austria, which is compiled by the eponymous SFB. On the long run, it shall also enable classification of varieties of German in Austria and provide a best practise example that might initiate the definition of community standards. Potentially, this best practise example can be extended to a new standard in terms of variationist linguistic variety annotation, if accepted and adopted by the community.

References

[1] Glauninger, M. M. (2012). Zur Metasozioseminose des ›Wienerischen. Aspekte einer funktionalen Sprachvariationstheorie. *Zeitschrift für Literaturwissenschaft und Linguistik*, 42(2):110–118.

[2] DiÖ (2017). Task-Cluster B: Variation. *DiÖ-Online*. Accessible at: https://dioe.at/en/projects/task-cluster-b-variation/, retrieved 2017-03-29.

[3] DiÖ (2017). Task Cluster C: Contact. *DiÖ-Online*. Accessible at: https://dioe.at/en/projects/task-cluster-c-contact/, retrieved 2017-03-29.

[4] DiÖ (2017). PP05: German in the context of the other languages of the Habsburg state (19th century) and the Second Austrian Republic. In *DiÖ-Online*. Accessible at: https://dioe.at/en/projects/task-cluster-c-contact/pp05/, retrieved2017-03-29.

[5] DiÖ (2017). PP06: German and the Slavic languages in Austria: Aspects of language contact. In *DiÖ-Online*. Accessible at: https://dioe.at/en/projects/task-cluster-c-contact/pp06/, retrieved 2017-03-29.

[6] Berger, Tilman (2007). Böhmisch oder Tschechisch? Der Streit über die adäquate Benennung der Landessprache der böhmischen Länder zu Anfang des 20. Jahrhunderts. In Nekula, M., Fleischman, I., and Greule, A., editors, Franz Kafka im sprachnationalen Kontext seiner Zeit. Sprache und nationale Identität in öffentlichen Institutionen der böhmischen Länder, pages 167–182, Böhlau, Köln – Weimar – Wien, Germany.

[7] DiÖ (2017): Task-Cluster D: Perception. In: *DiÖ-Online*. Accessible at: https://dioe.at/en/projects/task-cluster-d-perception/, retrieved 2017-03-29.

[8] DiÖ (2017). Task-Cluster E: Collaborative Online Research Platform. *DiÖ-Online*. Accessible at: https://dioe.at/en/projects/task-cluster-e-research-platform/, retrieved 2017-03-29.

[9] SYHD (2016). Syhd.info. Accessible at: http://www.syhd.info/startseite/, retrieved 2017-03-30.

[10] Schmidt, J. E., Herrgen, J., and Kehrein, R., editors (2008ff.) Regionalsprache.de (REDE). Forschungszentrum Deutscher Sprachatlas, Marburg. Accessible at: https://regionalsprache.de/, retrieved 2017-03-30.

[11] Text Encoding Inititative (2016). *P5:* Guidelines for Electronic Text Encoding and Interchange. Accessible at: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/, retrieved 2017-03-29.

[12] Hammarström, H., Forkel, R., and Haspelmath, M. (2017). Glottolog 3.0. Accessible at: http://glottolog.org, retrieved 2017-03-29.

[13] Haspelmath, M. (2013). Can language identity be standardized? On Morey et al.'s critique of ISO 639-3. In *Diversity linguistics comment. Language structures throughout the world*. Accessible at: http://dlc.hypotheses.org/610, retrieved 2017-03-29.

[14] Dalby, D. (2012). The Linguasphere Register of the World's Languages and Speech Communities. First online reprint. Linguasphere press. Accessible at: http://www.linguasphere.info/lcontao/bienvenue-welcome.html, retrieved 2017-03-29.

[15] ISO (2017). ISO/TC 37. Terminology and other language and content resources. Accessible at: https://www.iso.org/committee/48104.html, retrieved 2017-03-28.

[16] ISO (2017). ISO 639-6:2009. Codes for the representation of names of languages – Part 6: Alpha-4 code for comprehensive coverage of language variants. Accessible at: https://www.iso.org/standard/43380.html, retrieved 2017-03-17.

[17] Kamusella, T. (2012). The global regime of language recognition. *International Journal for the Sociology of Language*, 218:59–86.

[18] Dalby, D., Gillam, L., Cox, Ch., and Garside, D. (2004). Standards for language codes: Developing ISO 639. In *Proceedings of the LREC 2004. Forth International Conference on Language resources and evaluation.* Accessible at: http://www.lrec-conf.org/proceedings/lrec2004/pdf/327.pdf, retrieved 2017-03-30.

[19] ISO (2017). ISO/AWI 21636. Identification and description of language varieties. Accessible at: https://www.iso.org/standard/71300.html?browse=tc, retrieved 2017-03-28.

[20] Library of Congress (2017). ISO 639-2 Codes for the Representation of Names of Languages. Accessible at: http://www.loc.gov/standards/iso639-2/php/code_list.php, retrieved 2017-03-17.
[21] SIL International (2017). ISO 639-3 Downloads. Accessible at: http://www-01.sil.org/iso639-3/download.asp, retrieved 2017-03-17.
[22] Library of Congress (2017). ISO 639-5 Codes for the Representation of Names of Languages. Part 5: Alpha-3 code for language families and groups. Accessible at: http://www.loc.gov/standards/iso639-5/id.php, retrieved 2017-03-17.
[23] Library of Congress (2017). ISO 639-2. Frequently Asked Questions (FAQ). Accessible at: http://www.loc.gov/standards/iso639-2/faq.html, retrieved 2017-03-17.
[24] Library of Congress (2017). Development of ISO 639-2. Accessible at: http://www.loc.gov/standards/iso639-2/develop.html, retrieved 2017-03-17.
[25] ISO 639 Joint Advisory Committee (2000): Working principles for ISO 639 maintenance (ISO 639/JAC N3R). Accessible at: http://www.loc.gov/standards/iso639-2/iso639jac_n3r.html, retrieved 2017-03-28.
[26] Kloss, H. (1978). *Die Entwicklung neuer germanischer Kultursprachen seit 1800*. Pädagogischer Verlag Schwann, Düsseldorf.
[27] Library of Congress (2017). Criteria for ISO 639-2. Accessible at: http://www.loc.gov/standards/iso639-2/criteria2.html, retrieved 2017-03-17.
[28] Simons, G. and Fenning, Ch. D. (2017). Ethnologue: Languages of the World. 20th edition, SIL International, Dallas, Texas. Accessible at: http://www.ethnologue.com, retrieved 2017-03-29.
[29] Ethnologue (2017). History of the Ethnologue. Accessible at: https://www.ethnologue.com/about/history-ethnologue, retrieved 2017-03-30.
[30] Simons, Gary (2014). Terminology and language aspects in language coding. Presented at the *TKE 2014 Workshop: Language Codes at the Crossroads*. Berlin, Germany, 21 June 2014. Accessible at: https://tke2014.coreon.com/slides/2014_06_21_104_1030_Simons.pdf, retrieved 2017-03-16.
[31] Simons, G. (2013). ISO 639-3. Where are we and how did we get here? Presented at the *Workshop on Identifying Codes for Languages.* Newcastle, Australia, 9 February 2013. Accessible at: http://www-01.sil.org/~simonsg/local/ISO%20639-3.pdf, retrieved 2017-03-16.
[32] Haugen, E. (1966). Dialect, Language, Nation. In *American Anthropologist, New Series*, 6(4):922–935.
[33] Kloss, H. (1929). *Nebensprachen. Eine sprachpolitische Studie über die Beziehungen eng verwandter Sprachgemeinschaften*. Braumüller, Wien.
[34] SIL International (2017). Scope of denotation for language identifiers. Accessible at: http://www-01.sil.org/iso639-3/scope.asp, retrieved 2017-03-29.
[35] SIL International (2017). Change request documentation for: 2011-180. Accessible at: http://www-01.sil.org/iso639-3/chg_detail.asp?id=2011-180&lang=aeg, retrieved 2017-03-21.
[36] Library of Congress (2017). Codes for the Representation of Names of Languages. Part 5: Alpha-3 code for language families and groups. Introduction. Accessible at: http://www.loc.gov/standards/iso639-5/langhome5.html#intro, retrieved 2017-03-17.
[37] Phillips, A. and David, M. (2009). BCP 47. Tags for Identifying Languages. Accessible at: http://www.rfc-editor.org/rfc/bcp/bcp47.txt, retrieved 2017-03-29.
[38] IETF (2017). Mission Statement. Accessible at: https://www.ietf.org/about/mission.html, retrieved 2017-03-29.
[39] W3C (2017). Language tags in HTML and XML. Accessible at: https://www.w3.org/International/articles/language-tags/index.en#overview, retrieved 2017-03-29.
[40] MultiTree (2013). Indo-European: Composite. In *MultiTree: A digital library of language relationships*. Institute for Language Information and Technology, Ypsilanti, MI. Accessible at: http://multitree.org/, retrieved 2017-03-20.