CORPUS

Corpus

23 | 2022
Corpus et données en morphologie

Corpus linguistics for low-density varieties. Minority languages and corpus-based morphological investigations

Linguistique du corpus pour les variétés à faible densité. Langues minoritaires et enquêtes morphologiques basées sur corpus

Livio Gaeta, Marco Angster, Raffaele Cioffi and Marco Bellante

OpenEdition Journals

Electronic version
URL: https://journals.openedition.org/corpus/7345
DOI: 10.4000/corpus.7345
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

ELECTRONIC REFERENCE

Livio Gaeta, Marco Angster, Raffaele Cioffi and Marco Bellante**, "Corpus linguistics for low-density varieties. Minority languages and corpus-based morphological investigations"**, *Corpus* [Online], 23 | 2022, Online since 02 March 2022, connection on 05 March 2022. URL: http://journals.openedition.org/corpus/7345 ; DOI: https://doi.org/10.4000/corpus.7345

Corpus linguistics for low-density varieties. Minority languages and corpus-based morphological investigations*

Linguistique du corpus pour les variétés à faible densité. Langues minoritaires et enquêtes morphologiques basées sur corpus

Livio Gaeta, Marco Angster, Raffaele Cioffi and Marco Bellante

1. Introduction

Corpus linguistics grew up in the domain of written (and literary) language, while its recent methodological revolution is due to the computer-assisted capacity of elaborating massive amounts of text data (McCarthy, O'Keeffe 2010). On the other hand, spoken corpora, especially those coming from fieldwork, have also attracted the interest of corpus linguistics, although they have been confined to a rather marginal role. This is partially due to the technical problems connected with the transcription and elaboration of recorded speech, but also to the dramatic issue of comparability with huge written corpora. Similar problems arise dealing with written and spoken corpora of minority languages, whose size is dramatically small. Their data are scarcely normalized both in phonological and orthographic terms (Jones, Mooney 2017). Furthermore, the scarcity of language resources impedes the automatic processing of linguistic data for these so-called 'low-density varieties' (Maxwell, Hughes 2006). This term identifies those languages which – in neat contrast to the few 'high-density languages' – lack in computational resources, first of all texts provided with linguistic annotation, which are a prerequisite for Natural Language Processing. In addition, we prefer to speak of 'variety' instead of 'language' with the aim of including also (diamesic and diatopic) varieties of (written or standard) high- and middle-density languages which are not adequately provided with computational resources. By the

same token, the usage of 'variety' also aims at avoiding any discussion or stance-taking with regard to the different status of standard languages, regional languages or dialects, minority or endangered languages, etc., which in our view can all potentially provide examples of low-density languages, although in different respects.

1.1. Corpus linguistics for small and minority varieties

As a matter of fact, the spread of resources for NLP, combined with the large diffusion of computers and digital texts on the Web for a wide number of languages,[1] has definitely driven corpus linguistics towards dealing with low-density varieties. This is for instance the case of the languages of Ex-Yugoslavia[2] as well as of Swiss and Austrian German.[3] Besides, there are also examples of corpora of contact[4], minority[5] and migration[6] varieties, especially under the widespread term of 'heritage languages'.[7] These resources are an important extension of corpus linguistics in less common areas.

1.2. Corpus linguistics for endangered varieties

On the other hand, the interest of corpus linguistics for endangered varieties has remained scarce. This is probably due to their condition of under-represented and under-used varieties, which are often facing obsolescence or even subject to language shift, i.e. to their loss in favor of the majority's variety. Moreover, besides their condition of submission to an often radically different majority variety, they offer too small text corpora for being treated with the help of the stochastic methods of corpus linguistics. Such methods usually require a certain critical mass of annotated data, going far beyond the reduced size of the corpora available for minority varieties. These low-density varieties are highly interesting from a linguistic and a sociolinguistic point of view, as they witness of syntactic and morphosyntactic phenomena which are often very different from the written or spoken majority varieties.

At the same time, the possibilities opened by corpus linguistics are particularly valuable for endangered varieties because they offer a terrific chance for documenting and preserving a cultural and linguistic heritage which otherwise will be irremediably lost. In this vein, a number of projects have recently attempted to employ tools and methods of corpus linguistics for acquiring and analyzing the textual patrimony of the Walser German communities of Piedmont and Aosta Valley (Angster et al. 2017, 2020, Gaeta et al. 2019, Gaeta in press). The varieties of Highest Alemannic spoken there, dramatically exposed to language decay (Dal Negro 2004, Zürrer 2009), provide a limited but significant amount of data, which is accompanied by a substantial lexical documentation due to the active collaboration of the speakers' communities in collecting and compiling local dictionaries. It goes without saying that this documentation presents huge differences among the single varieties and texts with regard to their concrete elaboration into an archive (see for recent attempts on Walser islands Fazzini et al. 2004-, and on other Swiss German varieties, including Walliser German, Garner 2014, Samardžić et al. 2015, 2016, Scherrer, Ljubešić 2016, Honnet et al. 2018, as well as the work of the Zurich LORELAI initiative[8]).

As they were started with the aim of preserving, documenting and investigating the cultural and linguistic heritage of the Walser German communities, these projects constitute an attempt of combining the extremely dynamic realm of corpus linguistics with textual data coming from low-density varieties. In this paper the main focus will be on the ongoing project CLiMAlp (see http://www.climalp.org/) which essentially

expands and improves the old platform developed within the projects DiWaC and ArchiWals. In CLiMAlp data from other two low-density (Romance) varieties are also introduced, namely Franco-Provençal and Occitan.[9] In spite of the dramatic process of attrition, in the last three decades a remarkable process of cultural and linguistic revival took place, also supported by the Law 482/99 for the safeguard of linguistic minorities. This gave rise to a considerable text production, although of a different and heterogeneous nature and type across the single communities. This recent increase of written text production will provide the empirical basis of the present contribution, which will focus on crucial issues such as the presence/absence of standardization and the granularity of linguistic data, showing concrete solutions for problems which are potentially relevant also for spoken corpora. In particular, we will focus on Titsch, the Walser German variety spoken in Gressoney (Angster et al. 2017, Gaeta et al. 2019).

1.3. Small size and high granularity

Spoken corpora and corpora of minority varieties share the small size, which is clearly connected to a usually limited extent of oral documentation: as a matter of fact, corpora of minority varieties often result from the elaboration of transcribed conversations or interviews. On the other hand, texts written in minority varieties are often characterized by brevity and reduced complexity similar to that of spoken varieties. At any rate, they display a rather high degree of granularity. The latter can be intended in several ways. We emphasize the following three aspects without any claim of exhaustivity:

- granularity as complexity of the transcription;
- granularity as complexity of the metadata;
- granularity as complexity of the annotation.

The first aspect apparently concerns only spoken corpora, but in fact if we intend transcription in broader terms involving also the orthographic system we can also apply it to written texts.

The second aspect –the metadata[10]– concerns any sort of data, but it clearly increases as long as the distance between the observer and the data is reduced. For minority varieties metadata are usually very rich, similarly to spoken corpora directly recorded by the fieldworker, while this information is far less available for large spoken corpora –especially when they are indirectly acquired– and even more so for written corpora of a large dimension. In this latter case, metadata are often held to be irrelevant, as it is the large size of the corpus which warrants for the reliability of the generalizations captured. It is not by chance that for Web-based big corpora even residual metadata relating to genre or text types appear often superfluous.

The third aspect relating to annotation is multi-faceted. A basic annotation as for instance POS-tagging, usually accompanied by lemmatization, substantially increases the computability and the enrichment of corpus by means of other annotation levels (e.g. syntactic chunking). On the other hand, while POS-tagging and lemmatization proceed quite straightforwardly on a high-density variety provided with rich computational resources, they are extremely problematic for low-density varieties which are often idiosyncratic with regard to the most widespread orthographic standards. The only viable alternatives consist in an accurate manual annotation or in a new training stage on the basis of a manually annotated corpus of a sufficient size.

Thus, (manual) annotation presents initial difficulties of an empirical nature, but at the same time it smooths the path of any further level of automatic elaboration.
The relation between granularity –in the three perspectives discussed above, and especially the last one– and computability is probably the most relevant aspect regarding the application of corpus-linguistic methods to low-density varieties on which we will focus in the rest of the paper (cf. Gaeta in press for further discussion).

2. Towards orthographic standards for minority varieties

Since annotation is a crucial initial step for building corpora of small varieties, our projects had to deal with the peculiar issues concerning the orthographic standards adopted across the Walser German communities. While the overall token number of the corpus is quite reduced, their orthographic instability is pretty strong, as expected for this type of low-density varieties where the issues relating to language planning and ethno-linguistic identity also influence the development and the adoption of an autonomous orthographic system (Iannàccaro 2010). As observed above, this instability has a direct impact on the lowered degree of computability of the linguistic data.

2.1. Across norm and variation in Gressoney

The orthographic system employed in Gressoney was elaborated in the first days of the Walser revival when a few local enthusiasts founded the Walser Kulturzentrum (WKZ 1982) giving impulse to the publication of several printed works, and in particular the WKZ, a dictionary of the variety of Gressoney. Dictionaries in particular acquired the status of reference work especially for the written dimension, although in an implicit and unforced way. This moderate level of standardization did not improve in the following years, even if other attempts were made by cultural associations as well as by professional linguists to introducing common and more rational writing norms (see Antonietti 2010).
On the other hand, the reduced number of authors of the written texts limited the proliferation of orthographic variants, while the process of partial stabilization of the writing customs adopted within the community made even more varied the general picture observed through the years as well as across different documents of the same period, if not within the single texts. A similar degree of instability appears also in the transcriptions of recorded interviews, tales, conversations, etc., which were published in the bulletin of the community. In this light, one should not underestimate the role of the metalinguistic uncertainty of the speakers/writers who generally acquire literacy in the standard/majority's languages, where different orthographic norms are adopted, not straightforwardly usable in the minority variety.
In addition, the repertoire relating to the written varieties widespread in the community has changed radically in the last 150 years. German as H-code variety –which used to be taught in school– was slowly replaced first by French, and subsequently by Italian (cf. Angster 2014, Angster, Gaeta 2021). Thus, the actual orthographic system basically relies on phoneme/grapheme correspondences similar to the German system, but has been enriched by contributions drawn from other, and especially Romance, systems. For instance, the voiced palato-alveolar fricative /ʒ/, which is unknown in German or Italian with the exception of a few French loanwords –

compare respectively German *Jongleur* 'juggler' [ʒɔŋ(g)ˈløːɐ̯] and Italian *maquillage* 'maquillage' [makiˈjaʒ]–, was initially represented in the texts of the parochial bulletins in accordance with the French tradition. Accordingly, the phoneme/grapheme correspondence /ʒ/ ⇔ <j> is found in grammatical words as well as in other lexemes – e.g. *dije* for /ˈdɪʒe/ 'this' (see the German cognate *dieser*) and *hijer* for /ˈhiʒer/ 'houses'. Later, after the publication of the dictionary, a different option was adopted which relies on the consonantal cluster <sch> also used for the corresponding voiceless fricative /ʃ/ to which a diacritic sign on <ŝ> is added, providing the phoneme/grapheme correspondence /ʒ/ ⇔ <ŝch>, possibly also used for the affricate /dʒ/ ⇔ <dŝch> as in *dŝchi* for /dʒi/ 'she'.[11]

A pronounced variability is especially found with a number of morpho-phonological phenomena, and in particular those concerning prosodic words containing a clitic (group). For instance, in the easiest case containing only one clitic we find for the sequence *hät es* 'has it' forms like *häts* or *hätz* (see §3.4 below for further discussion). Further examples of orthographic instability concern morpho-lexical phenomena, as for instance the case of the so-called particle verbs (cf. Gaeta 2021), in which the preverbal particle can be either attached or not to the verb: *zròck chéeme/zròckchéeme* 'to come back', while other particle verbs always appear spelled as a single orthographic word: *achéeme* 'to arrive'.[12] Finally, an example of syntactic ambiguity mirrored by orthographic instability is given by the usage of the form *dass* or *daß* – which in Standard German orthography only stand for the conjunction 'that'– for the demonstrative expressed in Standard German only by *das* 'this'. These examples show the massive impact of orthographic instability on an automatic treatment of the linguistic data for such a small variety, which is at the same time at least moderately standardized.

Such orthographic instability is to a certain extent similar to the phonetic variability normally observed in spoken corpora, which raises the crucial question of the granularity of the representation discussed above. In both cases, the relation between text or phonetic string and its correspondence as a lexeme/lemma is far more complex than what is normally observed in the token/lemma relation in the corpus linguistics of high-density, largely standardized varieties.

3. The architecture of the ArchiWals platform

The absence of orthographic standardization discussed above, combined with the reduced number of available tokens, makes unprofitable the application of current models of POS-taggers (for instance the German version of TreeTragger)– even for the largest available corpus, i.e. that of Gressoney. A POS-tagger requires a training corpus of about 80,000 tokens, which roughly corresponds to the current size of the Gressoney corpus of written texts. In addition, the time-consuming manual annotation of a training corpus does not warrant any reliable results in view of the pronounced graphic instability.[13]

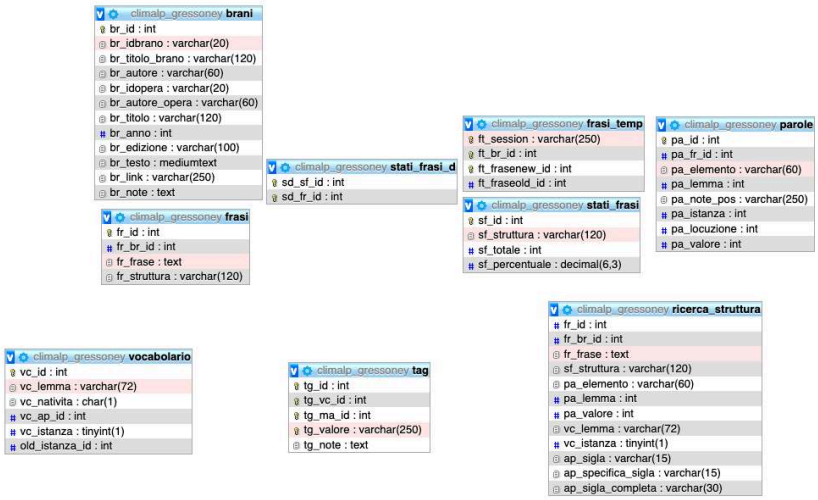
3.1. The general architecture of the database

Because of the difficulties of an automatic annotation, we decided to design a new platform in which data of low-density varieties might be easily accommodated. The platform consists of a multi-layered database containing in two interwoven structures

on the one hand the dictionaries collected by the single communities and on the other the corresponding text corpora.

Each lexical entry is equipped with its own structure containing all lexical and linguistic basic information (POS, lexical category, etc.) and with further fields containing specific tables for: nominal, verbal, etc. inflection; etymology; possible morphological relations (e.g. between bases and derivatives or compounds, the occurrence of affixes, particles, etc.). A further separate field is devoted to orthographic variants. Currently, no phonological representation, including stress, is provided. The multi-layered structure is directly interfaced with a second structure containing the digitized texts which are provided with metadata and sorted for genre and publication year (cf. Angster et al. 2017, Gaeta et al. 2019 for the details). This platform is meant to be flexible and easily adaptable –also in the future– to other minority varieties. What makes it flexible is its structure interfaced with different databases, where each database is arranged on different levels of data-consistency (the individual strata, see Fig. 1). This is extremely useful for dealing with the high granularity of the data which is typical of minority varieties. As can be gathered from Fig. 1, in the platform the different strata containing the whole texts, the metadata, the single lemmas, etc. are directly interfaced with each other.

Fig. 1. The structure of the platform



The substantial innovation of the platform lies in the capacity of treating each lexical entry together with its corresponding features, its single corpus occurrences (the tokens), as well as with the correspondence between lexical entries and corpus occurrences with the help of the logic of relational databases. Each lexical entry is provided with a unique ID which is linked to the whole range of features specified in the different strata of the database: orthographic variants, POS, morphological paradigms (possibly subdivided into single derivational elements linked to a unique ID), translations, etc. The stratification of the data allows us to acquire in the corpus texts with a different internal structure, applying a number of automatized operations which classify and optimize the texts for the linguistic analysis. Accordingly, the texts are first stored as such in a first layer, then as single strings in a second one, and finally as single occurrences of each single string in the last layer (so-called “explosion”). As mentioned above, each lexical entry is provided with its own ID to which all

information contained in the individual layers including metadata is linked. In this way, it is possible to acquire a text in the archive keeping it completely unaltered in order to be philologically accurate, and possibly to modify it only in specific layers where possible mistakes were found, only correcting the single "exploded" strings, namely the single tokens.
A further advantage of the platform is that it does not require to compile or re-compile the texts to carry out the lemmatization or the token/lemma matching. Finally, it is possible to create an unlimited number of further super-structures referring to syntactic or morphosyntactic elements to accommodate any possible linguistic variety. The structure and the algorithms on which the platform is based have been registered as technological license.[14]

3.2. Direct link between the lexicon and the corpus

The platform has been conceived with the aim of easing the management and the enrichment of the dictionary –which consists of different tables interconnected to each other and containing lexical and morphosyntactic information– and of concretely using it as a database for the lemmatization of the texts in the corpus. Accordingly, the lexicon of the single varieties (represented in the digitized dictionary) has been directly interfaced with the tokenized content of the corpus, allowing us a first automatic lemmatization. In this way, a consistent number of occurrences marked as 'null', i.e. not assigned to any lemma, was identified. In particular, besides 'null' tokens resulting from inflected forms which were connected to the extant entries in the corresponding inflectional tables, a significant number of fresh lexical entries was discovered, which reveals a certain discrepancy between the original dictionary and the lexicon used in the texts. They substantially enriched the original dictionary, especially with compounds and derivatives. For instance, we found new lexical entries like the simple verb *bieche* (1a) for which a meaning 'to roast gently, sauté' can be inferred from the context (cf. Middle High German *biuchen*, *būchen* 'to buck', Modern German *beuchen* 'to buck')[15], or the noun *lunò* (1b) directly corresponding to German *Laune* 'mood':[16]

(1)	a.	*Z'bròt*	*en*	*glichmessege*	*blettiene*
		DEF=BREAD	in	equal-measured.PL	leaf.PL

		hackò,	*ém*	*ange*	*biéche,*	(DOK_0236)
		chop.INF,	in.DEF	butter	roast.INF	
		'Cut the bread in uniform slices and roast it in the butter'.				

	b.	*Woa*	*d'huslitté*	*voller*	*lunò*	*sinn,*	(BEL_0476)
		where	DEF=HOUSE.PEOPLE	full	mood	be.3PL	
		'where the local people are good-humored'					

Furthermore, in the corpus we found new derivatives (2a), or new compounds with already attested lexical entries (2b), or new lexical entries or derivatives only occurring in compounds (2c):

(2)	a.	*lénne* 'to soften'	⟵	*lénn* 'gentle, soft'
		kòksò 'to storm'	⟵	*kòks* 'snow storm'
		wacher 'watchperson'	⟵	*wache* 'to watch'
		fannòtò 'panful'	⟵	*fannò* 'pan'
		ufschribetò 'annotation'	⟵	*ufschribe* 'to make a note'
		dròckeri 'print office'	⟵	*dròck* 'print'
		nòmmeriere 'to number'	⟵	*nòmmer* 'number'
		vergéftòng 'poisoning'	⟵	*vergéfte* 'to poison'[17]

	b.	*polléntòwasser*	'polenta:water, boiling water of the polenta'
		wuecherzénz	'usury:interest, usurious interest'
		bierbrouerei	'beer:brewery, beer brewery'

	c.	*reinégòng* 'clean' from *bluetreinégong* 'blood cleaning'
		wéerchma 'worker' from *gruebòwéerchma* 'miner'
		damò 'checker' from *damògschpél* 'checkers'

Furthermore, a certain number of new entries also comes from loanwords which have a correspondence in the German lexicon (like the example *damògschpél* which corresponds to the German compound *Damespiel* 'checkers', but in German the word *Dame* 'lady' is also found) or not. The following cases display different degrees of integration:

(3)	a.	*dominazion* 'dominio'	
		ònder Englésché dominazion	(DOK_0016)
		'under English dominion'	

	b.	*lievito* 'yeast'			
		z'mälòb	*mét*	*dem*	*lievito*
		DEF=FLOUR(N)	with	DEF	yeast

		zéemegsebòz	*dezuegä*	(DOK_0327)
		together.sieved.N	thereto.give.INF	
		'Add the flour sieved with the yeast'.		

	c.	*dado* ‘stock cube’	
		mét dem gschmolzne «dado»	(DOK_0241)
		‘with the fused stock cube’	

Although these words are clearly Italianisms, they display a certain degree of morphosyntactic integration. In this respect they are not prominently different from other loanwords which are contained in the dictionary like *kanellò* ‘cinnamon’ (also found in compounds like *kannellòpòlver* ‘cinnamon powder’) or *petrolio* ‘oil’ (found in the corpus only in the compound *petroliolampò* ‘oil lamp’). More generally, it is not easy to tease apart established loanwords and pure occasionalism in such a complex contact situation. At any rate, we excluded from lemmatization into the dictionary –besides proper nouns and toponyms– clearly non-integrated loanwords, onomatopoeic expressions, heavily idiomatic expressions, and –in certain cases– entire sentences in German or in a Romance variety. These entries will be placed on a separate level within the platform, where they can be searched and analyzed separately.

3.3. Morpho-lexical normalization and orthographic instability

During the lemmatization process we could pinpoint the orthographic variation without any massive normalization of the variants, which were all inserted and explicitly accounted for in a specific field of the dictionary. In this way we could map the different morpho-lexical and phonological domains for which the graphic variation is more pronounced. In this regard, one important issue is the representation of consonant clusters in which the plosive is preceded by a coronal fricative, which is normally realized as palato-alveolar in any position. The voiceless cluster swings between an overt or a covert coding of the allophonic palato-alveolar feature <scht/st> in words like *wòrscht/wòrst* ‘sausage’ or <schp/sp> in *schpäck/späck* ‘speck’, in which the latter forms correspond to the German orthographic standard, viz. *Wurst* and *Speck*. On the other hand, we already pointed out the difficulties posed by the phoneme /ʒ/, which are mirrored in the different variants of the reflexive pronoun *ŝché* (*sché*, *sche*, *je*), or in the noun *eŝchél* (*eschél*, *ejel*) ‘donkey’.

A further case of instability is provided by the (etymological) voiced or (phonological) voiceless representation of the prefix *b-/p-* attached to verbs beginning with a voiceless obstruent, which is normally realized as a voiceless /p/. This prefix is cognate with the German prefix *be-* found in verbs like *bedecken* ‘to cover’, *behalten* ‘to keep’, etc. Their corresponding verbs in Gressoney display both variants with different token frequencies:

Tab. 1. Verbs prefixed with *b-/p-* in Gressoney; the forms attested in the WKZ are in bold

bchenne	20	***pchenne***	18	‘to admit’
bchime	1	***pchime***	3	‘to take a breather’
bhacke	–	***phacke***	1	‘to grip’
bhälfe	2	***phälfe***	2	‘to make do’

bhefte	–	***phefte***	1	‘to button up’
bhelsò	–	*phelsò*	1	‘to uncover’
bhiete	1	***phiete***	–	‘to protect’
bhoalte	6	***phoalte***	4	‘to keep’
bhouptò	–	***phouptò***	1	‘to assert’
bscheibe	2	***pscheibe***	3	‘to plug’
bschétze	14	***pschétze***	2	‘to protect’
bschisse	1	***pschisse***	1	‘to dupe’
bschloa	1	*pschloa*	–	‘to dirty’
bschnétze	–	*pschnétze*	1	‘to prune’
bséche	9	***pséche***	1	‘to sprinkle’
bsénne	12	*psénne*	4	‘to bethink’
bstelle	3	*pstelle*	–	‘to order’
bstémme	2	*pstémme*	–	‘to determine’
bsueche	3	*psueche*	–	‘to visit’
btecke	2	***ptecke***	19	‘to cover’
btue	–	***ptue***	9	‘to shut’
Tot V = 21	79		71	

It must be added that in a couple of cases the variant displaying the etymological full form of the prefix is more frequent than the form represented in the dictionary, as for instance in the case of *bstémme* where the full form *bestémme* displays 7 occurrences in the corpus. As can be gathered from Tab. 1, there is basically a chaotic distribution. Several tendencies can be observed, however, which partially support the form represented in the dictionary indicated in bold in Tab. 1, e.g. for *ptecke*, *ptue* on the one hand and for *bsénne*, *bschétze* on the other. Nonetheless, the occurring instability is large, as in a couple of cases both forms are in the dictionary, e.g. for *bchime/pchime*, *bschétze/pschétze*, *bséche/pséche*, while in several other cases the form given in the dictionary is less frequent or not found in the corpus and vice versa, e.g. for *pchenne*, *phoalte*, *phiete*, *bschnétze*, *bschisse*.

These examples show the difficulty of the writers in the elaboration of a consistent writing system which should be able to account for competing forces and principles, and in particular the etymological comparison with the German Standard forms, the morphological level requiring a certain sign-homogeneity, and finally the uniform

phonological realization which is expected to potentially reduce any sort of instability or confusion.

3.4. Between word and lemma: the instances of phonological word

In §2.1 we mentioned the sequence of clitic pronouns attached to a finite verbal form like *häts* as a significant case of morpho-phonological variability. In particular, these phenomena call into play the relation between the lexical entries as they are represented in the dictionary and their mapping onto the corpus.[18]

Three aspects are important in this regard. First, the inventory of pronominal forms occurring in this as well as in other German varieties includes stressed and unstressed series which can also radically diverge from each other and display highly misleading homophonies. For instance, the unstressed form of the third person singular pronoun displaying masculine or neuter gender and dative case-marking (4a) merges with the unstressed form of the impersonal pronoun (4b), while their stressed forms are clearly different:

(4)	a.	*ém* 'him, 3SG.M/N.DAT'	(stressed form)
		-mò 'him, 3SG.M/N.DAT'	(unstressed form; cf. MHG *imu*)

	b.	*mò* 'one, IMPERS'	(stressed form)
		-mò 'one, IMPERS'	(unstressed form; cf. German *man*)

Second, the unstressed forms of the personal pronouns displaying nominative case-marking closely follow the verb and precede the unstressed forms of other personal pronouns displaying accusative and/or dative case-marking. They undergo assimilation processes involving the verbal endings with the effect that the boundaries among the different elements composing the verb+clitic group are highly opaque, as shown for instance by the corpus occurrence *hämmone ... gsèd* 'one has seen him' (DOK_0002) which has to be analyzed as *hät=mò=ne* 'has=IMPERS=3SG.M.ACC'.

In addition, peculiar forms are observed whose origin or function is difficult to account for analytically, as for instance the corpus occurrence *heiderdò erfreit* 'you have enjoyed' (DOK_0151), which is likely to be analyzed as *heid=er=dò* 'have.PRES.2PL=2PL.NOM=2PL.ACC/DAT'. In this example, the dental segment *-dò-* is likely to go back to the second plural verbal suffix found in *heid* 'have.2PL' and reanalyzed as part of the second plural pronoun *ou* '2PL.ACC/DAT' on the basis of clitic forms like *ier heid-ò* > *ier heid-dò* > *heid-er-do*, which are found also in other combinations like *wéntschen=dò* 'we wish you, lit. wish.PRES.1PL=2PL.ACC/DAT'.

Third, the disappearance of clear morpho-phonological boundaries between verbal endings and clitic pronouns, combined with the tendency of repeating the stressed form before the inflected verb, was interpreted as a first step towards the grammaticalization of new verbal endings, as shown by the following example in which the stressed pronoun *wier* 'we' precedes the finite verb containing the unstressed form of the pronoun *sibber* 'are.1PL:1PL.NOM' (cf. Giacalone 1989):

(5)	*Oanò*	*éndsché*	*sproach*	*wier*	*sibber*	*némme*	*Walser*

	without	1POSS.F.SG	language(F)	1PL	are.1PL:1PL.NOM	never	Walser
	‘Without our language we are no longer Walser’.						

Such a peculiar picture should give an idea of the complexity of the verb/clitic sequences, which are generally treated as a single word by the local orthography. In the logic underlying our platform, the only solution available would have been to treat these forms as variants of the respective verbs, possibly integrated in a further step into the verbal paradigm. This solution is however unsatisfactory for a number of reasons. First, in spite of their considerable interest these forms would have been difficult to search through the corpus because they would have been hidden in the verbal paradigms. Second, this treatment as part of the verbal paradigm is descriptively inadequate as these clitic pronouns are unlikely to be interpreted as pure suffixal forms as implied by such a morphological approach.

Therefore, we elaborated a different solution which constitutes an additive intermediate level in the database between the occurrences and the lemmas level. This level, which was labeled ‘instances of phonological word’, contains all cases in which clitic(ized) forms occur (including also the cases where prepositions are fused with the article): in this way they can easily be recovered for the analysis. On the other hand, this level makes the connection explicit between the specific occurrence in the corpus and the different lexical entries and word forms involved, substantially increasing the descriptive adequacy of the interface connection between corpus and the lexicon.

As with the single lexical entries, the instances of phonological word collect the whole range of orthographic variants occurring in the corpus. Accordingly, to any verb+clitic group a field ‘Variants’ was assigned, which was compiled during the corpus analysis, e.g. *tuemòne/tuemone* ‘does:IMPERS:3SG.M.ACC/PL.DAT’, *tuemòŝché/tuemòsche/tuemoje* ‘does:IMPERS:REFL, etc. The compilation of this intermediate level has substantially improved the treatment of corpus occurrences which were otherwise difficult to deal with by the lemmatization procedure. On the other hand, it also provides a rich inventory of prosodic words, which again shows the difficulty of the writers in the elaboration of a consistent writing system able to mirror the complexity of spoken forms in stable written correspondents.

4. Corpus linguistics and morphological variation: past participles in Gressoney

All the above-mentioned challenges underlying the creation of our data-base emphasize the granularity of the linguistic data. In this connection, we will now analyze some specific case-studies, starting from an example of variation which can only be accounted for if a quantitative view drives the researcher towards an empirically adequate conclusion.

In Titsch, the variety of Gressoney, the past participles are formed –similarly to Standard German– by means of a simultaneous process of prefixation and suffixation. The former involves the attachment of the prefix *g-* to the verbal stem unless the stem begins with an occlusive, while the latter distinguishes two possibilities, a nasal and dental suffix, respectively *-n* and *-t*, which traditionally characterize etymological strong and weak verbs:

(6)	a.	*bisse* 'to bite' / *bésset*
		fénne 'to find' / *gfònnet*
		éllade 'to invite' / *éngladen* or *éngladet*
		schribe 'to write' / *gschrében* or *gschrébet*
		vergässe 'to forget' / *vergässen* or *vergässet*
		verliere 'to lose' / *verlòren* or *verlòret*

	b.	*fiere* 'to lead' / *gfiert*
		läbe 'to live' / *gläbt*
		teile 'to divide' / *teilt*
		decke 'to cover' / *dackt*
		drécke 'to print' / *dròckt*
		féerbe 'to color' / *gfoarbt*

	c.	*moalò* 'to paint' / *gmoalòt*
		rächnò 'to calculate' / *grächnòt*

The verbs of the 1st class in (6a) reflect etymological strong verbs displaying a nasal suffix, while the other two classes in (6b-c) contain etymological weak verbs. Notice that besides prefixes and suffixes the verbs of the 1st (6a) and of the 2nd (6b) class also display root-vowel alternations which are typical of respectively etymological strong (the traditional ablaut) and weak (the traditional *Rückumlaut* 'backwards umlaut') verbs, while the 3rd class (6c) is completely regular. In keeping with a tendency observed throughout all Germanic languages, the etymological strong verbs of the 1st class either acquired the dental suffix coming from the weak classes or actually swing between the older nasal suffix of the strong macro-class as in *gschrében/gschrébet*, *vergässen/vergässet*, etc. Already in traditional descriptions this variation has been pointed out to be fairly widespread (Bohnenberger 1913: 232, Zürrer 1982: 90). Notice that the acquisition of the weak suffix in the 1st class (6a) does not involve the levelling of the root-vowel alternation. In fact, root-vowel alternation is also fairly widespread in the 2nd class (6b).

While this variation has been treated in the past as purely due to chance, the data drawn from the corpus reveal a well-behaved distribution according to the particular morphosyntactic environment in which the past participles are used. Before illustrating the pattern, however, it must be explained that in Titsch the past participles regularly agree with their morphosyntactic heads when a copula-like construction is found or in adnominal position, paralleling the behavior of adjectives (cf. Gaeta 2018, 2020).[19] Thus, in the BE-perfect (7a), in the BE- (7b), in the COME- (7c) and in the GO-passive (7d), as well as in any adnominal position (7e) inflected past participles are found:

(7)	a.	*Hilde*	*òn*	*Cristina*	*sinn*	*drobèr gsatzt-é*	(D_0010)

		Hilde(F)	and	Cristina(F)	are.3PL	thereon seated-PL	
		‘Hilde and Cristina are seated on that’.					

	b.	*em*	*Lido*	*vòn*	*Venedig*	*sinn*
		in.DEF	Lido	of	Venice	are.3PL

		ufbewart-é	*dschin*	*Reliquie*	(DOK_0002)
		preserve.PST.PTCP-PL	their	remains	
		‘In the Lido of Venice their remains are preserved’.			

	c.	*d’Wiehnachtsboumiéné*	*chéemen*	*kontròlliert-e*	(DOK_0202)
		DEF=CHRISTMAS.TREES	come.3PL	monitor.PST.PTCP-PL	
		‘The Christmas trees are monitored’.			

	d.	*De*	*toufnoamna*	*sin*	*of*	*franzesésch*
		DEF	forenames(M)	are.3PL	up	French

		abkändret-e	*kanget*	(DOK_0014)
		change.PST.PTCP-PL	gone’	
		‘The forenames were changed into French’.		

	e.	*En*	*wònderbar*	*glungn-e*	*oabe*	(DOK_0185)
		INDEF	wonderful	succeed.PST.PTCP-M-SG	evening(M)	
		‘A wonderfully successful evening’.				

In the other constructions no inflection is observed, and in particular in the HAVE-perfect:

(8)	a.	*éndschè*	*Sèndég*	*hät*	*fèr*	*éndsch*	*artòat*	*an*	*paar*
		POSS.1PL	mayor	has	for	1PL.OBL	open.PST.PTCP	INDEF	pair

		butèllè	*wi*	(DOK_0010)
		bottles	wine	
		‘Our mayor has opened for us a couple of bottles of wine’.		

	b.	*aber*	*héibèr*	*véll*	*glachet*	(DOK_0010)
		but	have.1PL	much	laugh.PST.PTCP	

		‘But we laughed a lot’.

It must be added that also in the BE-perfect the agreement of the past participle is scarcely or never observed, especially with certain verbs such as for instance *blibe* ‘to remain’ or *goa* ‘to go’:

(9)	a.	*Uf*	*em*	*obre*	*Platz*	*sinn*
		on	DEF	upper.M.SG	place(M)	are.3PL

		d’Medra	*bim*	*Ronkreschtentsch-Hus*	*gsetzt*	(DOK-0086)
		DEF=MOWERS	at.DEF	R.-house(N)	seat.PST.PTCP	
		‘At the upper place the mowers are seated close to the Ronkreschtentsch-house’.				

	b.	*mengé*	*chénn*	*sinn*	*en*	*de*	*Tache*
		many	children	are.3PL	in	DEF	roof

		én	*kanget*	*òn*	*andre*	*sinn*	*zem*	*hus*	*blébet*	(DOK_0192)
		in	gone	and	others	are.3PL	to.DEF	house	remain.PST.PTCP	
		‘Many children have gone under the roof and others have remained at home’.								

Let us now observe the distribution in the corpus. First, the corpus contains a substantial number of inflected verb forms and in particular of participles. From Tab. 2 we gather that about one quarter of the verbs contained in the dictionary displays a form of the past participle:

Tab. 2. Verbs in the corpus (types)

Verbs not attested in the corpus	1,799	61.0%
Verbs without attested participle	401	13.6%
Verbs displaying weak participles	674	22.9%
Verbs displaying strong participles	28	1.0%
Verbs displaying strong and weak past participles	45	1.5%
Total verbs	2,947	100.0%

Weak participles are clearly dominant across any inflectional class. Strong participles apparently compete with their corresponding weak participles insofar as a similar amount of verbs belonging to the 1st class either displays only the strong form or swings between the two. If we focus only on the 1st class –see (6a) above–, a clear pattern emerges as can be gathered from Tab. 3:

Tab. 3. 1st class verbs in the corpus

	H-Pe	B-Pe	Pa	AdN
a. W, +I	5 / 3.5 57 / 7.9	– –	5 / 9.4 12 / 9.6	7 / 13.7 8 / 7.1
b. {S, +I} & {W, +I}	8 / 5.6 196 / 27.1	1 / 1.4 35 / 6.0	9 / 17.0 32 / 25.6	9 / 17.6 21 / 18.8
c. S, +I	– –	– –	16 / 30.2 18 / 14.4	14 / 27.5 33 / 29.5
d. {S, +I} & {W, –I}	29 / 20.3 140 / 19.4	10 / 14.1 40 / 6.8	19 / 14.1 56 / 6.8	20 / 39.2 49 / 43.8
e. W, –I	101 / 70.6 330 / 45.6	60 / 84.5 512 / 87.2	4 / 7.5 7 / 5.6	1 / 2.0 1 / 0.9
Tot.	143 / 100.0 723 / 100.0	71 / 100.0 587 / 100.0	53 / 100.0 125 / 100.0	51 / 100.0 112 / 100.0

We report in Tab. 3 the past participles which are found in the corpus in the syntactic environments discussed above, namely the HAVE-perfect (H-Pe), the BE-perfect (B-Pe), the different sorts of passive (Pa) and the different adnominal constructions (AdN). Note that the past participles found in the H-Pe are always uninflected, independently of the group. The two rows in each cell contain the figures respectively for the types and for the tokens (including their percentage calculated with regard to the column). The verbs which have completely acquired the weak suffix form the tiny group in (a) displaying a weak inflected form in all contexts, except for the uninflected form in the HAVE-perfect {W, +I}. For instance, the verb *fénne* 'to find' belongs to this group because it only displays weak forms, possibly inflected in the corresponding environments. For the other verb groups of Tab. 3 we observe a crossed distribution. Apart from the about ten verbs of the group (b) in which both strong and weak forms of inflected past participles are found {S, +I} & {W, +I},[20] a consistent picture emerges. When the past participles are inflected the strong suffix is used, as shown by the group in (c) which displays a strong inflected form but lacks any attestation for the cases in which an uninflected form is expected to appear {S, +I}. The weak forms are also predominant in those contexts where the past participle does not display agreement, i.e. in the HAVE- and in the BE-perfect constructions, as shown by the group in (e) displaying a weak uninflected form {W, –I}. Notice that in this group uninflected weak forms are also marginally found in contexts where normally agreement is expected, namely in Pa and AdN. On the other hand, the group in (d) contains those verbs for which a strong inflected form in the agreement contexts is found, while a weak uninflected form appears elsewhere {S, +I} & {W, –I}. In the corpus, no examples are found of a strong participle in a context where no agreement is required: *{S, –I}. Especially this latter finding sheds light on the correct interpretation of the distribution of the allegedly casual variation observed in the literature. Etymological strong verbs were to a limited

extent reassigned to the weak model, adopting the dental suffix in any morphosyntactic environment. However, this is true only for the tiny group of verbs which only displays weak participles independently of agreement as found in the group (a) shown in Tab. 3. For the others, we normally find cases where two different participles, resp. a weak and a strong one, are used with one and the same verb in different and complementary environments, see resp. (10a-b) and (10c-d):

(10)	a.	*wenn*	*Benito*	*Leopold*	*Curtaz ...*	*hät*
		when	Benito	Leopold	Curtaz	has

		éndsch	*gschréb-et*	(DOK_0016)
		us	write.PST.PTCP	
		‘When Benito Leopold Curtaz ... has written to us’.		

	b.	*heintsch ...*	*d’hus-gspònnt-o*	*woll-schtrangn-a*
		have.3PL	DEF=HOUSE-WOVEN-PL	wool-skein-PL

		gwässch-et	(DOK_0295)	
		wash-PST.PTCP		
		‘They have ... washed their home-made wool-skeins’.		

	c.	*al-z*	*éscht*	*kanget*	*gschréb-en-z*	(DOK_0015)
		all-N.SG	is	gone	write-PST.PTCP-N.SG	
		‘Everything has been written’.				

	d.	*d’gròss-ò*	*lougò*	*ésch*
		DEF=BIG-F.SG	laundry(F)	is

		gwässch-n-e	*kanget*	(DOK_0348)
		wash-PST.PTCP-F.SG	gone	
		‘And the big laundry has been washed’.		

Note that this remodeling was also extended to etymological weak verbs where the strong suffix is not expected like for instance *bégleite* ‘to accompany’ (11a-b):

(11)	a.	*D’journalist-e*	*hein*	*désch-é*
		DEF= JOURNALIST-PL	have.3PL	this-F.SG

		familiò	*de*	*ganz*	*vorméttag*	*begleit-et*	(DOK_0430)
		family(F)	DEF	whole	morning	accompany-PST.PTCP	
		'The journalists have accompanied this family the whole morning'.					

	b.	*z'lied ...*	*és*	*vòn*	*der*	*gitarò*	
		DEF=SONG(N)	is	of	DEF	guitar	
		begleit-en-z				*gsid*	(DOK_0202)
		accompany-PST.PTCP-N.SG	been			been	
		'The song has been accompanied by the guitar'.					

Thus, most etymological strong verbs developed two participles, which are respectively found in the different environments. When the latter requires agreement, the strong form is used, while the weak form is used elsewhere. In other words, the group (d) in Tab. 3 represents the vast majority of etymological strong verbs in Gressoney, while the verbs found respectively in groups (c) and (e) lack (or almost lack) attestation for the crucial agreement environments, respectively the group (c) for the non-agreement contexts, and the group (e) for the agreement contexts. At the same time, they are likely to be assigned to the group (d), provided that further empirical evidence – directly elicited from the speakers– is found.
In sum, the corpus investigation allows us to pinpoint a clear distribution of the strong/weak forms which can be summarized by the following table:

Tab. 4. Syncretism in the Titsch past participles

	[+ AGR]	[– AGR]
1st class – strong	*-en*	*-et*
1st class – weak	*-et*	*-et*
elsewhere	*-t*	*-t*

A partial syncretism is observed in the 1st class, which holds true however only for syntactic environments not requiring agreement.
Finally, notice that on the basis of the corpus we can also provide empirical evidence in support of the independence of the syncretism shown in Tab. 4 from the root-vowel alternations mentioned above. In fact, if we consider the possible ablaut types attested in Gressoney, the following picture obtains:

Tab. 5. Strong and weak inflection in the 1st class verbs (types)

Ablaut-type	+S	+S, +W	+W	Tot.
A-1: Ø *lade* 'load' / *gladet*	7 7.4%	19 20.2%	68 72.3%	94 100.0%

A-2: *i/é* *blibe* 'stay' / *blébet*	8 29.6%	4 14.8%	15 55.6%	27 100.0%
A-3: *ie/o* *biete* 'bid' / *bottet*	8 12.5%	5 16.7%	17 56.7%	30 100.0%
A-4: *é/ò* *bénne* 'tie' / *bònnet*	4 12.9%	6 19.4%	21 67.7%	31 100.0%
A-5: *ä/o* *bräche* 'break' / *brochet*	1 7.1%	3 17.6%	13 76.5%	17 100.0%
A-6: residue	– –	8 19.0%	34 81.0%	42 100.0%

Although the distribution is not perfectly linear across the single Ablaut-types, both weak and strong forms are found within any single A-pattern to a comparable extent. Thus, the strong/weak distribution discussed above is apparently independent of any Ablaut-type, and the two types of morphological alternation (i.e. the infixal and the suffixal one) are not connected to each other.

5. Conclusion

The treatment of corpus-linguistic data coming from low-density varieties requires a peculiar approach. While their extension is strongly limited, low-density varieties normally present peculiar properties –such as for instance the orthographic instability accompanied by a substantial lack of normativity– which strongly influence their computability. On the other hand, their remarkable granularity constitutes a crucial factor for grasping their detailed (morpho-)phonological and morphological differences, as well as the strict connection between the oral and written dimension typical of minority varieties. In this perspective, in our platform we tried to develop original solutions able to account for these aspects in an adequate way, to a certain extent integrating them into the lemmatization process. In fact, it is our deep conviction that –in order to treat the peculiar state-of-affairs found in these communities– it is necessary to find a certain balance between opposite forces, namely the requirement of a precise lemmatization and the need of preserving the internal variety of the linguistic data. Exploiting the computational means, we pursued the possibilities opened by multi-layered databases, in which granularity is decomposed and represented on different layers. This allowed us to create a flexible and user-friendly tool which is able to carry out morphological, syntactic and lexical analysis, using refined instruments for querying the corpus, such as for instance the instances of phonological words. On the other hand, the corpus-based analysis helped us to discover unexpected patterns of variation such as those found in the past participles of the verbs which show surprising developments and deserve further and more detailed research.

BIBLIOGRAPHY

Angster M. (2014). "Lingue di minoranza e di maggioranza. 200 anni di lingue straniere a Gressoney (AO)", in V. Porcellana, F. Diémoz (eds.) *Minoranze in mutamento: Etnicità, lingue e processi etnografici nelle valli alpine italiane*. Alessandria: Dell'Orso, 105-121.

Angster M., Bellante M., Cioffi R., Gaeta L. (2017). "I progetti DiWaC e ArchiWals", *in* L. Gaeta, Livio (ed.) *Le isole linguistiche tedescofone in Italia: la situazione attuale e le prospettive future (Workshop, Torino 24 febbraio 2017)*. Special issue of *Bollettino dell'Atlante Linguistico Italiano* 41: 83-94.

Angster M., Cioffi R., Bellante M., Gaeta L. (2020). "Corpora e varietà minoritarie: le isole walser in Italia", *Rivista Italiana di Dialettologia* 44: 107-125.

Angster M., Gaeta L. (2021). "Contact phenomena in the verbal complex: the Walser connection in the Alpine area", *STUF - Language Typology and Universals* 74.1: 73-107.

Antonietti F. (ed.) (2010). *Scrivere tra i Walser. Per un'ortografia delle parlate alemanniche in Italia*. Formazza: Associazione Walser Formazza.

Baron A. & Rayson P. (2008). "VARD 2: A tool for dealing with spelling variation in historical corpora", in *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham: Aston University.

Baroni M., Bernardini S., Ferraresi A., Zanchetta E. (2009). "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora". *Language Resources and Evaluation* 43.3: 209-226.

Bohnenberger K. (1913). *Die Mundart der deutschen Walliser im Heimattal und in den Außenorten*. Frauenfeld: Huber.

Bildhauer F. & Schäfer R. (eds.) (2014). *Proceedings of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014*. Association for Computational Linguistics.

Dal Negro S. (2004). *The Decay of a Language: The Case of a German Dialect in the Italian Alps*. Bern: Peter Lang.

Dal Negro S. & Ciccolone S. (2018). "Il parlato bilingue: italiano e tedesco a contatto in un corpus sudtirolese", *in* F. Bermejo Calleja and P. Katelhön (eds.) *Lingua parlata. Un confronto fra l'italiano e alcune lingue europee*. Berlin: Peter Lang, 385-407.

Dammel A. (2011). *Konjugationsklassenwandel. Prinzipien des Ab-, Um- und Ausbaus verbalflexivischer Allomorphie in germanischen Sprachen*. Berlin / New York: De Gruyter.

Fazzini E. & Cigni C. (2004-). *Vocabolario comparativo dei dialetti walser in Italia*. Alessandria: Edizioni dell'Orso, 2004-ongoing.

Fuhrhop N. (2007). *Zwischen Wort und Syntagma*. Tübingen: Niemeyer.

Gaeta L. (2018). "Im Passiv sprechen in den Alpen". *Sprachwissenschaft* 43.2: 221-250.

Gaeta L. (2020). "Remotivating inflectional classes: an unexpected effect of grammaticalization", *in* B. Drinka (ed.), *Historical Linguistics 2017. Selected papers from the 23rd International Conference on Historical Linguistics, San Antonio, Texas, 31 July – 4 August 2017*. Amsterdam / Philadelphia: John Benjamins, 205-227.

Gaeta L. (in press). “The Observer’s Paradox meets Corpus Linguistics: Written and oral sources for the Walser linguistic islands in Italy”, *in* M. Genesin, G. Hempel, Th. Kahl (eds.) *Endangered linguistic varieties and minorities in Italy and the Balkans*. Wien: Austrian Academy of Sciences.

Gaeta L., M. Angster, M. Bellante & R. Cioffi (2019). “Conservazione e innovazione nelle varietà Walser: i progetti DiWaC e ArchiWals”, *in* R. Rosselli Del Turco (ed.) *Dall'Indeuropeo al Germanico: problemi di linguistica storica (atti del XVIII Seminario avanzato in Filologia Germanica, Torino, 18-20 settembre 2017)*. Alessandria: Dell’Orso, 141-193.

Garner Ph. N., Imseng D. & Meyer Th. (2014). “Automatic speech recognition and translation of a Swiss German dialect: Walliserdeutsch”, in INTERSPEECH-2014: 2118-2122.

Gatto M. (2014). *The Web as Corpus. Theory and practice*. London – New York: Bloomsbury.

Giacalone Ramat A. (1989). “Per una caratterizzazione linguistica e sociolinguistica dell’area Walser”, *in* E. Rizzi (ed.) *Lingua e comunicazione simbolica nella cultura Walser. Atti del VI convegno internazionale di studi Walser*. Anzola d’Ossola, Fondazione E. Monti, 37-66.

Hinskens F. (2011). “Emerging Moroccan and Turkish varieties of Dutch: Ethnolects or ethnic styles?”, *in* F. Kern, M. Selting (eds.) *Ethnic Styles of Speaking in European Metropolitan Areas*. Amsterdam/Philadelphia: John Benjamins, 102-131.

Honnet P. E., A. Popescu-Belis, C. Musat & M. Baeriswyl (2018). “Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German”, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*: 3781-3788.

Iannàccaro G. (2010). “Vita comunitaria e pianificaione linguistica: i Walser”, *in* F. Antonietti (ed.), *Scrivere tra i Walser. Per un'ortografia delle parlate alemanniche in Italia*. Formazza: Associazione Walser Formazza, 15-28.

Mari C. J. & Mooney D. (2017). *Creating Orthographies for Endangered Languages.* Cambridge: Cambridge University Press.

Jutta R., Mörth K. & Ďurčo M. (2013). “Linguistic Variation In The Austrian Media Corpus. Dealing With The Challenges Of Large Amounts Of Data”, *Procedia - Social and Behavioral Sciences* 95: 111-115.

Kilgarriff A. & Grefenstette G. (2003). “Introduction to the Special Issue on the Web as Corpus”, *Computational Linguistics* 29.3: 333-47.

Ljubešić N. & Erjavec T. (2011). “hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene”, *in* I. Habernal and V. Matousek (eds.) *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011*. Springer, 395-402.

Ljubešić N. & Klubička F. (2014). “{bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian”, *in* F. Bildhauer and R. Schäfer (eds.) *Proceedings of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014*. Association for Computational Linguistics, 29-35.

Maxwell M. & Hughes B. (2006). “Frontiers in Linguistic Annotation for Lower-Density Languages”, in *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*. Association for Computational Linguistics: 29-37.

Nagy N. (2017). “Documenting variation in (endangered) heritage languages: How and why”, *in* K. A. Hildebrandt, C. Jany and W. Silva (eds.) *Documenting Variation in Endangered Languages*. Honolulu: University of Hawai'i Press, 33-64.

O’Keeffe A. & McCarthy M. (eds.) (2010). *The Routledge Handbook of Corpus Linguistics*. London – New Tork: Routledge.

Samardžic T., Scherrer Y. & Glaser E. (2015). "Normalising orthographic and dialectal variants for the automatic processing of Swiss German", in *Proceedings of the 7th Language and Technology Conference*. Poznan.

Samardžić T., Scherrer Y. & Glaser E. (2016). "ArchiMob – a corpus of spoken Swiss German", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*: 4061-4066.

Scherrer Y. & Ljubešić N. (2016). "Automatic Normalisation of the Swiss German ArchiMob corpus using character-level machine translation", in *Proceedings of the 13th Conference on Natural Language Processing (KONVENS) 2016*: 248-255.

Schuppler B., Hagmueller M., Morales-Cordovilla J. A. & Pessentheiner H. (2014). "GRASS: the Graz corpus of Read And Spontaneous Speech", in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*: 1465-1470.

SI = *Schweizerisches Idiotikon. Wörterbuch der schweizerdeutschen Sprache*. Frauenfeld 1881-

WKZ = Centro Studi e Cultura Walser / Walser Kulturzentrum (ed.) 1988. *Greschòneytitsch. Italiano/ Titsch*. Quart: Musumeci.

Zürrer P. (1982). *Wörterbuch der Mundart von Gressoney. Mit einer Einführung in die Sprachsituation und einem grammatischen Abriß*. Frauenfeld: Huber.

Zürrer P. (2009). *Sprachkontakt in Walser Dialekten: Gressoney Und Issime Im Aostatal (Italien)* (Zeitschrift für Dialektologie Und Linguistik – Beihefte). Stuttgart: Steiner.

NOTES

***.** * The paper results from the joint work of all co-authors. However, for academic purposes, Livio Gaeta carries the responsibility for the sections 1, 3.2, 4 and 5, Marco Angster for the sections 1.1, 1.2, 2 and 3.4, Raffaele Cioffi for the sections 1.3, 3 and 3.3, Marco Bellante for the section 3.1. We thank two anonymous reviewers for their comments and remarks.

1. On the recent success of initiatives devoted to the construction of big corpora made out of texts drawn from the Web, see Kilgariff, Grefenstette (2003), Baroni et al. (2009), Gatto (2014).

2. Most varieties spoken in the territories of Ex-Yugoslavia display Web-based corpora of a size comparable to those of English, German, Italian, etc. (cf. Ljubešić, Erjavec 2011, Ljubešić, Klubička 2014).

3. For Swiss German see the ArchiMob Corpus: https://www.spur.uzh.ch/en/departments/research/textgroup/ArchiMob.html. For Austrian German see among others: the GRASS (Graz Corpus of Read and Spontaneous Speech, Schuppler et al. 2014) and the Austrian Media Corpus (Jutta et al. 2013): https://www.oeaw.ac.at/acdh/tools/amc-austria-media-corpus/.

4. See the multilingual corpus Kontatto in which the Italian/German contact in Alto Adige is documented (Dal Negro, Ciccolone 2018): https://kontatti.projects.unibz.it/before-kontatti/.

5. As for minority languages like Breton, see the project Breton Text Corpora accessible on the platform of SketchEngine (https://www.sketchengine.eu/corpora-and-languages/breton-text-corpora/), as well as the corpus contained in the Leipzig Corpora Collection (https://corpora.uni-leipzig.de/en?corpusId=bre_wikipedia_2007). Furthermore, on endangered Slavic minorities in non-Slavic speaking countries see the activities of Lacito (https://lacito.vjf.cnrs.fr/partenariat/euroslav/index_en.htm).

6. See the project "The roots of ethnolects. An experimental comparative study" (Hinskens 2011) centering on the emergence of two ethnolects among Turkish and Moroccan young people in Amsterdam and Nijmegen.

7. See the corpus HerLD ("Heritage Language Documentation Corpus"; Nagy 2017) developed within the project "Heritage Language Variation and Change in Toronto" which collects conversations in 10 different heritage languages, including Italian and Franco-Provençal varieties (from Faeto): http://projects.chass.utoronto.ca/ngn/HLVC/0_0_home.php.

8. See the link https://www.cl.uzh.ch/en/texttechnologies/research/Low-Resource-NLP/LORELAI.html.

9. Currently, the platform is under rebuilding: every scholar interested to have access to the documentation of the project and to the corpora of Gressoney and Issime are invited to contact the members of the research group (https://www.climalp.org/index.php/contatti/).

10. Under metadata we intend the set of descriptive data relating to a document uploaded into an archive. They are a semantic system providing the background of a document's content (descriptive and structural metadata), as well as the context in which it appears (administrative metadata). The metadata allow a straightforward organization and management of the documents, a quicker retrieval of the information and an easier interoperability of the managing system and of the archive (see in this regard the Dublin Core Metadata Initiative: https://dublincore.org/).

11. We generally adopt the orthographic norms used in the dictionary in which <é>, <ä> and <ò> roughly correspond respectively to [ɪ], [æ] and [ʊ] while vowel sequences like <ie>, <ée>, etc. correspond to true (falling) diphthongs: [i̯ɐ̯], [ɪ̯ɐ̯], etc. It must be added that the texts acquired in our data-base do not always follow these orthographic norms, also because to a large extent they have been written before their adoption (cf. Angster et al. 2017 for discussion).

12. As is well known, the separability and its orthographic coding are hotly-debated issues also in the Standard German variety. See the discussion in Fuhrhop (2007).

13. The token number available to us was too limited for attempting the application of techniques of automatic recognition and lemmatization (as well as of orthographic normalization) adopted within the Swiss-German projects mentioned above (Garner et al. 2014, Honnet et al. 2018, Samardzić et al. 2015).

14. License deposited on the 21.11.2019 with priority number 102019000021837.

15. One anonymous reviewer contends that *bieche* might be related to MHG *bæhen* 'to heat', cf. Modern German *bähen* 'to roast'. We see two problems with this view. First, MHG *bæhen* is also found in Alemannic forms like *bājen* 'to braise' (cf. SI, s.v.). Second, the Titsch diphthong /ie/ found in *bieche* is the normal outcome of the umlauted OHG *ū* or *uo* resulting into MHG *iu* and *üe* as shown by OHG *fūhten*, *bluoen*, MHD *viuhten*, *blüe(je)n* > *(a-)fiechte* 'to dampen', *blieche* 'to blossom', etc., while MHG *æ* gives normally rise to /ɪ/: MHG *blæjen* > *bléche* 'to swell', *kræ(je)n* > *chréche* 'to crow', etc. (cf. Zürrer 1982: 75-76).

16. In the paper, examples drawn from the corpus are marked with an abbreviation identifying the genre (DOK for 'documents', BEL for 'fiction', etc.) followed by the corresponding text number. The examples are glossed according to the Leipzig Glossing Rules.

17. In this case, in the dictionary the derivative is already found in the compound *bluetvergéftòng* 'blood poisoning' but the corpus provides the attestation for the derivative alone.

18. In the corpus several proclitic elements are also found, for instance the form *z'* which can serve as definite article or preposition. They were generally treated as separate tokens connected to their respective lexical items.

19. In this connection, the role of the resultative/stative value of copula-like constructions resembling the value of typical predicative adjectives has been repeatedly emphasized in the literature, while the HAVE-perfect construction rather conveys tense (cf. Dammel 2011: 249ff. for a discussion). This distinction has especially been discussed for the different past participles found with *Rückumlaut* verbs (see (6b) above) like *wentä* 'to turn' in Bosco Gurin, where the root-vowel alternation is only found in the inflected form: *wentä* 'turned' / *gwant-s*. However, while the difference in tempo-aspectual terms between the two constructions might capture the

diachronic origin of the actual distribution, this cannot hold for the current situation, insofar as copula-like constructions do not necessarily convey a resultative/stative value, as shown by the example (11) below containing the atelic verb *bégleite* 'to accompany'. In this paper we will generically refer to the morphosyntactic environment without embarking into such a complex distinction, which requires further and more detailed research.

20. Note that the only verb belonging to group (b) in which forms of the B-Pe occur is *blibe* 'to remain': in this case, however, only uninflected weak forms are found, while strong inflected forms are found in adnominal position.

ABSTRACTS

Corpus linguistics grew up in the domain of written (and literary) varieties, while its recent methodological revolution is due to the computer-assisted capacity of elaborating massive amounts of text data. On the other hand, the so-called 'low-density varieties', including spoken varieties as well as varieties spoken in minority communities, have been confined to a rather marginal role. Among others, this is due to the technical problems connected to the scarce degree of normalization in linguistic –including graphemic– terms, as well as to the scarcity of language resources for automatic processing. In this paper, we will exploit the possibilities opened by corpus linguistics for acquiring and analyzing the textual patrimony of the Walser German communities of Piedmont and Aosta Valley. The varieties of Highest Alemannic spoken there, dramatically exposed to language decay, provide a limited but significant amount of data, which is accompanied by a substantial lexical documentation due to the active collaboration of the speakers' communities in collecting and compiling local dictionaries. After briefly introducing our archive and discussing the peculiar solutions adopted for the construction of the platform, we will also present corpus-based morphological investigations regarding the representation of verbal prefixes, of the clitic group, as well as of the inflectional behaviour of verb classes.

La linguistique de corpus s'est développée dans le cadre des variétés écrites (et littéraires), tandis que sa récente révolution méthodologique est due à la capacité assistée par ordinateur d'élaborer des quantités massives de données textuelles. D'autre part, les variétés dites 'à faible densité' comprenant les variétés parlées ainsi que les variétés parlées dans les communautés minoritaires, ont été confinées à un rôle plutôt marginal. Cela est dû, entre autres, aux problèmes techniques liés au faible degré de normalisation en termes linguistiques, y compris graphémiques, de ces variétés ainsi qu'à la rareté des ressources linguistiques pour leur traitement automatique. Dans cet article, nous allons exploiter les possibilités offertes par la linguistique de corpus pour acquérir et analyser le patrimoine textuel des communautés allemandes Walser du Piémont et de la Vallée d'Aoste. Les variétés d'alémanique supérieur qui y sont parlées, dramatiquement exposées à des processus avancés de décadence linguistique, fournissent une quantité limitée mais significative de données, qui s'accompagne d'une documentation lexicale substantielle due à la collaboration active des communautés dans la collecte et la compilation de dictionnaires locaux. Après une brève présentation de nos archives et la discussion des solutions particulières adoptées pour la construction de la plate-forme, nous présenterons également des investigations morphologiques basées sur corpus concernant la

représentation des préfixes verbaux, du groupe clitique, ainsi que du comportement flexionnel des classes de verbes.

INDEX

Mots-clés: patrimoine culturel, langues minoritaires, documentation linguistique, préfixes verbaux, pronoms clitiques, classes de flexion verbale
Keywords: cultural heritage, minority languages, language documentation, verb prefixes, clitic pronouns, inflectional verb classes

AUTHORS

LIVIO GAETA

University of Turin

MARCO ANGSTER

University of Zadar

RAFFAELE CIOFFI

University of Turin

MARCO BELLANTE

University of Turin