AperTO - Archivio Istituzionale Open Access dell'Università di Torino

“Going standard” on a blank page. A corpus-based approach to the written varieties of the Italian Western Alps minorities (Occitan, Francoprovençal and Walser)

This is a pre print version of the following article:

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/1930831 since 2023-09-07T16:45:14Z

Publisher:

John Benjamins

Published version:

DOI:10.1075/scl.110.04rai

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

17 October 2023

"Going Standard" on a Blank Page. A Corpus-Based Approach to the Written Varieties of the Italian Western Alps Minorities (Occitan, Francoprovençal and Walser)[1]

Gianmario Raimondi[++], Marco Angster[+++], Marco Bellante[+], Paolo Benedetto Mas[++], Raffaele Cioffi[+], Livio Gaeta[+], Aline Pons[+], Matteo Rivoira[+]

Abstract

The paper investigates non-standard languages, i.e., those which are dialectal, non-standardised – or standardised to a very limited extent, represented by the local linguistic varieties that populate the Italian Western Alps. Despite the fact that these have almost exclusively existed as spoken languages throughout their history, our particular aim is to discuss methods and problems raised by the investigation of the written corpora of these varieties from a corpus linguistics perspective. This is especially challenging because corpus linguistics usually employs methods and standards elaborated for standard(ised) written varieties. Focusing on Occitan and Francoprovençal varieties, it is shown that the different historical backgrounds of the two

[1] This contribution was collectively conceived but its parts are separately attributed, for academic purpose, to the following authors: *Introduction* and § 4: Livio Gaeta, Gianmario Raimondi and Matteo Rivoira; § 1: Gianmario Raimondi; § 2: Marco Angster, Marco Bellante and Raffaele Cioffi; § 3.1: Matteo Rivoira; § 3.2. and § 3.3. Paolo Benedetto Mas (FP) and Aline Pons (OC). As for the affiliations: + = University of Turin, ++ = University of Aosta Valley, +++ = University of Zara.

languages also have an impact on their speakers' attitude towards standardisation and on how texts are produced and accordingly made accessible for corpus linguistics methods.

Keywords

corpus linguistics, linguistic minorities, standardization, linguistic islands, language contact

1. Introduction

This contribution is somehow *peripheral* to the general scope of this volume, for two good reasons. First, we will focus on largely non-standardised languages, spoken in the Italian Western Alps. As peripheral members of wider linguistic groups, both Romance (Occitan and Francoprovençal, on which the contribution will duly focus) and Germanic (Walser), these varieties essentially defy the adoption of a shared standardised norm, even when (as for Occitan) such a norm is available. The second reason is that the corpora in question are based exclusively on *written* instances of said varieties. This could initially appear awkward, as they have been almost exclusively *spoken* languages throughout their history. Hence the idea of the *blank page* in the title: this, to our mind, expresses the double challenge we want to focus on in this paper. First, the writers of such varieties are challenged to cope with the lack (or uneven diffusion) of an endogenous, stable and reliable norm for transposing their spoken languages into a written form. Second, from the perspective of a corpus-based linguistics, the challenge consists in how to combine methods and standards typically employed with standard(ised) written varieties with the methodological issues raised by written corpora issued from these largely non-standardised varieties.

The paper is structured as follows: in Section 2 we offer an overview of the minority languages spoken in the Western Alps, particularly focusing on the Romance varieties which are currently the object of the CLiMAlp project (see the project website and Gaeta et al. 2022 for a brief presentation), briefly presented in Section 3. In Section 4 the problems raised by the employment of corpus linguistics methods with Occitan and Francoprovençal texts are discussed in detail, showing that their different linguistic landscapes also have an impact on how texts are produced and accordingly made accessible for corpus linguistics methods. The final Section 5 draws the conclusion.

2. The Western Alps minority languages: an overview

As a natural consequence of its borderland nature, the Italian Western Alps is a typical example of a *linguistic crossroads*. Historically, two different linguistic boundaries are observed, respectively between Germanic and Romance languages on the North-South geographical axis and between Gallo-Romance and Italo-Romance on the West-East axis. Their convergence has resulted in the formation of three linguistic alpine minorities within the north-western regions of Italy, i.e., Piedmont and Aosta Valley (see Figure 1). Firstly, the Cisalpine Occitan (= OC) minority is found in the southern and central Piedmont highland valleys, from the head of Susa Valley southward. Secondly, the Cisalpine Francoprovençal (= FP) minority populates the northern Piedmont valleys and the Aosta Valley. Both minorities belong to the Gallo-Romance group. Finally, the Walser (= W) minority is found in isolated and small villages surrounding the Mont Rose, in particular the Lys Valley in the Aosta Valley region and the High Sesia and Ossola Valleys in Piedmont. These varieties belong to the so-called *Highest Alemannic* group of the West-Germanic family, namely the *Southern Walser*, which date back to medieval migrations of settlers originating from the northern part of the Swiss Valais, or *Kanton Wallis*.[2]

[2] For general historical and sociolinguistic surveys, see Sumien 2006, Oliviéri & Sauzet 2016, Rivoira 2016 and Regis 2020 (Occitan), Favre 2010, Kristol 2016 and Benedetto Mas & Regis 2022 (Franco-Provençal), Zürrer 2009 (Walser dialects). According to the *Mountain Linguistics* principles (Nichols 2015: 262-270), the whole Italian Western Alps can be defined as a *Central Mountain Crest* area, presenting typical phenomena such as high linguistic diversity, high inner micro-variation, uphill sociolinguistic isolation, asymmetrical

@@ Insert FIG01.jpg

The three linguistic minorities were only officially recognised by the Italian state at the end of the 20th century thanks to the Law 482/1999 on *Historical Linguistic Minorities*, although already in the Republican constitution of 1948 their use is not forbidden as it was during the fascist period. The three linguistic groups share a historical condition of *dialectal diglossia*, without showing any tendency towards spontaneous koineization, while we constantly observe the recourse to the use of a "Dachsprache" (Kloss 1967) or roof-language at the high/written levels (mainly Latin until the 16th century, especially for the Romance varieties; later Italian or French, plus German for the Walser communities; see Angster & Gaeta 2021) and at the middle/spoken levels (Italian/French; also lowland Piedmont dialects, until the half of 20th century) of the repertory. This condition has also determined in the speakers a low degree of self-consciousness of their ancestral linguistic identity, which in many cases was only developed in the second half of the 20th century. This condition has also determined in the speakers a low degree of self-consciousness of their ancestral linguistic identity, which in many cases was only developed in the second half of the 20th century.

As for geolinguistic features, while Walser communities can be seen as fairly independent *linguistic islands*, the two Romance minorities represent an

vertical bilingualism, *Burushaski* and *leapfrogging* distribution of language spread (see also Urban 2020 and Cioffi et al. 2021).

extension toward the East of the Gallo-Romance dialectal *continuum*, dating back to the very formation of Romance linguistic groups. Both the linguistic areas of these *Cisalpine Gallo-Roman* varieties exhibit internal subdivisions. Traditionally, the OC continuum, belonging as a whole to the Vivaro-Alpine section of general Occitan (Oliviéri & Sauzet 2016: 320), is furtherly divided into Northern (Susa, Chisone and most of the Waldensian Valleys), Central (around the Monviso: Po, Varaita, Maira, Grana and Stura Valleys) and Southern dialects (Gesso and Vermenagna Valleys). The FP area is generally split between the two regions of Piedmont and Aosta Valley, but finer distinctions can be made, for example between the varieties of High and Low Aosta Valley (Raimondi 2020: 114-115).[3]

3. The CLiMAlp project

CLiMAlp (*Corpus Linguistics Meets Alpine Cultural Heritage*) is a partnership between the Universities of Turin and Aosta Valley. This initiative aims to investigate the Germanic (W) and Romance minorities (OC and FP) of Piedmont and Aosta Valley, applying the methods and

[3] Recent geolinguistic distinctions focus on an opposition between *Inalpine* and properly *Cisalpine* dialects, the former referring to the varieties spoken in the Italian inland of the more frequented mountain passes (as Montgenèvre and Mont Cenis in Piedmont, Little and Grand Saint Bernard in Aosta Valley), more subject to the influence of neighbouring transalpine dialects and less subject to that of Piedmontese (see Garnier 2020, about OC but appliable also to FP).

technologies of corpus linguistics to the written texts produced by their speakers' communities. The texts are varied in nature, ranging from parish bulletins, journals issued by the local cultural centres, as well as books collected for special occasions on specific subjects, for example cookbooks, ethnological materials, and so on. Its expected outcome is a series of web databases related to the different languages, which will facilitate further corpus-based or corpus-driven investigations.[4]

The present multilingual nature of the project is actually the result of the expansion of earlier projects. In particular, at the base of CLiMAlp there are two previous research projects, DiWaC and ArchiWals (Angster et al. 2017) focused on five Walser communities of Aosta Valley (Gressoney and Issime) and Piedmont (Formazza, Rimella, Alagna).

CLiMAlp represents an application of the previous methods to varieties that are different in many aspects. Apart from the obvious genealogical and typological differences, the German and Romance minority languages in question differ firstly in the dimension of the text corpora historically produced and potentially available, which is dramatically higher for Romance varieties, partly due to their geographic extension.[5] Even more significantly,

[4] The three-year project (2019-2022) is financed by the Italian Ministry of Universities, within the initiative PRIN 2017-*Progetti di Rilevante Interesse Nazionale*.

[5] For the observed Italian context, the estimated number of speakers (OC: 15–20 000, FP: 35–45 000) and municipalities (OC: at least 68, FP: at least 122; Regis 2020 and Benedetto Mas & Regis 2022 for Piedmont, Raimondi 2020: 111 for Aosta Valley) of the Romance area are in fact not comparable with the Walser ones. The written production of the latter is, moreover, more recent.

unlike the W area (which consists of independent, albeit closely related, linguistic islands scattered across several valleys), the OC and FP local varieties are actually part of two well-defined dialectal *continua*, each of which can potentially be referred to as a *model* language, namely two general converging varieties resulting from partial processes of koineization.

This has introduced two new issues for the CLiMAlp project: on the one side is the *diatopic variation* within a single dialectal *continuum*, which is huge both in OC and in FP areas; and on the other, the evaluation, within the Romance corpora, of the emerging *standardisation solutions*, possibly balancing between the aspiration towards *supradialectal* models (such as those implied by the various OC or FP linguistic and orthographic standards; see section 4.1) and the use of more local solutions or of mere *idiolectal* elaborations.

The CLiMAlp platform is designed for the treatment of so-called *low-density varieties* (for which electronic resources are scarce; see Maxwell & Hughes 2006) and for managing their high degree of *granularity*, i.e. their inherent complexity regarding such aspects as transcription, metadata and annotation (Gaeta et al. 2022).[6]

[6] *Metadata* refers to 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 allows for straightforward organisation 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). *Annotation* refers to the enrichment of the text stored in the corpus by means of detailed linguistic information concerning the grammatical class of the words, the morphosyntactic environment, etc.

@@ Insert FIG02.jpg

Its multi-layered structure (Figure 2) is conceptually contained within the wider context of a two layered structure, consisting of an implementable textual corpus and a dictionary for each language. The Dictionary, which can be manually enriched with new lexical types, is the grammatically annotated (Part of Speech, basic meaning in Italian and French, and others) lexical grid which permits the partial automatic recognition and lemmatisation of the tokens occurring in the texts.

As for the type-token relation, the lemmatisation procedure provides the possibility to manage both *type inflection* and *type variation* (as, respectively, *vèyen* ‘they see’ > *vére* ‘to see’ and *aoura* > *aouva* ‘hour’ in Figure 3), the latter referring to the variation (mostly graphic or phonetic) which is typical in low-density languages.

@@ Insert FIG03.jpg

While in the previous stage of the project, which was focused only on Walser linguistic islands, in which this variation was designed to be managed only within the limits of single varieties, the introduction of OC and FP in the project enforced some reconsiderations. In fact, the *one Dictionary to one Corpus* relation previously adopted for W didn’t perfectly match with the geolinguistic and sociolinguistic status of minority languages spread through a dialectal *continuum* which is hugely diversified, such as the two Romance ones. In the cases of OC and FP, the probable prospect was rather a *one*

Dictionary to many Corpuses relation, as the alternative (one Dictionary/Corpus for each OC and FP community) was obviously not viable. Despite the increased set of issues that this solution was likely to entail, the Romance section of CLiMAlp accepted the challenge, approaching it with two different strategies, as we will see.

4. The "corpus-based" approach: Romance languages

The construction of the Romance databases was the testing ground for the newly introduced (and above mentioned) issues regarding diatopic variation and normalisation options. This preparatory stage, which structurally underpins the construction of the database, was confronted with two main choices: (1) the standards to adopt for the first upload of the base dictionary, which involves both the choice of the lemmas and the choice of the written standard; then (2) the populating strategies concerning the texts to progressively submit to the machine-learning process.

In the definition of the process, the choice of the orthographic standard for the two (OC and FP) dictionaries was quite naturally the first aspect to be discussed, keeping in mind the different traditions which characterise the two languages (see Section 4.1) but also the availability of lexicographic instruments for the two areas and (in the prospect of corpora construction), the consistency of the texts corpus available for the different language communities (see Section 4.2).

The present state of the two databases and their limited quantitative consistency (OC: 18,503 tokens; FP: 8,596 tokens) obviously doesn't currently allow for any satisfactory data-driven approaches. Nevertheless, some interesting evaluations can already be made, mostly with respect to the observation of the platform's *learning process* in this initial stage, and the responses given by the databases to some simple *automatic Recognition Tests* (RT) focused on independent variables such as graphic, diatopic and textual genre (see Section 3.3).

4.1 Written OC and FP standards

The Cisalpine Gallo-Roman area is characterised by plenty of modern written standards,[7] which are different in both the moment and cultural context of their original elaboration, as well as for their diffusion. In addition to being writing models, these standards may also be regarded as implicit answers to the issue of linguistic normalisation of such uneven ensembles of spoken varieties and, from this point of view, OC and FP areas have followed different pathways.

[7] Regarding the past, the medieval *scripta* of Waldensian Occitan-speaking communities (Borghi Cedrini 2017) is probably the only outstanding example of a local writing norm elaboration.

In the OC area, the writing of spoken varieties begins in the second half of 20th century, alongside the political claims arising from the local *Occitanist* movement. In this cultural climate, since the 1970s, there have been two contending writing systems, distinguished by graphic solutions, historical background and related linguistic policy.

The first (referred to as *Concordata* or *dell'Escolo dòu Po*) was based on an expanded version of the Provençal *Mistralian* writing, in order to render the whole inventory of phonemes resulting from the analysis of the OC varieties, independently from *etymological* considerations. The *Concordata* writing system was further developed in the 1980s by the linguist Arturo Genre and frequently adopted by scholars (but also used in the written production of everyday speakers), also in the Piedmont FP area (Benedetto Mas & Pons 2016).

The second was the *Classical* (or *Alibertina*) writing, based on the literary Occitan of Languedoc and originally designed by Louis Alibert in 1935.[8] In this writing norm (conceived as a *unifying*, etymological and archaic-oriented one), each grapheme may correspond to different phonological realisations in different local varieties. The regular phonological correspondences between the different dialects, as well as the regularisation of inflectional morphology, are obtained via a strict application of the Weinreichian *diasystematic*

[8] This writing norm was progressively adapted to the main Occitan varieties (Provençal, Gascon, Northern Occitan). On this process see Regis & Rivoira 2016: 268 and Rivoira 2021: 141-142.

approach of structural linguistics. These operations result in an idealised writing model, of which the most accomplished realisation is found in the modern literary Occitan of Languedoc. A driving impulse for the use of this writing system for Cisalpine varieties was provided by its complete and coherent exposition in the introduction of DOC (2008), which also contained a set of morpho-syntactic and lexical regularisation proposals aimed at sketching out the ideal variety to be assumed as a local standard. This ideal variety proved to match up, to a large extent, with those of the Central Cisalpine area.

As for the FP Cisalpine area, the first comprehensive writing model is a subregional one: it comes from Aosta Valley and is closely connected with the *Felibrige* cultural climate of the early 20th century. The writing norm (referred to as *Cerlogne writing*), which was created by the abbot Jean-Baptiste Cerlogne and adopted in his literary and grammar opus (Cerlogne 1907), was based on the urban and High Valley varieties and is still used today for spontaneous written productions by local amateurs. This writing system served then in the 1980s as a base for the elaboration of a new one, which was promoted in particular by the Swiss linguist Ernst Schüle. After being refined by the work of a committee of scholars and speakers from different FP-speaking areas (Schüle 1992), this writing norm was adopted by the *BREL-Bureau Régional pour l'Ethnologie et la Linguistique* (the Aosta regional institute for FP promotion; hence the short name of *BREL writing* *system*) and has been promoted via its publishing and teaching activities up to the present day.

Although based on French orthography in its general features, this phonological writing system, whose stated objective is to permit to every speaker "to write his own local *patois* and to read that of others" (PatoisVdA, *Grafia*; our translation) abides by the principle of the biunivocal correspondence between graphic symbols and phonological values.

Outside Italy's borders, other writing models for Francoprovençal exist, though these are of little or no use in the Italian FP area. The so-called *Graphie de Conflans* is another phonological writing norm based on French orthography, elaborated for the Savoy area with recourse to supplementary phonological devices (stress indication, use of ‹k› for /k/, etc.). Last to be mentioned, also for its lack of success, the ORB writing norm (Stich et al. 2003) was proposed as a supradialectal standardised model for the whole Francoprovençal speaking area, with the objective of permitting the written inter-comprehension between varieties which are not mutually intelligible at the spoken level.[9]

For the constitution of the two corpora, the discussion about the writing system to be adopted led to two different methodological solutions. The *Normalised* (OC) and the *BREL* (FP) writing systems (which were chosen for

[9] See also Kristol 2016: 350 for the *Graphie commune pour les patois valaisans* and for his criticism about ORB writing system.

the dictionaries and the corpuses' first phase of population; see Sections 4.2 and 4.3), in fact occupy different slots in the scale proposed by Iannàccaro & Dell'Aquila (2008: 315-318) in their theoretic analysis of dialectal writing norms, mainly based on their variable tendency towards normalisation. It is fair to place the former among the *Polynomic Writings*, which are characterised by a normalisation process which touches on all language levels (including the morphological) and proposes a unique written word form for various actual spoken expressions. On the other hand, BREL's writing system was originally conceived as a simple method to encode the locally spoken varieties into a written form ("grafia dialettale riflessa", i.e. reflected dialectal spelling), mainly concerned with phonetic issues. For this reason, this spelling can at best be classified as a locally elaborated speech-to-text writing system, without any serious attempt at normalisation regarding morphological and lexical aspects.

4.2 Populating strategies

4.2.1 The Dictionaries

Occitan. DOC 2008, the selected lexicographic instrument in *Normalised* writing system for the OC Basic Dictionary upload "collects the Occitan translations of about 10,000 Italian words" (DOC 2008: 5), 8,000 of which are taken from a frequency vocabulary of Italian, the rest being the result of the integration with words allegedly as “of not very high frequency” but considered to have outstanding conceptual relevance for the specific alpine context. The proposed Occitan lemmas come from the screening of previous scholarly works about the OC varieties (grammars and vocabularies) and from field enquiries purposely conducted for the publication.

The upload of the Basic Dictionary produced an inventory of 10,391 lemmas, plus almost 2,000 orthographic or phonetic variants, in some cases two or even three forms for a single lemma. For example, the lemma *cognom* ‘family name’ (an evident loanword from Italian, observed here only regarding the phonetic level of its adaptation) presents *cognòm*, *conhom* and *conhòm* as variants. In this case, while the graphemes ‹o› and ‹ò› are introduced to render the phonetic alternation between [u] and [o], the alternation between ‹gn› and ‹nh› does not correspond to different pronunciations, but only to different orthographic traditions.

Francoprovençal. Considering the huge linguistic variation of the inter-regional (Piedmont and Aosta Valley) FP domain and the absence of a supralocal lexicographic instrument, the fundamental choice for FP was to focus initially on the Aosta Valley varieties, considering: (1) their relative linguistic homogeneity; (2) their coverage under a shared writing system (the above mentioned *BREL*); (3) a consistent and well distributed (both geographically and in terms of genres) textual production. Nevertheless, this choice turned out to conflict with the absence of a reference dictionary capable of satisfying the two conditions of being written in the standard writing and, at the same time, being representative of the whole regional domain. The most complete and recent dictionary for Aosta Valley's *patois* (Chenal & Vautherin 1997) uses the traditional *Cerlogne* writing system.

For this reason, the initial content of the Basic Dictionary was implemented via a selectively reduced word set, based on the 1,584 lemmas drawn up in standard writing on the *BREL* site for the *patois* of the regional chief town, Aosta (PatoisVdA, *Glossari*). This selected set includes lemmas that largely belong to the basic vocabulary, though no explicit selection criteria are given, but it almost entirely lacks words belonging to *grammatical* parts of speech (articles, pronouns, adverbs and other determiners). The local dictionaries set up by the *BREL* (one for each of the 71 Francoprovençal speaking municipalities of the region), were in fact compiled starting with the heritage vocabulary connected to local culture (e.g., agriculture and farming), and are still in progress.

4.2.2 The Corpora

*Occitan***.** The OC corpus population was implemented with recourse to a selection of texts belonging to different genres and written by authors from different zones of the linguistic area in *Normalised* writing system. Along with a majority of texts from the Central Valleys, which are closer to the linguistic standard adopted by DOC 2008, documents from Chisone and Susa Valleys and Val Vermenagna (respectively North and South peripheral varieties) were also selected. As for the genres, the corpus includes original literary prose, poetry, and a drama, as well as translations of excerpts from world literature and articles on science, history, or current affairs. At the end of the first phase of the populating process, the OC corpus reached 18,503 tokens. After excluding from this count the inflected forms of already existing lemmas (which were tagged with the pertinent detailed grammatical information and reduced to the corresponding lemma), the manual screening produced 801 new lemmas (+7.7%), in addition to almost 500 orthographic or phonetic variants.[10] The normalised *New Lemma/Token ratio* (NL/T*100: number of new lemmas added for every 100 tokens processed), which essentially shows the Basic Dictionary *response* to the submitted corpus, is thus 4.3 (4.3 new lemmas every 100 tokens) for this first phase of upload.

Among these new lemmas, we find terms marked on the one hand by diatopic variation, like *cocho* ‘rabbit’ found only in Val Vermenagna (see Artusio et al. 2005), or *bleton* ‘larch’ which is typical of the Northern Valleys (see Pons & Genre 1997). On the other, we also find terms characterised by diaphasic

[10] Within the count of the *new* lemmas also fall the new so-called *Instances of phonological words* (in the CLiMAlp platform jargon ‘tokens merging two or more lemmas and forming one phonological word’) of already present lemmas, for example *plo* < *per* + *lo* ‘for the’, or *trobant-me*, inflected form of *se trobar* ‘to find oneself’ with enclitic pronoun.

variation, especially connected to special lexical fields, for example *allòdi* ‘freehold property’, from the legal-historical lexicon, or *coperton* ‘tyre’, from the automotive sector.

The second group of examples show that many of the new lemmas consist of loanwords, mostly from Italian (*distanciament* < It. *distanziamento* ‘distancing, spacing’, *azerar* < It. *azzerare* ‘reset, reduce to zero’), but also from other surrounding languages (*sagrin* ‘worry’ < Piedm. *sagrin* or Fr. *chagrin*).

Another simple statistical indicator we used to measure the databases’ response is the *Exploitation Index*, i.e., the percentage of Basic Dictionary *empty* lemmas after the corpus lemmatisation. Regarding this index, less than 25% of the lemmas contained in the original OC thesaurus were found in the Corpus at the end of the first populating process. The data appears to be strictly related to the corpus size, as it rose to 73,750 tokens following the Recognition Tests (see Section 4.3). The exploitation percentage also rose to 32%.

*Francoprovençal***.** Following the strategy chosen for FP, the Corpus population was carried out via the upload of orthographically normalised texts contained in the virtual library of the *BREL* (20 texts). Since these texts belong to the subset of the main town Aosta, they can be held to refer to a diatopically marked variety, though a highly influential one. The most frequent genres are local or general fairy tales (12 texts), but translations of common issues articles (5), ethnographic essays (2) and poetry (1) are also present. The number of tokens thus obtained was 8,596, which makes up less than the half of the OC corpus. Using the same calculation method previously applied, 869 new lemmas were introduced, with an incremental percentage of 54.8% of the Basic Dictionary (1,584 lemmas); the New Lemmas/Tokens ratio is more than double that of the OC database one (10.1).[11]

Of these new lemmas, 138 (15.9%) refer to *grammatical* Part Of Speech (articles, pronouns, adjectival determiners, prepositions, conjunctions, interjections), whilst the *lexical* lemmas count 283 nouns (38.8%), 219 verbs (29.9%), 147 lexical adjectives (20.1%) and 82 lexical adverbs (11.2%). The new grammatical lemmas cover almost entirely the initial gap of the Basic Dictionary, while the lexical part of the vocabulary should of course be considered as strictly related to the genre of the uploaded texts. Besides, 244 new variants (mostly phonetic and graphic) were introduced, despite our expectations in light of the diatopic correspondence between Dictionary and Corpus and the previous normalisation work done by the *BREL*. Among these variants, the phonetic ones are more frequent in loanwords (L: *université* 'university' > V: *universitó*, *universitoù*; L: *eumpléyà* 'employee' > V: *eumplèyà*, *eumployé*), which frequently exhibit an alternance between spoken

[11] As for the FP corpus, an intermediate incremental value was also observed: 407 new lemmas were added for the first ten texts (+20.4% for 3,463 tokens; NL/T ratio: 11,8), 462 for the second half (+18.8% for 5,133 tokens; NL/T ratio: 9.0).

and learned forms, but also occur for common vocabulary (L: *totsé* ‘to touch’ > V: *totché*, *toutchì*). In other cases, the variants are simply due to the uneven compliance with the *BREL* writing conventions (L: *étudiàn* ‘student’ > V: *étudian*, with no accent on the stressed final vowel followed by nasal).

As for the Basic Dictionary exploitation, the selective character of the FP dictionary determines, in the first upload phase and despite the absence of words belonging to grammatical POS, a coverage percentage higher than the OC one (31% *vs.* 25%). A similar value for the OC database (32%) was obtained only after the second upload phase, i.e., with 73,750 tokens, compared with only 8,596 tokens of the FP after the first.

4.3 “Machine-learning” performances

We used quotation marks for the term *machine learning* in the chapter’s title. The tests we conducted to test out the abilities of our database in fact lacked the dimensions which are essential to proper machine learning methods. These dimensions refer to those *superior* levels of language (namely morphosyntax and, above all, semantics and pragmatics) which are necessary in order to actively and correctly respond to given linguistic tasks (Mooney 2005: 377). According to their present state of basically annotated corpora, we tested whether our databases were able to recognise and correctly lemmatise the linguistic forms occurring within the other new written texts which were gradually submitted.

*Occitan***.** In general terms, the OC database response to the Recognition Tests has been quite satisfying: more than 75.5 % of the above 52,000 tokens added for the test were automatically recognised. Considering that the corpus collects texts produced within the hugely articulated linguistic (dia)system of Cisalpine Occitan, and moreover were created by writers whose primary literacy was achieved only in another language (Italian) and who often deviate from the orthographic norms of the writing they are learning, the results allow us to be optimistic about the compliance of such an instrument in the view of developing a written corpus fully functional to the study of the language.

What the results firstly (and expectedly) show, is the relevance of the submitted texts' writing model. In fact, the lowest recognition percentages were not obtained from texts from linguistic sub-areas which were particularly divergent from the central one, nor from texts characterised by the highest presence of special language. Instead, the lowest recognition rates were obtained from a prose text belonging to the central area of OC (Val Varaita), arguably because the text was written using a different writing system. We refer to the two available versions (strictly *Mistralian* and *Concordata*; see Section 4.1) of the Giovanni Bernard's novel *Steve* (Bernard 2007) which both scored, without any meaningful difference, a less than 50%

percentage of recognised tokens, many of which were also affected with issues of homonymy.

Regarding the diatopic variation within the OC area, the written texts (common affairs articles and poetry) from the central area (Stura, Maira and Varaita Valleys) receive, as expected, recognition percentages that are on average higher than those for texts produced in the Northern Valleys. On the contrary, the percentage recognition for the texts from Vermenagna Valley, in the south, were comparable to those of the central ones.[12]

Remaining on the issue of dialectal variation but focusing also on the relationship between Cisalpine and Transalpine linguistic standards, an interesting result is represented by the submission of two different translations of Jean Giono's novel *L'homme qui plantait des arbres* (Giono 1980: 754-767), both using Normalised writing. The first was translated directly from the French original into the Valle Stura local variety; the second was adapted to the general norms of Cisalpine Occitan via a previous translation from French to standard Occitan of Languedoc. Despite the local characteristics of the first, both texts exhibited very satisfying and quite comparable recognition percentages (76% *vs.* 78%), thus demonstrating the satisfying diasystematic coherence of this *ideal* linguistic ensemble, balancing on the two axes of the

[12] The reliability of this data is however questionable because these texts, formerly written in *Concordata* system, have been converted into the *Normalized* one by writers originally from Central Valleys.

core area of Occitan in a broad sense (Languedoc) and the central area of Cisalpine Occitan (Valle Stura).

The system has also been tested by two articles published in the local journal *Ousitanio Vivo* but written by Transalpine Occitan speakers, and a review from *Linguistica Occitana*, a journal published in Montpellier. The three texts have scored quite high percentages (from 64% to 69%), in any case comparable to those observed for the most peripheral Cisalpine Occitan area, the Northern Valleys.

Moving on to the genre analysis, it seemed interesting to test the difference in recognition between poetry and current affairs articles two text types that can be considered as significatively distant on the axis of the discourse structure (or *conception*, according to Koch & Oesterreicher; see also Benedetto Mas & Pons 2016). In general terms, the result of the two subsets (10 poems, 10 articles) were similar: the poetry subset registered a percentage of 71% recognition of its almost 3,150 tokens, in line with the 70% scored by those of the articles (8,200 tokens). Nonetheless, on closer inspection, the two subsets differ for the distance between the lowest and highest scores, which is considerably wider for the first, reaching 21 points (59-80%) against 10 (64-74%) for the second. This could be easily explained considering the difference between the lexicon of poetry (which is generally more constrained but also very idiosyncratic, based also on the individual poems' themes) and of generic prose lexicon, which corresponds more to the normative model which underlies the creation of DOC 2008. With respect to this, it is also to be noticed that the texts produced after 2008 are on average better tokenised and recognised than the others: as obvious as it may seem, these data demonstrate the diffusion of this instrument among the writers who had previously adopted the *Classical* writing.

Francoprovençal. As for the FP database, which is less advanced than the OC one in quantitative terms (8,596 tokens) after the first Corpus upload phase, the Recognition Test was conducted with the specific goal of verifying its responsivity to diatopic variation, and to address the subsequent population phase.

The first test was conducted by feeding the database with the same text (the fairy tale *La vache partagée*, already uploaded and lemmatised in the Aosta variety version) in the 56 diatopically characterised versions, one for each municipality, published on the *BREL* website (PatoisVda, *Tresor*).

Since the text was already *known* by the machine, we may say that this Recognition Test (*Old_Text RT*) was aimed at assessing the machine's ability to deal with *pure* dialectal variation, and the results have been quite interesting. The 19,364 new tokens were correctly recognised and attributed to an existing lemma in 46.63% of cases. This relatively low average value shows in general terms the huge linguistic variation of the regional area, but what is important to underline is that the percentage distribution (see Figure

3) follows very closely the pattern of the variation established by scholarly findings. The High Valley varieties, to which Aosta's *patois* belong, show in fact an average recognition of 60%, while the percentages decrease progressively when we move towards the Low Valley, and goes under 35% at the south-eastern border with Piedmont and in some side valleys of the same zone. More in detail, the highest values (often more than 80%) are recorded in the restricted sub-area of the Grand Saint Bernard valley (north of Aosta, towards the homonymous mountain pass), thus confirming the existence of a core-zone of pronounced linguistic convergence in the territory extending from the chief town to Swiss Valais, a linguistic spatial configuration already suggested by previous geolinguistic and dialectometric studies (Raimondi 2019: 42, 51-52, based on data coming from the first volume of APV-*Atlas des Patois Valdôtains*; Favre & Raimondi 2020).

@@ Insert FIG04.jpg

In parallel with the OC database, a recognition test based on new texts was conducted also on the FP data. For this *New_Texts RT*, good results were recorded for the Aosta variety (76.3% of the new tokens), while the values for other eight variants of the same text (Figure 4) ranged from a minimum of 41.6% (Challand-Saint-Anselme, Low Valley) to a maximum of 69.2% (Saint-Christophe, near to the main town), with a general average value of 57.7%.

@@ Insert FIG05.jpg

A final test on Francoprovençal exogenous written specimens, combining linguistic and writing system variety, was conducted. The submitted texts were a text from Mezzenile (Piedmont) in *BREL* writing system, one from La Rochette (Savoy) in *Conflans* writing and one from Saint-Maurice (Valais) in the local *Graphie Commune* (see above, Note 8). Though generally lower scores were expected, their inner ranking is, in a way, surprising: the lowest, in fact, appears to be that of the Piedmont text (only 18.9%, against 32% for the French and 22.3% for the Swiss texts), which was the only one to share the writing system with the database and whose unsatisfying recognition is shown to be motivated mainly by linguistic distance. The average percentage for the whole set of new texts (both from Aosta Valley and from abroad) was 32.7%.

5. Summary and outlook

The results of the stress tests described above, though necessarily provisional due to being conducted at an early stage of the population process of the two Romance databases, may be of some use with respect to their future implementation, and will be briefly summarised.

The populating strategies adopted separately for the OC and FP corpora were, as we have seen, different in their roots and they correspond to a different position of the two linguistic minorities in facing the issues of *normalisation* and of the relation between *standard language* and *linguistic variation*, which can be fundamentally described as an opposition between *top-down* (OC) and *bottom-up* (FP) attitudes, respectively.

As shown above (see Section 4.1) for the OC context, the top-down attitude manifests itself in the progressive orientation, from the year 2000 onwards, towards the adoption of a supralocal writing norm, well backgrounded in the mainland French Occitan (culturally rooted in the medieval Provençal literature tradition and with modern Catalan as a remote roof-language model; Regis 2020), promoted through a dictionary which is conceived not so much as a mere descriptive instrument, but as a lexicographic mean for fostering orthographic and linguistic normalisation among the Italian Occitan communities.

The better general recognition test results for OC are clearly related to this political-linguistic operation, and the most significant ones are probably the increasing automatic recognition values for the texts produced after the publication of DOC (2008) on one side and the good results for exogenous French Occitan texts on the other, the latter testifying the convergence of the adopted model with general Occitan standards. On the contrary, it will be interesting to further test the OC databases' responsive capacity with regard to textual corpora both from previous or divergent states of writing elaborations and from the most peripheral area of the OC domain.

In the absence of a *polynomic writing system* comparable to the OC Normalised, the results of the model adopted for the FP database construction, based on a local writing model (regional *BREL* writing) and on a local variety (Aosta), are naturally much lower in terms of general average automatic recognition (32.7% *vs.* 75% for OC). A finer analysis nevertheless enables us to observe at least two computational aspects, concerning the general performances and the *Basic Dictionary/Corpus* relation.

Regarding the first aspect, it is interesting to note that the general average recognition performance of the OC database (75%) is almost equal to the one for the only *diatopically coincident* new text in FP database (Aosta: 76.3%). As the OC database was built instead with a *diatopically varied* set of texts, these data suggest for OC the substantial matching of its chosen standard with an ideal, and not particularly geographically connotated, *standard Cisalpine Occitan*. As for the second aspect, a deeper analysis of the *New Lemmas/Tokens ratio* (which is very different for the two databases) and of the *Exploitation Index* (which is on the contrary quite similar, in spite of the different dimensions of the two databases) might give the possibility to measure the difference between the top-down and the bottom-up approach, both in terms of future implementation strategies and to compare our results to those obtained by computational approaches to other low-density minority languages (Gaeta et al. 2022, Note 3-7) or to national standard languages.

Regarding other perspectives, the results obtained by the FP database in the diatopic variation recognition test (*Old_Text RT*) confirm the overall view of Aosta Valley's dialectal variation emerging from inquiries conducted with traditional geolinguistic methods and suggest pursuing the ongoing bottom-up approach adopted here. This can be done drifting away from the chosen initial centre (Aosta) in a sort of *radial enlargement* of the linguistic space that should perhaps be gradual, in order to better monitor the significant linguistic variation of FP *continuum*. This perspective, incidentally, corresponds to the current vision of Francoprovençal in a broader sense: not as *a* language, but rather as "a collection of speech varieties displaying a common linguistic typology yet an extremely high degree of dialect fragmentation" (Kristol 2016: 350).

As for OC, the path seems easier and better traced, and future implementations of the corpus will add data to the analysis of the diffusion and of the actual employment of a standard that seems already to be on its way, thus finally opening the door to further corpus-driven approaches. Returning to the initial metaphor, we might conclude that when it comes to writing, in Italy the page for OC is much less *blank* than for FP.

References

Angster, Marco, & Gaeta, Livio. 2021. Contact phenomena in the verbal complex: the Walser connection in the Alpine area. In *The Alps as a linguistic area* [Special Issue of *Language Typology and Universals (STUF)* 74.1], Livio Gaeta & Guido Seiler (eds), 73-107. Berlin: De Gruyter/Mouton.

Angster, Marco, Bellante, Marco, Cioffi, Raffaele & Gaeta, Livio. 2017. I progetti DiWaC e ArchiWals. *Bollettino dell'Atlante Linguistico Italiano* 3/41: 83-94.

Artusio, Lorenzo, Audisio, Piermarco, Giraudo, Gianni & Macario, Eliano. 2005. *Dizionario occitano Robilante – Roccavione*. Roccabruna: Chambra d'Òc.

Benedetto Mas, Paolo & Pons, Aline. 2016. Expériences d'écriture du francoprovençal en Piémont: continuité et originalité au regard de la réalité occitane. In *Transmission, revitalisation et normalisation*, 75-85. Aosta: Région Autonome de la Vallée d'Aoste.

Benedetto Mas, Paolo & Regis, Riccardo. 2022. Il francoprovenzale in Piemonte: qualche appunto. In *La Suisse romande et ses patois. Autour de la place et du devenir des langues francoprovençale et oïlique*, Dorothée Aquino-Weber & Maguelone Sauzet (eds)*,* 165-183. Neuchâtel: Alphil.

Bernard, Giovanni. 2007. *Steve. Roumans ousitan*. Cantalupa (TO): Effatà.

Borghi Cedrini, Luciana. 2017. *Ai confini della lingua d'oc. Nord-Est occitano e lingua valdese*, ed. by Giraudo, Andrea, Meliga, Walter & Noto, Giuseppe. Modena: Mucchi.

Cerlogne, Jean-Baptiste. 1907. *Dictionnaire du Patois Valdôtain précédé de la petite grammaire*. Aoste: Imprimerie Catholique.

Chenal, Aimé & Vautherin, Raymond. 1997. *Nouveau dictionnaire de patois valdôtain.* Aosta: Musumeci.

Cioffi, Raffaele, Angster, Marco, Bellante, Marco, Benedetto Mas, Paolo, Gaeta, Livio, Murelli, Adriano, Pons, Aline, Raimondi, Gianmario & Rivoira, Matteo. 2021. Mountain linguistics: The Western Alpine Landscape. Paper presented at the *Online Workshop on "Mountain*

Linguistics", 54th International Annual Meeting of the Societas Linguistica Europaea, 30.8. - 3.9.2021 <https://osf.io/4sthq/>.

CLiMAlp = CLiMAlp project website <www.climalp.org>.

DOC 2008 = Commissione Internazionale per la normalizzazione linguistica dell'Occitano Alpino. 2008. *Dizionario Italiano-Occitano, Occitano-Italiano*. Cuneo: +Eventi.

Dublin Core Metadata Initiative. *Dublin Core Metadata Innovation home page* <https://dublincore.org/>.

Favre, Saverio. 2010. Francoprovenzale, comunità. In *Enciclopedia dell'italiano* , Simone, Franco (ed). Roma: Istituto dell'Enciclopedia Italiana Treccani <https://www.treccani.it/enciclopedia/comunita-francoprovenzale_%28Enciclopedia-dell%27Italiano%29/>.

Favre, Saverio & Raimondi, Gianmario (eds). 2020. *APV-Atlas des Patois Valdôtains, vol. 1: Le lait et les activités laitières*. Aosta: Regione Autonoma Valle d'Aosta/Le Château.

Gaeta, Livio, Angster, Marco, Cioffi, Raffaele & Bellante, Marco. 2022. Corpus linguistics for low-density varieties. Minority languages and corpus-based morphological investigations. *Corpus* 23 <https://doi.org/10.4000/corpus.7345>.

Garnier, Quentin. 2020. Le vivaro-alpin: progrès d'une définition. *Géolinguistique* 20 <https://doi.org/10.4000/geolinguistique.1992>.

Giono, Jean. 1980. *Œuvres romanesques complètes*. Paris: Gallimard.

Iannàccaro, Gabriele & Dell'Aquila, Vittorio. 2008. Per una tipologia dei sistemi di scrittura spontanei in area romanza. *Estudis Romànics* 30: 311-331.

Kloss, Heinz. 1967. *Abstand* Languages and *Ausbau* Languages. *Anthropological Linguistics* 9: 29-41.

Koch, Peter & Oesterreicher, Wulf. 2008. Comparaison historique de l'architecture des langues romanes. In *Histoire linguistique de la Romania*, Gerhard Ernst, Martin-Dietrich Gleßgen, Christian Schmitt & Wolfgang Schweickard (eds), Vol. III, 2575-2610. Berlin/New York: De Gruyter.

Kristol, Andres. 2016. Francoprovençal. In *The Oxford Guide to the Romance Languages*, Martin Maiden & Adam Ledgeway (eds), 350-362. Oxford: Oxford University Press.

Maxwell, Mike & Hughes, Baden. 2006. Frontiers in Linguistic Annotation for Lower-Density Languages. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, 29-37. Stroudsburg (PA): Association for Computational Linguistics.

Law 482/1999. Parlamento Italiano, *Legge 15 Dicembre 1999, n. 482 "Norme in materia di tutela delle minoranze linguistiche storiche"*, pubblicata nella Gazzetta Ufficiale n. 297 del 20 dicembre 1999 <https://web.camera.it/parlam/leggi/99482l.htm>.

Mooney, Raymond J. 2005. Machine Learning. In *The Oxford Handbook of Computational Linguistics*, Mitkov, Ruslan (ed), 376-394. Oxford: Oxford University Press.

Nichols, Joanna. 2015. Types of spread zones. In *Language Structure and Environment*, Rik De Busser & Randy J. LaPolla (eds), 261-288. Amsterdam/Philadelphia: John Benjamins.

Oliviéri, Michèle & Sauzet, Patrick. 2016. Southern Gallo-Romance (Occitan). In *The Oxford Guide to the Romance Languages*, Martin Maiden & Adam Ledgeway (eds), 319-349. Oxford: Oxford University Press.

PatoisVdA, *Glossari*. *Glossari per comune*. In Regione Autonoma Valle d'Aosta, *PatoisVdA. Il sito del francoprovenzale in Valle d'Aosta* <www.patoisvda.org/it/glossari-per-comune/aosta_a/>.

PatoisVdA, *Grafia*. *Grafia: per principianti, per esperti*. In Regione Autonoma Valle d'Aosta, *PatoisVdA. Il sito del francoprovenzale in Valle d'Aosta* <https://www.patoisvda.org/it/grafia>.

PatoisVdA, *Trésor*. *Trésor de textes*. In Regione Autonoma Valle d'Aosta, *PatoisVdA. Il sito del francoprovenzale in Valle d'Aosta* <https://www.patoisvda.org/tresor-de-textes/>.

Pons, Teofilo G. & Genre, Arturo. 1997. *Dizionario del dialetto occitano della Val Germanasca*. Alessandria: Edizioni dell'Orso.

Raimondi Gianmario. 2019. Atlanti interpretativi, cartografia sintetica, distanza linguistica. Il banco di prova dell'APV-Atlas des patois valdôtains. *Géolinguistique* [online] 19 <https://doi.org/10.4000/geolinguistique.1170>.

Raimondi, Gianmario. 2020. ALEPO et APV: la contribution de l'Italie à l'étude de la *Galloromania peripherica*. *Bien dire et bien aprandre* 34: 109-130.

Regis, Riccardo. 2020. Profilo dell'occitano in Piemonte: aspetti sociolinguistici. *Estudis Romànics* 42: 101-125.

Rivoira, Matteo. 2016. L'occitano nelle valli del Piemonte. *Bollettino della Società Storica Pinerolese* 33: 173-185.

Regis, Riccardo & Rivoira, Matteo. 2016. Ortografie e lingue tetto: qualche appunto. *L'Italia Dialettale* 77: 261-283.

Rivoira Matteo. 2021. Note linguistiche a *Vautres que m'avetz tuada*. In Joan Ganhaire, *Voi che mi avete uccisa.* Introduzione, traduzione e note di Monica Longobardi, 138-158. Arenzano (GE): Castel Negrino.

Schüle, Ernst. 1992. *Comment écrire le patois? (Principes et conseils pratiques)*. Saint-Nicolas/Aoste: Centre d'Études Francoprovençales "René Willien".

Stich, Dominique, Gouvert, Xavier, Favre, Alain & Walter, Henriette. 2003. *Dictionnaire des mots de base du francoprovençal. Orthographe ORB supradialectale standardisée.* Paris: Le Carré.

Sumien, Domergue. 2006. *La standardisation pluricentrique de l'occitan. Nouvel enjeu sociolinguistique, développement du lexique et de la morphologie*. Turnhout: Brepols.

Urban, Matthias. 2020. Mountain linguistics. *Language and Linguistics Compass* 14(9): e12393.

Zürrer, Peter. 2009. *Sprachkontakt in Walser Dialekten. Gressoney und Issime im Aostatal.* [Zeitschrift für Dialektologie und Linguistik, Beihefte 173], Stuttgart: Steiner.