A digital, retro-standardized edition of the Tableaux phonétiques des patois Suisses romands (TPPSR)

Abstract

This study presents a digital, retro-standardized edition of the Tableaux Phonétiques des Patois Suisses Romands (TPPSR), an early collection of lexical dialect data of the Suisse romande, which was compiled by Louis Gauchat, Jules Jeanqaquet, and Ernest Tappolet in the beginning of the 20th century and later published in 1925. While the plan of Gauchat and his collaborators to turn their data into a dialect atlas could never be realized for the lack of funding, we show how consistent techniques for digitization, accompanied by transparent approaches to retro-standardization can be used to turn the original data of the TPPSR into a modern interactive dialect atlas. The dialect atlas is not only publicly available in the form of a web-based application, but also in the form of a dataset that offers the data in standardized, human- and machine-readable form.

1. Introduction

The Tableaux Phonétiques des Patois Suisses Romands (TPPSR) are a collection of lexical data of 62 dialect varieties spoken in the Western part of Switzerland in the beginning of the 20th century. Compiled by Louis Gauchat, Jules Jeanqaquet, and Ernest Tappolet, who published the data in tabular form in 1925, the TPPSR are a unique witness of the former linguistic diversity that once characterized the Suisse romande. In the early 20th century – quite in contrast to today, when digital maps and geographic information services dominate people’s lives – it was very difficult and expensive for linguists to create high-quality dialect maps, and Gauchat and his colleagues never succeeded in showing the linguistic diversity of the Suisse romande in geographic space. In this study, we will illustrate how the use of modern techniques for digitization along with new, transparent techniques for retro-standardization can be applied to fulfill the dream of Gauchat and his colleagues by turning the original TPPSR data into a rich web-based application that serves as a modern interactive dialect atlas. In the following, we will first provide a short background on the TPPSR (Section 2) and then illustrate how we digitized the data (Section 3) and retro-standardized them (Section 4). The result of our endeavour is a new, digital and interactive edition of the TPPSR, which can be freely accessed at https://tppsr.clld.org. In addition, we provide the data underlying the web-based resource in the form of a human- and machine-readable dataset which can be used for quantitative and qualitative analyses.

2. Background

2.1. Fading voices of Suisse romande

Until the second half of 19th century the Western part of Switzerland, called Suisse romande or Romandy, showed a puzzling variety of Franco-Provençal and Franc-Comtois dialects that was nearly unparalleled in the Romance-speaking world. However, this vital linguistic diversity came to an abrupt end as the vast majority of local dialects gave way to
high variety Standard French for socio-economic reasons. When Suisse dialectologists Louis Gauchat and his collaborators Jules Jeanjaquet and Ernest Tappolet started collecting dialect data for the newly founded *Glossaire des patois de la Suisse romande* (GPSR) in 1899, they saw many local dialects in full dissolution and they often had to rely on aged speakers with only a limited competence of their patois.

In his endeavor to save the cultural heritage of Suisse romande, Gauchat designed a series of short questionnaires that were sent at monthly intervals to specially trained informants in 400 locations throughout Romandy. During a period of four years, from 1899-1903, a multitude of more than one million lexicographic notes was collected that provide the material basis of the GPSR to this day. Unfortunately, the phonetic transcriptions of the returned notes showed many disappointing results and needed permanent corrections. Therefore, Gauchat and his collaborators decided to make their own on-site phonetic survey of 62 characteristic Romandy dialects (*patois-types*) to be able to correct and complement the incoming results (Gauchat and Jeanjaquet 1920, 22). They collected the data for their survey in sections of several weeks per canton from July 1904 to September 1907.

When the three investigators toured the country anew at the beginning of the 20th century, they observed how dialect use had continued to decline (Gauchat et al. 1925, 160-168). In Vaud, local dialects had only survived in some remote areas. In the north and southwest of the canton they had almost completely disappeared. In Berne, the industrial south had already switched to French, whereas in some northern parts, like Ajoie et Franches-Montagnes, the local Franc-Comtois dialects resisted better. In Neuchâtel, local dialects had already disappeared during the second half of the 19th century and only a handful of aged people kept rudimentary knowledge of their patois. In Fribourg, dialect use was still frequent with aged people living in rural districts not influenced by greater cities. In Geneva, the local Savoyard dialects had disappeared in the city of Geneva and in surrounding areas. Dialect use had only survived in a few rural catholic districts. Only in Valais local dialects were still vigorous in rural districts. Nonetheless, French gained influence quickly because parents began to speak French to their children to avoid social disadvantages.

In 1914 Gauchat announced that their phonetic survey would soon be published as a complement to the GPSR with the title *Relevés phonétiques* (Gauchat 1914, 26). Unfortunately, due to the considerable costs to typeset the complex transcription system, the survey could not be published until 1925 under the modified title *Tableaux phonétiques des patois suisses romands. Relevés comparatifs d’environ 500 mots dans 62 patois-types* (Gauchat et al. 1925). Furthermore, Gauchat wanted to publish a dialect atlas called *Atlas linguistique de la Suisse romande* that should contain 80 colored, large format maps showing the distribution of phonetic features of local dialects (Gauchat 1914, 26, Gauchat et al. 1924, I, 7). This project could never be realized due to lack of funding. There exists a single map called *Spécimen de l’Atlas Phonétique de la Suisse Romande* which shows the development of Latin K/G+A (see Gauchat 1925, V, Fn 1, and Pop 1950, 256, planche XXIV).

---

1 Gauchat made plans for such an atlas right from the start of GPSR in 1899. In 1903 he published a sample map synthesizing 20 phonetic features of local dialects (Gauchat 1903, 393, reproduced in Gauchat 1904, 17).
The TPPSR remained the only comprehensive dialect survey of Suisse romande. It would have been impossible to make a comparable survey later on, because most local dialects had completely disappeared in the twenty years from data collection in 1904 to 1907 until publication of the TPPSR in 1925 (Gauchat 1925, VI). Today, the UNESCO Atlas of the World’s Languages in Danger lists Franco-Provençal as “definitely endangered” (Moseley 2010), with about 16,000 speakers estimated for Switzerland according to a census from 2000 (Lüdi and Werlen 2005), and only 7,000 speakers according to Ethnologue, based on a census from 1998 (Lewis 2009).

2.2. Benefits of the TPPSR

The TPPSR represent a rare type of in-depth phonetic investigation that nearly seamlessly covers a whole dialect area and offers a valid cross-section of characteristics of the dialects in question. Furthermore, they contain a number of theoretical innovations, initiated by Gauchat, that were only taken up more than half a century later in sociolinguistics. Especially Labov and Chambers gave Gauchat the appreciation he deserved in praising him as a forgotten forerunner of variationist linguistics, emphasizing his carefully conceived attempts to study language in its social context (Labov 1963, Chambers et al. 2008, Chambers 2013). But it should be noted that both referred exclusively to Gauchat’s perspicacious study on Charmay (Gauchat 1905) and did not even mention his comprehensive dialect survey. Thus, the TPPSR remain a forgotten milestone in the description of dialects2, which are worth being made accessible again to scientific research and an interested public in the form of new, interactive media experience.

The questionnaire of the TPPSR contains approximately 170 simple sentences arranged in tabular form. For an optimal presentation in print, the sentences were split into phrases, each containing a central concept that rests embedded in the grammatical and prosodic structure of the whole sentence. The phrases are arranged in continuous columns with normally six columns per page. The queried points are listed horizontally with points 1-31 on left sides and points 32-62 on right sides of the book (see the original pages on TPPSR Online). A column header in French specifies the meaning of the phrases. A secondary column header indicates the etymological Latin or reconstructed Proto-Romance form. Overall, there is a matrix of 480 columns and 62 rows. The tabular form is database compliant with just a few adaptations. The resulting TPPSR database consists of 29464 items3.

The questionnaire is composed of sentences dealing with everyday rural life (“It’s hot today”; “Let’s go there together”; “He sweeps in front of the barn door”; etc.). Moreover, the sentences were related thematically to create a normal conversation atmosphere during the

---

2 The only significant recourse to TPPSR is found in Goebl (1985, 1987) who used tables 1-249 as a database for one of his well-known dialectometrical investigations. Some hundred macro- and microstructural phonetic features of TPPSR entries were used in nominalized form to reconstruct the genealogical relationships between Romandy dialects by means of a similarity index.

3 296 items in the source are indicated as not having received a response, which explains why we do not receive 480 x 62 = 29760 items.
interviews. Although the questionnaire focused primarily on phonetic features, it also covers outstanding morphological and lexical features of the Romandy dialects.

The questionnaire of whole sentences presented as phrases in columns has a triple advantage. Firstly, the words denoting the actually interesting concepts are placed in a natural context of speaking that selects a specific meaning of the word. Secondly, the phrasal embedding of these words in varying contexts produces a large number of grammatical forms (articles, pronouns, verb paradigms, etc.), that allow for the building of dialect grammars. Thirdly, some Romandy dialects show an unstable stress placement that becomes evident only by indicating the phrasal prosodic structure (see stress shifts in TPPSR C 30+31 on voit la lune realized in Ormont-Dessus (P 15) as ô vaê là lënæ and in L’Auberson (P 3) as ô vê là länã).

The study strongly emphasizes the description of phonetic variation. Therefore, the TPPSR main tables have two different footnote types with listings of phonetic variants. Footnote type 1 called Remarques contains a) phonetic variants and phonetic tendencies given by Jeanjaquet in relation to his own main table entries, b) additional morpho-syntactic and lexical information to the main table entry. Footnote type 2 called Notations divergentes contains phonetic variants and phonetic tendencies given by the respective co-investigator Gauchat or Tappolet who accompanied Jeanjaquet (see the original page with footnotes to columns 1-6 on TPPSR Online). This ambitious procedure allows double-checking and validating all main entries given by Jeanjaquet. As Jeanjaquet participated in all investigations only his transcripts were listed in the main tables in order to get a coherent view of a single investigator.

The questionnaire also contains many repetitions to test phonetic stability of forms (see C 1-2 il fait chaud, C42-43 il fait du vent, C455-457 l’oiseau fait son nid). In addition, to better document phonetic particularities of local dialects that were not covered by the general questionnaire, the investigators could freely add words to increase the flexibility of documentation. For example, some local realizations for ‘cage’ are additionally quoted, because they show a noticeable development for stressed long Lat. A: La-Côte-aux-Fées (P 50) këvâ and Develier (P 60) ūv. Additional words are listed in a separate chapter and grouped in the phonetic category they serve to illustrate (Gauchat et al. 1925, 168-176 Mots supplémentaires, and 196-197 Index des mots supplémentaires).

3. Digitizing the TPPSR

Given that – as mentioned in the preceding paragraph – already the typesetting from the manual notes to printed form took several years and vast resources before it could be accomplished, it should not be surprising that the TPPSR with their so-called impressionistic transcription system pose great challenges for a consistent digitization. In the following, we...
will briefly discuss the major characteristics of the TPPSR transcription system, in order to present how we decided to render the data in Unicode.

Until the introduction of Unicode as a new IT standard for text encoding, dialect processing with computers was a tedious task. The widespread American Standard Code for Information Interchange (ASCII) only contained a Latin based alphabet with 26 lowercase letters (still reflected in Unicode block Basic Latin) with missing code points even for diacritical characters in common orthographies (like French ç or German ü). Thus, complex transcription systems had to be digitized in ASCII by serializing base character and diacritical marks. For example, TPPSR transcriptions like Longirod tsô ‘hot’ [1:6] or Collombe tsô ‘hot’ [1:18] had to be encoded as tsô- and tsô^/. For typesetting it was necessary to use programs which rebuild the original glyphs and printed them on maps (Geisler 1993).

Unicode, which was especially designed to encode the world’s writing systems (Moran and Cysouw, 2018), changed the situation completely from 1991 on. Thousands of characters for Latin based scripts became suddenly available (like <œ>, <æ>, <ø>, <ß>, etc.). Even phonetic transcriptions were easily possible with a complete set of IPA characters at hand (see the block of IPA Extensions). Most important for traditional transcription systems was the introduction of the Unicode block Combining Diacritical Marks whose non-spacing characters combine freely with other characters to build multi-level graphemes like <â>, <ê>, <œ̃>, etc. Unfortunately, earlier versions of Unicode did not yet cover all TPPSR diacritics and missing characters are added only slowly. Some very useful superscript letter diacritics came with version 3.2 in 2002 which allowed to represent high-frequency vowel graphemes like <â>, <ê>, etc. Special characters to indicate sound ligatures were introduced in version 4.1 in 2005 to model frequent vowel ligatures like <ao>, <ou> <âi>, etc. And finally, a character for the base vowel <u> was introduced in 2014 (see U+AB52 LATIN SMALL LETTER U WITH LEFT HOOK). Still today, some code points for low-frequency diacritics are missing and must be replaced by substitute characters (see 3.3.6). Nevertheless, it seems appropriate to present a TPPSR retro-digitization which can be transliterated to IPA for reuse and comparison.

3.2. The TPPSR transcription system

Gauchat and his collaborators took over the transcription system of Rousselot and Gilliéron – commonly used in Gallo Romance dialectology at that time – and moderately adapted it to the needs of their investigation area (see Rousselot 1887, Rousselot and Gilliéron 1888, Gauchat et al. 1924, I, 7, Gauchat 1925, VII-X, XX). In particular, they used many superscript letters like <â>, <ê>, <œ̄>, etc. to denote the frequent “voyelles intermédiaires” of Romandy dialects. When they decided in 1910 to print their handwritten notes, there was an urgent need to reduce transcription granularity to make it comparable with the transcription system of the Atlas Linguistique de la France (published between 1903 and 1910), and to reduce costs for typesetting. For example, quantity is only exceptionally indicated with intermediate and centralized vowels (but see Prahins tséra ‘expensive’
3.2.1. Vowels

The TPPSR have seven base vowel qualities <a, e, i, o, u, œ, ꭒ>. These vowel qualities can be fine-graded by a system of vowel quality changing diacritics <◌̃, ◌́, ◌̇, ◌̈> to denote opening/closing, fronting/backing and centralizing of base vowel qualities. These can even be more fine-graded by building intermediate vowels using superscript letters like <a�滘, ā̇, ē> or <œ̇>. In total, these combinations span a grid of 32 vowel qualities (Figure 1).

![Figure 1: TPPSR vowel chart](image)

All vowel qualities allow for a further modification by six generic diacritics <◌̃, ◌̄, ◌̅, ◌̆, ◌̇, ◌̈>, which denote nasality, length, weak articulation and stress. As some of these generic diacritics combine freely with each other, like <short> and <nasal> or <stress> and <nasal>, it is possible to build a large number of vowel graphemes. So, the TPPSR have as many as 58 a-based and 59 e-based vowel graphemes alone (see the TPPSR Grapheme Inventory). This is possible because vowel graphemes can have up to four layers of diacritics and superscript letters built around a base vowel quality (see Figure 2).

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Phonetic value</th>
<th>Example</th>
<th>TPPSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ṝ</td>
<td>stressed long nasalized open-mid front unrounded vowel</td>
<td>tu̞ⁿa 'chain'</td>
<td>149:14</td>
</tr>
<tr>
<td>ḝ</td>
<td>bound weak long near-open front unrounded vowel</td>
<td>pr̈oŋm 'plum'</td>
<td>251:13</td>
</tr>
<tr>
<td>ṝ</td>
<td>stressed long raised close-mid front unrounded vowel</td>
<td>t̛̤s̛̤ 'tête'</td>
<td>421:42</td>
</tr>
<tr>
<td>ḝ</td>
<td>stressed short open-mid back rounded vowel</td>
<td>dr̠b̠va 'double'</td>
<td>143:24</td>
</tr>
</tbody>
</table>

Figure 2: TPPSR multi-level vowel graphemes

6 TPPSR entries are cited in square brackets with indication of chart number and point number, i.e. [181:11] means TPPSR chart 181 ‘expensive’ at point 11 ‘Prahins’. Concept and locality related to Chart number and point number can be looked up at TPPSR Online Concepts and Dialects, where the entries are also linked to the original print form.
In total, the TPPSR main tables use 265 vowel graphemes to capture the wealth of vowel production that had characterized Romandy dialects7.

3.2.2. Consonants

The TPPSR have a consonant transcription system built on 19 Latin small letter based consonant qualities: \(\text{<b, c, d, f, g, h, j, k, l, m, n, p, r, s, t, v, w, y, z>}\). These consonants are fine-graded by 8 consonant quality changing diacritics \(\text{<◌̈, ◌̑, ◌̣, ◌̧, ◌̮, ◌̯, ◌̵, ◌̷>}\) to denote palatalisation, velarisation, retraction, etc. In addition, there is a rather astonishing increase through formation of intermediate consonants, like \(\text{<d̟, g̟, k̟, r̍, s̍, ...>}\), that results in a grid of ca. 40 consonant qualities (Figure 3). Moreover, some consonant qualities allow further modification by generic diacritics \(\text{<◌, ◌̥, ◌̱>}\) for length, syllabicity and weak articulation.

TPPSR notations show a noticeable spread of consonants in the palatal range due to palatalization of Lat. \(\text{K,G+I,E, and K,G+A}\). Apart from palatal consonants, the Romandy dialects have relatively moderate consonant systems which sometimes contrast in a striking way with their multifaceted vowel systems (see the TPPSR Online Sound Inventories).8

3.3. Representing the TPPSR data in Unicode

The general requirement in representing a given dataset in digital form is that the transfer should be unambiguous and lead to no loss of information. A more specific concern for modern digitization projects should be to represent the original data in a standardized form in Unicode to ensure strict reusability. In addition to a clear encoding, this includes that the digital transcription should visually reflect the original data in an appealing way.

---

7 It is to remember, that main table notations are solely based on Jeanjaquet’s notations. The Notations divergentes contain even more graphemes, like a grapheme \(\text{<u̥>}\), used by Gauchat to describe the sound that Jeanjaquet noted \(\text{<i̥>}\) (see Évolène il fait [1:30]).

8 The respective grapheme inventories also provide an indication of this imbalance. There are 265 vowel graphemes and only 109 consonant graphemes in the main tables.
Besides the demanding task to encode hundreds of complex graphemes the printed version of the TPPSR posed additional problems which required some ad-hoc decisions. There are a few graphemes that are not documented in the introductory chapter on transcription techniques nor listed in the summarizing grapheme inventory (see Gauchat 1925, VI-X, XX). For example, Orvin əvį ‘have’ [34:55] shows the grapheme <ỹ> that is used once in the text (see Figure 4.1). It could be a a simple misprint or it could possibly stand for a *velarized voiced palatal approximant*, because Combining Dot Above is used for velarization of consonants like <ṙ> and <ṅ>. Furthermore, there are 11 examples for a grapheme <ɛ> that is not mentioned either and whose phonetic value remains unclear (it doesn’t seem to mark a syllabic consonant, see Martigny-Combe *frwɛ* ‘fruit’ [247:20]).

Some graphemes build variants with identical or nearly identical phonetic value, as Gauchat points out for a series like pőma, pőma, pőma or pőmā. This possibly applies to a pair of graphemes like <û> and <œ̈>. Both graphemes denote a near-close front rounded vowel, maybe with a slight phonetic difference because intermediate vowels tend to emphasize more the articulation of the base vowel (Figure 4.2). The editors did not smooth the individual manuscripts for typesetting to pass on their originality (Gauchat 1925, p. XVIII).

As is to be expected, there are also misprints and textual distortions. Some diacritics for weak articulation printed in a small font are often difficult to distinguish, especially frequent e and o with grave accent or dot above (Figure 4.3 and 4.4). But in general, there are only rare and minor problems that do not considerably hinder digitization.

---

1. Undocumented grapheme <ỹ> [34:55]
2. Unclear phonetic value of graphemes <û> [165:57] and <œ̈> [153:22]
3. Distorted glyphs for short vowel and intermediate vowel <ê̄> [201:31]
4. Distorted glyph for weakly articulated centralized vowel <ɛ̈> [416:19]

Figure 4: TPPSR encoding problems

### 3.3.1. Strict canonical decomposition coding (NFD)

Unicode normalization is required because sequences of code points can be equivalent but differ in its visual and logical order. For example, grapheme <û> is ambiguous to whether it has the combining character sequence <o+û+̃> or the combining character sequence <o+û+̃+̲>. Although these two layerings are visually homoglyphs, they encode differently as <ou0304ou030du0329> vs. <ou0329ou0304ou0300>. Unicode normalization should be applied to this string to reorder the code points into a canonical order, allowing the data to be used adequately for search and comparison (Moran and Cysouw 2018, 61).

The early years of Unicode saw an abundant use of code points for so-called composite or precomposed characters like <ã> = U+00E2 LATIN SMALL LETTER A WITH CIRCUMFLEX or <ɛ> = U+00E9 LATIN SMALL LETTER E WITH ACUTE. These characters typically represent letters with diacritical marks of common orthographies like French â, Spanish ç, German û, Danish ð, etc. The simultaneous introduction of the Unicode
Combining Diacritical Marks led to an unwanted duplication of encodings because visually identical graphemes can have different code points. Later on Unicode has given up creating precomposed characters (NFC) and strongly recommends use of non precomposed encoding (NFD) (see https://unicode.org/reports/tr15/).

### 3.3.2. Canonical Combining Classes (ccc) and stacking order

Stacking order of diacritical marks is based on the Unicode Canonical Combining Class values. Thus, complex multi-level graphemes like `<ê>` have a combining character sequence `<c+◌̂+◌̅>` depending on its canonical combining class values: \u0335 has ccc = 1 [Overlays and interior], \u033E has ccc = 2 [Above] and is therefore coded as `<c\u0335u033E>` (see Unicode Normalization Chart Latin available at https://www.unicode.org/charts/normalization/).

### 3.3.3. Interacting and non-interacting combining classes

Diacritics pertaining to different canonical combining classes have a stacking order determined by the normalization form (NFD). For example, the TPPSR grapheme `<â>` has a non-canonical stacking order with combining character sequence `<a+◌́+◌̱>`, coded as `<a\u0301\u0331>`, that is equivalent to the canonical stacking order with combining character sequence `<a+◌̱+◌́>`, coded as `<a\u0331\u0301>` (Figure 5). The diacritics are non-interacting because they have different canonical combining classes: `<◌́>` has ccc=220 and `<◌̱>` has ccc=230. In case of TPPSR grapheme `<â>` both diacritics have an identical ccc=230. Therefore, they are interacting and the stacking order is regulated by normalization and two encoding sequences with visual difference `<â>` vs. `<ã>` are possible (Figure 6). The TPPSR transcription has chosen grapheme type `<ã>` to transcribe long open-mid vowels but inverse stacking `<ã>` for nasalized open-mid vowels (see Évolène bonne ‘good’ [8:30]).

<table>
<thead>
<tr>
<th>Non-canonical order</th>
<th>Equivalence</th>
<th>Canonical order</th>
</tr>
</thead>
<tbody>
<tr>
<td>a + ◌́ + ◌̱</td>
<td>≡</td>
<td>a + ◌̱ + ◌́</td>
</tr>
<tr>
<td>a\u0301\u0331</td>
<td>equivalent</td>
<td>a\u0301\u0331</td>
</tr>
<tr>
<td>ccc=0 ccc=230 ccc=220</td>
<td>(non-interacting diacritics)</td>
<td>ccc=0 ccc=220 ccc= 230</td>
</tr>
</tbody>
</table>

*Figure 5: TPPSR homoglyph grapheme `<â>`*

<table>
<thead>
<tr>
<th>Non-canonical order</th>
<th>Equivalence</th>
<th>Canonical order</th>
</tr>
</thead>
<tbody>
<tr>
<td>a + ◌̃ + ◌̤</td>
<td>≠</td>
<td>a + ◌̤ + ◌̃</td>
</tr>
<tr>
<td>a\u0301\u0304</td>
<td>not equivalent</td>
<td>a\u0304\u0301</td>
</tr>
<tr>
<td>ccc=0 ccc=230 ccc=230</td>
<td>(interacting diacritics)</td>
<td>ccc=0 ccc=230 ccc=230</td>
</tr>
</tbody>
</table>

*Figure 6: TPPSR non-homoglyph graphemes `<ã>` and `<â>`*

### 3.3.4. Encoding complex TPPSR graphemes

The TPPSR have a transcription system with complex multi-level graphemes built on a slightly modified Latin small letter script used in traditional French orthography. These
graphemes can be remodeled in Unicode with the help of non-spacing diacritical marks which recursively combine with a chosen base character.

The following example illustrates the principles that are effective in encoding Combining Diacritical Marks for the complex vowel grapheme escaped. The grapheme is based on a Latin small letter character and has four layers of diacritical marks with a combining character sequence and a coding sequence . The stacking order of the combining diacritical marks is intrinsically determined by the following canonical combining class values: , , , , . This implies that combining diacritical mark for stress has to be encoded directly after base vowel . However, the remaining three diacritical marks have an identical ccc=230 and can be ordered freely. The resulting stacking order of the TPPSR follows a script tradition widespread in 19th century Romance dialectology.

<table>
<thead>
<tr>
<th>Grapheme</th>
<th>Layer</th>
<th>Code Point</th>
<th>CCC</th>
<th>Stacking Order</th>
<th>Phonetic Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ẽ</td>
<td>&lt;e&gt;</td>
<td>u0303</td>
<td>230</td>
<td>4</td>
<td>nasalized</td>
</tr>
<tr>
<td>ẽ</td>
<td>&lt;e&gt;</td>
<td>u0304</td>
<td>230</td>
<td>3</td>
<td>long</td>
</tr>
<tr>
<td>ẽ</td>
<td>&lt;e&gt;</td>
<td>u0301</td>
<td>230</td>
<td>2</td>
<td>open</td>
</tr>
<tr>
<td>ẽ</td>
<td>&lt;e&gt;</td>
<td>e</td>
<td>-</td>
<td>0</td>
<td>mid front unrounded</td>
</tr>
<tr>
<td>ẽ</td>
<td>&lt;e&gt;</td>
<td>u0329</td>
<td>220</td>
<td>1</td>
<td>stressed</td>
</tr>
</tbody>
</table>

The foregoing example shows that the Unicode stacking order does not necessarily reflect the phonetically motivated layering of diacritics which would place the vowel quality changing diacritic before the generic diacritics for nasality and length and would place the suprasegmental diacritic for stress as outermost layer. For grapheme this would give a combining character sequence which has no effect on visibility but has an impact on reusability and comparison.

3.3.5. Encoding the TPPSR vowel ligature

The TPPSR make frequent use of a tie bar below to indicate a vowel ligature (see Veyrier ‘waé ‘today’ [3:35]). From a logical point of view ligatures equally belong to both elements they connect. Nonetheless, in Unicode ligature ties are not independent characters but are encoded as Combining Diacritical Marks that are obligatorily bound to a preceding character (Moran and Cysouw 2018, 59). Thus, we chose the ligature tie U+035C COMBINING DOUBLE BREVE BELOW (= ) and encoded it as the uttermost diacritical layer of the preceding vowel grapheme. The following “bound” vowel grapheme becomes part of the preceding vowel only by exclusively defining grapheme clusters like escaped. The stacking order of the first part of the cluster is shown in Figure 8. The vowel grapheme cluster is partially canonically ordered with the combining character sequence and the coding sequence .

9 The phonetically motivated ordering of features is also reflected in grapheme descriptions like escaped = stressed long nasalized open-mid front unrounded vowel.
3.3.6. Substitute characters

Some phonetic features of the TPPSR transcription system do not have Unicode code points. In these cases, code points of shape-like characters will serve as substitutes in order to make the TPPSR database compliant and to ensure the uniqueness of the TPPSR encoding (see the TPPSR Online Dialects)\(^\text{10}\).

In some other cases missing code points for the TPPSR phonetic features were introduced in recent extensions of Unicode. For example, the Unicode blocks Latin Extended-E and Combining Diacritical Marks Supplement contain code points U+1DDD COMBINING LATIN SMALL LETTER L, U+1DE4 COMBINING LATIN SMALL LETTER S, etc. Nevertheless, if generic fonts yet have no glyph for these new code points or do not render these code points optimally, we continue to use substitutes for practical reasons and define compatibility equivalences for the proper new code points. Compatibility encoding allows to define equivalence between characters with different visual appearance and behavior: e.g. TPPSR \(<\texttt{ʋ}>\) U+028B as substitute for TPPSR \(<\texttt{ꭒ}>\) U+AB52 (see Unicode Standard Annex #15, Unicode Normalization Forms). The following paragraphs describe some substitute forms with high frequency.

3.3.6.1. Weak articulation

The TPPSR use a small font to indicate weakly articulated sounds. Unfortunately, the use of font properties is purely representative. To get full database compliance the font property was replaced by code point U+0331 COMBINING MACRON BELOW. Thus, original TPPSR forms like \(\texttt{séntē̂}\) are encoded as \(<\texttt{se}\u0306\u0300\texttt{n}\u0331\texttt{ti}\u0304\texttt{e}\u0331\u0307>\) and rendered as \(\texttt{séntē̂̂}\) (Figure 9.1).

3.3.6.2. Weak nasalization

The TPPSR phonetic feature of weakly nasalized sound is substituted with code point U+034B COMBINING HOMOTHETIC ABOVE (see La-Côte-aux-Fées \(\texttt{gr̄ú} \texttt{bārn} \texttt{[53:48]}\)). The feature has a low frequency of only 17 tokens with six grapheme types: \(<\texttt{ə}>\) (3), \(<\texttt{ɛ}>\) (hapax [149:62]), \(<\texttt{ɛ}>\) (10), \(<\texttt{ɛ}>\) (hapax [228:21]), \(<\texttt{i}>\) (hapax [227:28]), \(<\texttt{o}>\) (hapax [364:8]) (Figure 9.2).

3.3.6.3. Latin Small Letter U With Left Hook

In the TPPSR database the grapheme was encoded with the substitute character \(<\texttt{ʋ}>\) U+028B LATIN SMALL LETTER V WITH HOOK (see Landeron \(\texttt{djṓ} \texttt{‘untó} \texttt{[53:48]}\)).

---

\(^{10}\) We hope that future extensions of Unicode will fill these code point gaps for Romance transcription systems (as it has been done with the Teuthonista transcription system, see Revised proposal to encode “Teuthonista” phonetic characters in the UCS).
There was no possibility to properly encode this grapheme until the introduction of code point U+AB52 LATIN SMALL LETTER U WITH LEFT HOOK in block Latin Extended-E (Unicode 7.0, June 2014). Nonetheless, the TPPSR Online (version 1.1) has still the substitute encoding (see ʋ̆, ʋ̃, ʋ̱́, etc. with 3782 tokens). The reason is that, although newer versions of generic fonts have glyphs for code point U+AB52, they often show faulty or insufficient rendering of diacritics (see Times New Roman ꭒ̆, ꭒ̃, ꭒ̱́, etc.). Standardization can be ensured in this case by defining a compatibility equivalence of code point U+028B with code point U+0AB52: <ʋ> ≡ <ꭒ>.

This peculiarity of the transcription system is motivated by French orthography which uses a grapheme <u> to denote the phonetic value [y] and a grapheme <ou> to denote the phonetic value [u] (see French mur [myʁ] ‘wall’ vs. tour [tuʁ] ‘tower’). Therefore, the French based transcription system of Rousselot (1887) introduced a new grapheme <ꭒ> with phonetic value [u] to avoid confusion with grapheme <u> with phonetic value [y].

3.3.6.4. Latin Small Letter C With Latin Small Letter H Inside

The TPPSR use a complex grapheme Latin Small Letter C With Latin Small Letter H Inside to denote the voiceless postalveolar fricative consonant [ʃ] (see also Rousselot 1887, 4). Because there is no adequate Unicode code point the grapheme is replaced with substitute grapheme <c̵> (LATIN SMALL LETTER C WITH COMBINING SHORT STROKE OVERLAY). There are the following four grapheme types: <c̵> (2106), <ç̵> (38), <ć̵> (2), <c̵̾> (419) with a total of 2565 tokens (Figure 9.3).

4 Retro-standardizing the TPPSR

While one may consider a pure digitization of the data presented in the TPPSR as an enterprise that is worth to be pursued by itself, recent advances in the computer-assisted curation and standardization of linguistic data enable us to lift the digital dataset of the TPPSR to a level where it can not only be conveniently inspected, but where it can also be compared with other datasets and integrated with other resources. This procedure, which we call retro-standardization, has been successfully applied to numerous lexical datasets in the past (Rzymski et al. 2020) and is supported by standard formats for data sharing (Forkel et al. 2018) and software packages that facilitate data curation (Forkel and List 2020). In the following, we will introduce the major steps that were needed for the retro-standardization of the TPPSR in more detail. After giving a quick background on the concept of data lifting and retro-standardization, we will show how an enriched, retro-standardized version of the TPPSR was created, and how this version was then deployed in the form of an interactive online application.

---

11 This phenomenon is historically motivated by the sound change from [u] > [y] in Early Old French without subsequent change of orthography.
4.1. Background on (Retro-)Standardization of Linguistic Data

Given the large number of languages spoken and signed, as well as the very small number of languages which have been thoroughly documented, linguists have long felt the need to standardize cross-linguistic data in order to allow for a quick access to all the relevant resources describing a given language variety. Despite long-standing standardization efforts, such as the establishment of the ISO 693-3 language codes\(^{12}\), most linguistic datasets are still largely incomparable from a cross-linguistic perspective. While the lack of comparability may be in part due to the problem of finding the right balance between what is generalizable and what is specific in linguistic transcription itself (Van Gysel et al. 2019), another important reason can be found in the great variety of objectives which motivate scholars to collect linguistic data (List et al. forthcoming, 1/15).

In order to increase the comparability of cross-linguistic data, the Cross-Linguistic Data Formats initiative (https://cldf.clld.org, Forkel et al. 2018) has proposed guidelines, standards, and best practice examples for the most fundamental types of linguistic datasets which are collected and curated in linguistics. Starting from smaller data collections illustrating how data could be rendered in the basic formats for wordlists and structural datasets proposed by the CLDF initiative, the application range and the tools supporting the generation, curation, and exploration of CLDF datasets has greatly increased over the past years. This increase is reflected in (a) large collections of lexical datasets, culminating in the third installment of the Database of Cross-Linguistic Colexifications (Rzymski et al. 2020, https://clics.clld.org, see also List et al. 2018 for a more detailed description of the database), (b) growing collections of structural datasets, including CLDF versions of popular structural datasets such as the World Atlas of Language Structures Online (Dryer and Haspelmath 2013), and (c) new application examples illustrating how additional data types can be modeled in CLDF (see List et al. forthcoming as an example of how to handle interlinear-glossed text in CLDF).

The basic formats proposed by the CLDF initiative are based on a tabular data model which – in its simplest form – consists of triples of language, parameter, and value. An example for a lexical parameter would be an elicitation gloss, as we find them in the TPPSR, but it could also be an elicitation phrase, or a structural item. In order to increase the comparability of languages, parameters, and values, CLDF allows linking certain aspects of the data to reference catalogs. Reference catalogs are linguistic meta-data collections providing information on basic entities of linguistic research. Glottolog (https://glottolog.org, Hammarström et al. 2020), for example, is a catalog for language varieties, providing identifiers for language varieties along with geographic locations and basic genetic classifications. Concepticon (https://concepticon.clld.org, List et al. 2020) is a catalog for concepts which are frequently elicited in linguistic field work, offering identifiers for concept sets which are themselves used to link lexical questionnaires and concept lists which have been proposed over the past four centuries by linguists, anthropologists, and psychologists (see List et al. 2016 and List 2018 for details)\(^{13}\). Finally, CLTS (https://clts.clld.org, List et al.

\(^{12}\) Revised in 2002, see https://www.iso.org/standard/22109.html.

\(^{13}\) The Concepticon itself is currently being expanded by additional datasets on norms, rankings, and relations from psychology and comparative linguistics, providing interesting additional
2019) is a reference catalog for speech sounds, providing identifiers for more than 8000 different sounds which have been documented in the linguistic literature and offering unique feature descriptions for each sound, as well as translating these sounds into different transcription systems (see Anderson et al. 2018 for details).

Preparing new or converting existing datasets into the standards proposed by the CLDF initiative goes beyond a mere standardization or retro-standardization of the original datasets, since the resulting data are integrated with reference catalogs offering additional information that was not available from the data themselves and also allows to compare the newly standardized data with other standardized datasets. Data are lifted to a higher level of internal and external comparability. Thanks to software libraries designed to handle CLDF datasets, they can be easily accessed and reused, thus conforming to the fundamental requirement of interoperability of research data14.

4.2.1. Languages, concepts, and phrases

When rendering a tabular lexical dataset such as the TPPSR in CLDF format, each lexical entry can be represented by a triple of language, concept, and form. Additional information on languages and concepts can be provided in additional tables. When retro-standardizing the information on the dialect varieties in the TPPSR, we therefore augmented the data by supplying Glottolog identifiers, providing additional information on the internal classification of the language varieties, and adding geographic information on the dialect locations (both in the form of geographic coordinates and the canton in which the variety was spoken). Given that the TPPSR provides rich information on the informants for each dialect variety, we also included this information, which details the year of recording, basic characteristics of the speakers (gender and age), their use of the dialect variety in question, and their proficiency (see TPPSR Online Dialects).

The list of 480 concepts in the TPPSR questionnaire was added to the Concepticon project by translating all French elicitation glosses into English (thereby also disambiguating those glosses which occurred several times in the data), and then linking those entries to Concepticon concept set identifiers, which could be most transparently identified. Since most entries in the TPPSR are still lexical phrases, even if larger phrases were broken up into smaller units, we refrained from linking ambiguous cases, also acknowledging that the questionnaire underlying the TPPSR was compiled for the purpose of documenting a particular subgroup of dialect varieties.

In addition to enriching the information on languages and concepts with internal information found in the non-tabular part of the TPPSR resource and external information found in Glottolog and Concepticon, we also added a detailed reference for each word form, indicating on which page it can be found in the original resource, and we added a table that

---

14 Interoperability is one of the four fundamental requirements for open data in scientific research, also known as FAIR data (findability, accessibility, interoperability, and reusability), as proposed by Wilkinson et al. (2016).
indicates how the original phrases can be restored from the concepts as they are rendered in
the individual tables of the original data. While the page references have the advantage of
allowing users to compare our digital forms with the original scans of the data directly in the
web application, the restoration of the phrases provides users with a convenient comparison
of specific language entries beyond the level of the word (or the noun or verb phrase).

4.2.2. Transliterating TPPSR Transcriptions to Broad IPA

The transcription system underlying the original TPPSR is an adaptation of the
transcription system of Rousselot and Gilliéron, which was widely used in Romance
dialectology at the time the TPPSR dataset was printed (see 3.2). The transcription system
makes some basic assumptions regarding the production of speech sounds, which date back
to Rousselot’s investigations on speech sound production in the late 19th century.

In order to make the lexical entries transcribed in the TPPSR transcription system
comparable to lexical entries in other datasets, it is important to transliterate the TPPSR
transcription to some standard system serving as a reference point. Here, the B(road) IPA
transcription system underlying the CLTS reference catalog is a very good starting point for
initial transliteration experiments, given the large number of sounds which the system
describes and the numerous datasets on sound inventories to which the catalog links.

CLDFBench allows for a convenient transliteration between transcription systems with
the help of orthography profiles as introduced by Moran and Cysouw (2018). An orthography
profile is a look-up table for a given transcription system, which essentially lists all
graphemes in the original transcription system (including graphemes composed of more than
one character) in one column, and corresponding transliteration values for user-defined
transcription systems in additional columns. In order to apply an orthography profile, a
greedy algorithm is used that starts from the left-most character of a given string, and then
proceeds searching for the longest matching graphemes in the orthography profile. Once the
grapheme has been identified, the corresponding transliteration value is appended to the
output string and the search is repeated from the rest of the string. Orthography profiles
provide two essential services: they allow to convert strings from one transcription system (or
orthography) to another one, and they segment a string into its basic graphemes. Especially
the aspect of segmentation is of great importance when dealing with lexical data, since it
allows us to identify those character combinations in the transcription which qualify as sound
segments.

While orthography profiles tend to appear more complex for the task of transliteration
than a well-chosen list of regular expressions, they have huge advantages due to their

15 After he had become acquainted with Étienne-Jules Marey’s (1830-1904) graphic techniques
for data display, Rousselot established the discipline of experimental phonetics with many
groundbreaking insights into tongue and lip movement, pharyngeal and nasal involvement,
degrees of palatalization, vowel length, etc. (see Schmidt 1907, 308s.).

16 We are aware that the question of how to segment a word into sounds does not necessarily
have a straightforward solution. However, we also know that for any linguistic treatment, be it
quantitative or qualitative, the question of the segmentation of a sound sequence into sound
segments is indispensable.
transparency and explicitness. Since the algorithm proceeds in a greedy fashion, the order of
graphemes is irrelevant, while the order of applications is important for regular expressions
and may drastically change the output. Since the algorithm is transparent, one can easily
understand why a given output was achieved when applying a given orthography profile to a
given text. Since the algorithm is easy to implement, different versions of the orthography
profile algorithm are now available in different programming languages (for R see Cysouw
2015; for Python see Moran and Forkel 2020; for JavaScript see List 2019).

Before converting the entries in the digitized TPPSR version into the B(road) IPA
transcription system underlying the CLTS reference catalog, we had to identify all graphemes
in the original data. Already this step required some specifically pragmatic decisions, since
we had to decide which units we wanted to treat as one sound and which we wanted to treat
as being composed of several sounds. For example, we had to decide how to treat diphthongs
and triphthongs, as they could be treated as two or more distinct vowels or as single units
instead. Given that CLTS offers a class for diphthongs but none for triphthongs, we decided
to render all combinations of two vowels as diphthongs, while the triphthongs where treated
as monophthong plus diphthong or diphthong plus monophthong sequences. Similar
decisions had to be made with respect to certain consonant combinations. All in all, this
yielded a list of 1195 unique graphemes in the data, which formed the basis of our
orthography profile. Of these 1195 graphemes, 795 (67%) are diphthongs, 265 (22%) are
vowels, 106 (9%) are consonants, 29 (2%) are triphthongs converted to diphthongs plus
vowels or other character sequences difficult to segment.

With all basic sound units assembled, we had to identify their closest counterparts in the
BIPA system of the CLTS reference catalog. Thanks to the very detailed descriptions of the
transcription systems underlying the TPPSR transcription system, this could be done in a
rather straightforward way, although it resulted inevitably in a certain loss of information,
due to missing specifications for certain fine-grained pronunciation distinctions in the BIPA
system in CLTS. As a concrete example, consider the grapheme <a> in the TPPSR
transcription system, which would be best specified as an open central unrounded vowel.
However, since there is no dedicated single-character symbol for this vowel in the
International Phonetic Alphabet (which is between the open front unrounded vowel [a] and
the open back unrounded vowel [ɑ]), linguistic transcription practice tends to transcribe it as a
simple open front unrounded vowel. In order to avoid further complications, we decided to
follow this practice in our conversion to BIPA, but we hope that we can provide more fine-
grained transcriptions in future versions of the digitized TPPSR.

---

17 The total orthography profile contains 1209 rows, since we occasionally list sounds occurring
only word-initially explicitly. It is furthermore possible that this number will change in future
versions, when the original digitization is further adjusted and the Remarques and Notations
divergentes are integrated.

18 A more correct (but barely used) transcription would be [ä] (with the COMBINING
DIAERESIS <◌̈> indicating the centralization of the front vowel). In the current version of the
CLTS transcription data, however, centralization is treated as a feature itself, and the sound is
therefore called centralized open front unrounded vowel.
The mapping of the original TPPSR transcription system to the BIPA system of the CLTS reference catalog yields a considerable reduction of phonetic complexity, with only 398 distinct sound segments remaining. However, given that the original transcription was quite impressionistic and phonetic in its nature, it is quite possible that our BIPA version of the TPPSR covers enough signal for future phonetic analyses of the datasets. It should furthermore be kept in mind that we did not render stress markers in our conversion (since we think that these should be better handled independently), and also deliberately ignored specific aspects, such as weak pronunciation and weak nasalization. We hope, however, that we will be able to handle these aspects more properly in future updates of the TPPSR.

4.2.3. Prosodic structures

The great advantage of having lexical data linked to standardized transcription systems such as the BIPA system of CLTS is that the entries can be easily processed by dedicated software packages. In order to introduce this, we computed what we call prosodic structures for each lexical form in the data. The basic idea of prosodic structures goes back to the hypothesis that the probability that a given sound will be affected by sound change processes depends on the prosodic environment in which the sound occurs. While sophisticated hierarchies for weak and strong prosodic environments have been proposed in the literature (Geisler 1992), we follow a very simplified version of this approach (List 2014) which only distinguishes three basic environments: increasing sonority (C), decreasing sonority (c), and vocalic peaks (V). We used the LingPy Python library (List et al. 2019, http://lingpy.org) to compute these environments from the segmented BIPA transcriptions and add them as additional information for each triple of language, concept, and form. Once prosodic structures have been computed, the lexical data can not only be conveniently queried for prosodic properties (for example by searching for all instances of CV inside the lexical entries for a given language), they can also be automatically analyzed for general prosodic properties (basic syllable structure, maximal onset, etc.).

4.3. Deployment with the help of CLLD

Exploiting the similarities of the cross-linguistically linked (CLLD) data model (Forkel and Bank 2014) and CLDF, we can easily ingest the TPPSR data into a CLLD application, thereby turning TPPSR into a true dialect atlas.

The current CLLD application serving our digital, retro-standardized version of the TPPSR offers three major perspectives on the data, a dialect perspective, a concept perspective, and a sentence perspective. The dialect perspective (https://tppsr.clld.org/languages/) provides the major information for all 62 dialect points in the dataset, including their geographic distribution shown on a map, their classification, and the year of recording. From here, individual dialect points can be selected and inspected along with more detailed information on the dialect variety, all lexical entries, the sound inventory, derived from our conversion to the Broad IPA, and the sentences reconstructed from the lexical entries (see https://tppsr.clld.org/languages/1).

The concept perspective lists the lexical data for all 480 items of the original questionnaire. When using this perspective to inspect individual items, data can be conveniently viewed on a geographic map or in tabular form. In addition, we also link the
original source as a scan, which makes the digitization process more transparent and allows users to verify the data in detail (see https://tppsr.clld.org/parameters/2).

The sentence view, finally, offers users the possibility to inspect the larger phrases which we reconstructed from the original questionnaire in due detail. In addition to rendering the data simply side by side, they can also be inspected in horizontally aligned form (see https://tppsr.clld.org/sentences).

5. Conclusion and Outlook

In this study, we have presented a digital version of the Tableaux Phonétiques des Patois Suisses Romands which does not only render the original dataset in digital form, but also enriches it by linking additional resources and adding sophisticated layers of semi-automated annotation. Thanks to this process of retro-standardization, the dataset can be straightforwardly presented in the form of an interactive CLLD application, which allows for a convenient exploration of the data both in the form of tables and geographic maps – fulfilling more than a hundred years later Gauchat’s dream of an Atlas linguistique de la Suisse romande.

Heinrich-Heine-University, Düsseldorf Heinrich-Heine-University, Düsseldorf
Max Planck Institute for Evolutionary Anthropology, Leipzig Max Planck Institute for Evolutionary Anthropology, Leipzig
Max Planck Institute for Evolutionary Anthropology, Leipzig Max Planck Institute for Evolutionary Anthropology, Leipzig
Robert FORKEL Johann-Mattis LIST

Bibliography


Wilkinson, Mark D. / Dumontier, Michel / Aalbersberg, I. J. / Brand Jan / Appleton, Gabrielle / Axton, Myles / Baak, Arie et al., 2016. «The FAIR Guiding Principles for scientific data management and stewardship», *Scientific data* 3, 160018, DOI: 10.1038/sdata.2016.18.

**Supplementary Material**

The digitized version of the TPPSR along with the retro-standardized data submitted in CLDF format has been curated on GitHub (<https://github.com/lexibank/tppsr/>) and archived with Zenodo (Version 1.0, <https://doi.org/10.5281/zenodo.3988472>). A digital copy of the original TPPSR has been archived with the Internet Archive (<https://archive.org/stream/gauchat-et-al-1925-tppsr>). The interactive CLLD application in which the data are presented for interactive inspection can be accessed via https://tppsr.clld.org.