Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 245–249
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Automatic Extraction of Verb Paradigms in Regional Languages: the case of the Linguistic Crescent varieties

Elena Knyazeva[1], Gilles Adda[1], Philippe Boula de Mareüil[1],
Maximilien Guérin[2], Nicolas Quint[2]

[1] Université Paris-Saclay, CNRS, LIMSI, Orsay, France
[2] LLACAN - UMR 8135 (CNRS/INALCO/USPC) Villejuif, France
{knyazeva, gadda, Mareuil}@limsi.fr
{maximilien.guerin, nicolas.quint}@cnrs.fr

Abstract

Language documentation is crucial for endangered varieties all over the world. Verb conjugation is a key aspect of this documentation for Romance varieties such as those spoken in central France, in the area of the Linguistic Crescent, which extends overs significant portions of the old provinces of Marche and Bourbonnais. We present a first methodological experiment using automatic speech processing tools for the extraction of verbal paradigms collected and recorded during fieldworks sessions made *in situ*. In order to prove the feasibility of the approach, we test it with different protocols, on good quality data, and we offer possible ways of extension for this research.

Keywords: language documentation, speech processing, dialectology, Linguistic Crescent, Romance languages

1. Introduction

An important and costly step in the process of language documentation is the transcription (total or partial transcripts) of speech data collected in the field. Several projects adopt a methodology involving the use of speech transcription systems (Adda et al. 2016; Michaud et al. 2018); in such an approach, it is necessary to adapt the systems so that they can transcribe (at least phonetically) speech collected during fieldwork. However, within the data gathered, some have either an approximate transcription (e.g. in the case of reading), or more or less precise information on its content, for example in the case of verb conjugations: the linguist proposes a verb, and the informant must give all the possible inflections, most often in a fixed order for tenses and persons. The question addressed in this paper is to explore whether it is possible to use a transcription system developed for a given language (here French) without precise adaptation of acoustic models, in order to produce both segmentation and transcription of verbal paradigms of a closely related language (here several Romance varieties spoken in central France), and the conditions under which the system will or will not require post-processing.

Verb conjugation is a major difficulty of the grammar of Romance languages. This holds true for the varieties spoken in the centre of France, in a transition area between *Oïl* and *Oc* varieties called *Croissant* 'Crescent', named after Ronjat (1913) because of its geographical shape, see Figure 1). Knowing that there are about 40 distinct types of verb inflections for a given local Crescent variety, to be multiplied by about 60 forms for different tenses, moods and persons, the descriptivist has to deal with at least 40×60 = 2,400 different verbal inflection for each local variety. Speech processing can facilitate and speed up the analysis of huge amounts of data collected in the field.

Within the framework of an ongoing project described in Section 2, many fieldwork sessions were done with native speakers in order to record these highly endangered varieties. In this paper, we will present automatic segmentation methods of recordings collected *in situ* in the linguistic Crescent, containing both Crescent and French data in order to extract the targeted content for linguistic studies, namely verb paradigms. The data consist of short recordings (typically less than one minute of speech) where the surveyed speaker conjugates a verb for all possible subjects in a given tense and mood.

In the ideal case, the classical order is followed: 1SG, 2SG, 3SG-M, 3SG-F, 1PL, 2PL, 3PL-M, 3PL-F.[1] An example, much less straightforward than in English (a poorly inflected language), is, in the commune of Dompierre-les-Églises (Guérin 2019:183): [i sori] 'I would know', [tə sorjɑ] 'you (SG) would know', [u sori] 'he would know', [al sori] 'she would know', [nə sorjɑ̃] 'we would know', [u sorjɛ] 'you (PL) would know', [i sorjɑ̃] 'they (M) would know', [al sorjɑ̃] 'they (F) would know'. The actual recordings however, are quite different: they may contain French (from the investigator), digressions (on the part of the interviewee, mainly in French), hesitations, errors (corrected or not), repetitions, a different order from the classical order, gaps, etc. The challenge is therefore to extract what interests us in the presence of these various types of noise and artefact. In the following, solutions are proposed, depending on whether the pronunciation of the searched paradigm is known (through previous descriptive work) or not, in which case we base ourselves on the paradigms already available for neighboring survey points. The two scenarios will be considered successively in Section 3 and will be evaluated in Section 4. Future

[1] List of abbreviations used in this paper: AZER = Azerables, CLFR = Cellefrouin, COND = conditional, F = feminine, FUT = future, HYP = hypothesis, M = masculine, PL = plural, SG = singular.

work will be envisioned in Section 5.

2. Background

Field work was carried out in about 30 survey points (or local varieties), 16 of which will be considered here: Archignat, Azerables, Bonnat, Chaillac, Châteauponsac, Cellefrouin, Dompierre-les-Églises, Dunet, Jouac, La Châtre-Langlin, Luchapt, Naves, Oradour-Saint-Genest, Prissac, Saint-Léger-Magnazeix, Saint-Sornin-Leulac (see Figure 1). All the informants are elderly people (mean age > 70) who are also fully proficient in French. With the exception of Naves and Cellefrouin, we have 2 to 5 informants per survey point. Regarding pronunciation, phonological systems may vary from one survey point to another, but in most cases the phonetic realisations hardly differ from (regional) French.[2]

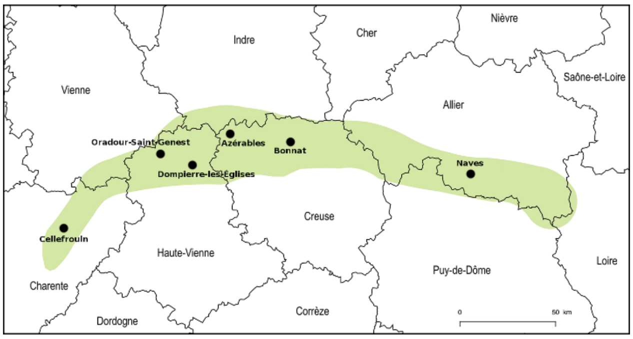

Figure 1: Map of the linguistic Crescent with the localisation of some of the surveyed varieties.

From a dialectological point of view, most of the 16 varieties considered here belong to two different groupings: (1) Archignat and Naves are *Bourbonnais* (i.e. eastern Crescent) varieties, showing closer affinities with Auvergnat Occitan; (2) most of the remaining points belong to the *Marchois* (i.e. western Crescent) group, which shows closer affinities with Limousin Occitan. In addition to Bourbonnais and Marchois, some survey points lying outside but in the close periphery of the Crescent have been taken into account in order to serve as control information. Among our 16-point sample, Châteauponsac represents a Limousin Occitan variety, spoken immediately South to the Crescent, while Dunet and Prissac represent Poitevin-Saintongeais (*Oïl*) varieties spoken immediately North to the Crescent.

All these varieties are being documented and described within the scope of two projects, namely *Les parlers du Croissant*[3] and VC2 - *Central Gallo-Romance: linguistics and ecology of a transitional zone*,[4] respectively funded by the French ANR (National Research Agency) and the Labex EFL. These two projects regroup a team of researchers — including the authors of the present paper — working in different branches of language sciences and studying the Crescent varieties within various approaches and frameworks.

The descriptivist linguists working in these projects have developed standard questionnaires[5]aiming at providing a set of comparable data for as many varieties as possible. The verbal paradigms considered in this paper have been collected resorting mainly to one of these questionnaires (Brun-Trigaud, Guérin & Quint, 2018) as well as to several monographs devoted to Crescent verbs (e.g. Guérin, 2019; Lavalade, 1987; Quint, 1991; 1996). At least 24 different verbs were administered to the informants for each surveyed variety: regular and irregular patterns (including asyllabic models such as /kwɑ(ː)/ 'brood' or /(l)j ɑ(ː)/ 'tie together (oxen)'[6]) as well as auxiliaries (the local equivalents of French *avoir* 'have' and *être* 'be'). Each verb paradigm was audio-recorded with a separate file for every tense or mood (see Section 1 above). Note that, in most cases, only one informant was recorded for each variety: as a matter of fact, the Crescent varieties are now on the verge of extinction and it is often not possible to find several active speakers for a given survey point.

3. Methodology

Under the assumption that the inventory of the dialect's phonemes is included in that of French, extracting verb forms pronounced in a Crescent variety could be defined as a task similar to searching for French words. However, adaptations of this technique were felt necessary. A system was consequently implemented, combining two recipes: automatic speech recognition (ASR) in French, provided by a public `git` repository and word spotting provided by the Kaldi distribution (Povey et al., 2011: `egs/babel/s5b/`).

The objective of ASR is to transcribe an audio stream or, more generally, to transform it into a lattice presenting the most likely pronunciation variants in a compact form. Three components are included:

- acoustic models (SGMMs learned from data collected within the framework of the LibriVox project: 13,620 sentences corresponding to 42 hours of read literary works);

[2] Some Southern Crescent varieties may exhibit phoneme inventories which happen to be much more at variance with French, including diphthongs such as /a^{ø}, aw/ and palatal plosives /c, ɟ/. However, such phonemes are absent from most of the varieties studied herein — or may be equated to French units for an automatic processing.

[3] http://parlersducroissant.huma-num.fr/projet.html

[4] http://www.labex-efl.com/wordpress/2020/01/15/vc2central-gallo-romance-linguistics-and-ecology-of-a-transitional-zone/?lang=en

[5] http://parlersducroissant.huma-num.fr/participer.html

[6] The so-called asyllabic pattern is observed for verbs whose lexical root does not include any vocalic element and therefore cannot be stressed as such. For instance, in most Crescent varieties /kwɑ(ː)/ 'brood' has an asyllabic root /kw/, contrasting with the regular verb /ʃɑ̃ˈtɑ(ː)/, whose root /ʃɑ̃t/ is syllabic and includes a vocalic element /ɑ̃/. When the stress is supposed to lay on the root in a given paradigm (e.g. 1SG present indicative), most syllabic roots remain unchanged (e.g. /i ˈʃɑ̃t/ 'I sing') while asyllabic roots necessarily undergo a change in order to host the stress (e.g. /i ˈku/~/i ˈkwe/~/i ˈkwø/… 'I brood' according to the variety considered). For more details, see Guérin (2019: 127, 129, 149-150), Quint (forthcoming).

- a language model, trained on the text of the same corpus using the IRST Language Modeling Toolkit (Federico et al., 2008);
- a French pronunciation dictionary provided in the `git` repository (including variants).

The recognition lattices, once they are obtained, are indexed in order to search for keywords. The result of the procedure is a list of hypotheses, each one containing the detected keyword, its start and end point in frames (one frame corresponding to 10 milliseconds), and the associated confidence score, which is the opposite of the posterior probability logarithm (see the examples below). Choosing the hypotheses is the final step of recognition. For all scenarios, French acoustic models are used.

3.1 Scenario 1: the Paradigm is Known

As the paradigm is known, we can supplement the language model and the pronunciation dictionary with this information. The language model is a mere unigram containing the French words as well as the searched verb paradigms, with constant weights (empirically set at 100, 1,000 or 10,000 times the weights of French words, depending on the dataset). The French pronunciation dictionary is supplemented with the searched paradigms, amalgamated with the associated subject personal pronouns, to avoid further confusions. For instance, for the verb *dire* 'say' in the Crescent variety of Cellefrouin (conditional present, singular subject), we added the following entries with their respective pronunciations in the pronunciation dictionary:

• dire-CLFR-COND-1SG	idiri
• dire-CLFR-COND-2SG	tidiri
• dire-CLFR-COND-3SG-M	udiri
• dire-CLFR-COND-3SG-F	adiri

An example of output of the system, for the pronunciation of 6 forms[7] of this verb in the conditional present, is as follows:

```
dire-CLFR-COND-1SG 197 254 0.0
dire-CLFR-COND-2SG 320 395 0.0
dire-CLFR-COND-3SG-F 473 550 0.0
dire-CLFR-COND-1PL 571 648 0.0
dire-CLFR-COND-2PL 678 750 0.0
dire-CLFR-COND-3PL-F 800 851 0.0
```

This example is almost an ideal case, because all paradigms were perfectly detected without any ambiguity. Yet, in other cases, we may have several hypotheses for the same form, as in the following example for the verb *voir 'see'* in the Crescent variety of Naves when inflected in the future tense:

```
voir-NAVES-FUT-1SG 95 169 0.0
voir-NAVES-FUT-2SG 217 302 0.0
voir-NAVES-FUT-3SG-M 324 407 0.0
voir-NAVES-FUT-3SG-F 451 538 3.89
voir-NAVES-FUT-3PL-F 451 544 0.02
voir-NAVES-FUT-1PL 565 648 0.0
voir-NAVES-FUT-2PL 695 783 0.0
voir-NAVES-FUT-1PL 825 913 0.0
voir-NAVES-FUT-3PL-M 957 1044 0.0
voir-NAVES-FUT-3PL-F 1082 1161 0.0
```

There may be three possible sources for these multiple hypotheses: (1) a form is repeated several times, (2) some forms are phonetically similar or identical to each other; (3) for some reason the speech recognition system failed to make accurate detections. To choose among these hypotheses, the confidence score could help us make a correct choice: for instance, the second 3PL-F form (from 1082 to 1161 frame) features a better confidence score (0.0 vs. 0.02). However, this is not always sufficient. In the previous example, the knowledge of the paradigm order allows us to choose the correct 1PL form, given that the two hypotheses have the same confidence score: the first hypothesis (from 565 to 648) is the only one correct with regard to the order, while the second hypothesis (from 825 to 913) results from a confusion with an alternative pronunciation for the 2PL form.

More generally, when selecting from various hypotheses, we added the following two structural constraints. (1) The conjugation is complete: if the 8 forms are not present (the feminine forms are quite often neglected, for instance[8]), this should be specified in a text file associated with the audio. (2) The remaining forms are pronounced in the classical order. These constraints, which are verified in the majority of high-quality recordings significantly contribute to improve the results, as exemplified above.

Technically, for each paradigm, a list of hypotheses with a confidence score greater than a given threshold is built. Then, these hypotheses are organised in a research lattice, where each path leading from the initial state to the final state consists of the hypotheses for the searched conjugation sequences. Only hypotheses compatible with the structural constraints are considered. Finally, the shortest path in this lattice is calculated, which enables the system to determine the best sequence of hypotheses.

3.2 Scenario 2: the Paradigm is Unknown

Scenario 2 represents a difficult task, namely when the verb paradigm of a given local variety is unknown. Yet, work done in nearby varieties may help to segment recordings from a new variety. The excerpt reported in Table 1 below shows that some forms may be repeated from one variety to another.

[7] In the Cellefrouin recordings, the speaker systematically omits 3SG-M and 3PL-M forms: therefore, the output contains only 6 persons.

[8] See however the previous note for an opposite case in which the masculine forms are omitted.

Cellefrouin	Dompierre-les-Églises	Oradour-Saint-Genest	Bonnat
i diri	i diri	i diri	i diri
ty diri	tə dirjɑ	tə diri	tə dirjɑ
u/a diri	u/al diri	ø/al diri	u/al diri
nə dirjɑ̃	nə dirjɑ̃	nə dirjɑ̃	nə dirjɛ̃
və dirje	u dirjɛ	u dirjɛ	u dirje
i/a dirjɑ̃	i/al dirjɑ̃	i/al dirjɑ̃	u/al dirjɛ̃

Table 1: Excerpt of known verb paradigms from four local varieties of the Linguistic Crescent.

Technically, for a new local variety (e.g. Azerables), a pronunciation dictionary is completed by combining the personal pronouns specific to this variety (e.g. *i, ti, u, al, n(ə), (v)u, i, al*) with the different verb forms of the surroundings.

- dire-AZER-HYP-COND-1SG idiri
- dire-AZER-HYP-COND-2SG tidiri tidirjɑ
- dire-AZER-HYP-COND-3SG-M udiri
- dire-AZER-HYP-COND-3SG-F aldiri
- dire-AZER-HYP-COND-1PL nədirjɑ̃ nədirjɛ̃
- dire-AZER-HYP-COND-2PL vudirje vudirjɛ
- dire-AZER-HYP-COND-3PL-M idirjɑ̃ idirjɛ̃
- dire-AZER-HYP-COND-3PL-F aldirjɑ̃ aldirjɛ̃

This approach presupposes a certain proximity between the different points, which can lead to errors. For instance, Naves (one of our two only survey points in Bourbonnais, in the East of the Linguistic Crescent), does not have many closely related varieties, as the majority of our sample comprises Marchois varieties (the Western part of the domain, see Section 2 above). This may result in a lower detection quality.

4. Evaluation

The results achieved were manually evaluated on two sets of data by two experts of the corresponding areas. The first one was collected at Azerables; 15 recordings were selected, each containing the 8 persons of the present indicative of different verbs.[9] The pronunciations of these paradigms were approximated by the known pronunciations of the other local varieties spoken in the same region (the set of transcribed varieties reported in Section 2). These data are used in the first evaluation protocol.

The second set of data was collected in Naves. In a similar way, we selected 15 recordings containing the 8 persons of the present indicative of different verbs.[10] As Naves is part of the set of transcribed local varieties, we designed two protocols with this second set of data: for the first one the exact pronunciation is known, while for the second one it is approximated as in the case of Azerables (for this reason, Naves was excluded from the set of varieties used in order to approximate the pronunciations of Naves). Table 2 summarises the three evaluation protocols.

protocol	place	pronunciations
1	Azerables	approximate
2	Naves	exact
3	Naves	approximate

Table 2: The different evaluation protocols.

The quality of the data collected for Azerables is comparable to that of Naves: both provide clean data where all forms[11] are pronounced properly, in the classical order, with few hesitations and unnecessary words.

The results are shown in Table 3. Regarding the correctly recognised paradigms, some happen to be segmented erroneously. The boundary problem may be addressed with appropriate post-processing; detections with imperfect boundaries are also an aid to manual processing by reducing the time required to extract a paradigm.

protocol	# of pronounced paradigms	# and % of paradigms correctly segmented	# and % of paradigms correctly recognised
1	118	80 (67.8%)	101 (85.6%)
2	120	112 (93.3%)	117 (97.5%)
3	120	65 (54.2%)	90 (75.0%)

Table 3: Results of the manual evaluation: (from left to right) the protocol, the total number of paradigms present in the processed recordings, the number and percentage of paradigms segmented correctly, as well as the number and percentage of paradigms which are recognised correctly but whose boundaries may be misplaced.

Here are some remarks regarding these results:

- By comparing the two protocols with approximate pronunciations (1 and 3), we can conclude that the results for Azerables (protocol 1) are better than for Naves (protocol 3). This can be explained by the fact that in the set of transcribed varieties, there are many localities whose varieties are similar to Azerables. Naves, on the other hand, is an atypical example (it is one of the two eastern Crescent varieties contemplated in this study[12]), so its pronunciations are less well approximated.
- By comparing the two protocols of Naves (2 and 3), we notice that knowing the exact paradigm (protocol 2) helps recognition a lot. The effect is further amplified by the fact that, as explained previously, pronunciation approximations for Naves are of poor quality.

[9] The first set consists of the following verbs: *acheter* ‘buy’, *aller* ‘go’, *avoir* ‘have’, *blanchir* ‘whiten’, *chanter* ‘sing’, *couver* ‘brood’, *devoir* ‘have to’, *dire* ‘say’, *être* ‘be’, *partir* ‘leave’, *pouvoir* ‘be able to’, *prendre* ‘take’, *savoir* ‘know’, *venir* ‘come’, *vouloir* ‘want’.

[10] The second set consists of the following verbs: *aller* ‘go’, *avoir* ‘have’, *couver* ‘brood’, *faire* ‘do’, *lier* ‘tie together (oxen)’, *partir* ‘go away’, *pouvoir* ‘be able to’, *prendre* ‘take’, *savoir* ‘know’, *tenir* ‘hold’, *vendre* ‘sell’, *venir* ‘come’, *voir* ‘see’, *vouloir* ‘want’.

[11] With a few exceptions (2 paradigms are missing in the Azerables set).

[12] The other one is Archignat, see Section 2 above.

- Finally, a comparison of the last two columns shows that boundary errors are relatively frequent, especially when the pronunciations are of lower quality. In the case of exact pronunciations, we find less than 5% of erroneous boundaries, while this rate rises up to over 15% in the case of better quality approximations (Azerables) and to more than 20% in the case of poorer quality approximations (Naves).

5. Conclusion and Future Work

In summary, we proposed a method for automatically extracting verb paradigms from audio recordings in Romance varieties spoken in France, in the area of the Linguistic Crescent. The searched pronunciations may be known a priori, in which case classical techniques can be improved by taking into account the particular structure of the data. When the verb paradigm is not known a priori, we may benefit from the knowledge of the conjugation in nearby survey points, in which case the quality of the results depends on the degree of similarity of the pronunciations of the varieties that are contemplated.

We can further relax the structural constraint when forms are missing or when the conjugation is not in the classical order. Future work will also combine speaker identification and speech recognition. Indeed, the information we are looking for comes from a native speaker of a given Crescent variety, but the system is sometimes disturbed by the interviewer who can repeat or even suggest verb forms himself. Diarisation can thus filter the speaker who interests us. More powerful neural acoustic models than SGMMs may also be used. Finally, further quantitative assessment of the system performance is highly desirable, as well as the development of a user interface to help linguists exploit the results presented herein.

6. Acknowledgements

This research is part of projects ANR-17-CE27-0001-01 (“The Linguistic Crescent: A Multidisciplinary Approach to a Contact Area between Oc and Oïl varieties”) and ANR-10-LABX-0083 (program “Investissements d’Avenir”, Labex EFL, Strand 3, Workpackage VC2 – “Central Gallo-Romance: linguistics and ecology of a transitional zone”). It contributes to the IdEx Université de Paris – ANR-18-IDEX-0001.

7. Bibliographical References

Adda, G, Stüker, S., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, F., Idiatov, D., Kouarata, G.-N., Lamel, L., Makasso, E.-M., Annie Rialland, Van de Velde, M., Yvon, F., Zerbian, S. (2016) Breaking the Unwritten Language Barrier: The BULB Project, *Procedia Computer Science*, 81:8–14.

Brun-Trigaud G., Guérin, M. & Quint, N. (2018). Questionnaire « Parlers du Croissant » — Conjugaison. Projet ANR « Les Parlers du Croissant ». http://parlersducroissant.huma-num.fr/docs/Croissant_Questionnaire_Conjugaison.pdf (last accessed 14/02/2020) or http://tulquest.huma-num.fr/sites/default/files/questionnaires/156/Croissant_Questionnaire_Conjugaison.pdf (last accessed 14/02/2020).

Can, D. & Saraclar, M. (2011). Lattice indexing for spoken term detection. *IEEE Transactions on Audio Speech and Language Processing*, 19(8):2338–2347.

Federico, M.., Bertoldi, N., Cettolo, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. *Proc. 9th Annual Conference of the International Speech Communication Association*, Brisbane, pages 1618–1621.

Guérin, M. (2019). *Grammaire du parler marchois de Dompierre-les-Églises (Haute-Vienne)*. Collection « Les Parlers du Croissant ». L’Harmattan, Paris.

Lavalade, Y. (1987). *La Conjugaison occitane* (*Limousin*). La Clau lemosina, Limoges.

Michaud, A., Adams, O., Cohn, T., Neubig, G., Guillaume, S. (2018) Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit, *Language Documentation & Conservation*, University of Hawaii Press, vol. 12, pages 393–429.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motl´ıcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. (2011). The Kaldi Speech recognition Toolkit. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Hawaïï, pages 1–4.

Quint, N. (1991). *Le parler marchois de Saint-Priest-la-Feuille (Creuse)*. La Clau lemosina, Limoges.

Quint, N. (1996). *Grammaire du parler occitan nord-limousin marchois de Gartempe et de Saint-Sylvain-Montaigut (Creuse)*. La Clau lemosina, Limoges.

Quint, N. (forthcoming). *La question des radicaux asyllabiques et des paradigmes verbaux de la 1re conjugaison dans le Croissant limousin et dans d'autres variétés occitanes ou romanes*. In Esher, L., Guérin M., Quint N. & Russo M. (eds), *Le Croissant Linguistique : nouvelles perspectives aux confins oc / oïl*. L’Harmattan, Paris.

Ronjat, J. (1913). *Essai de syntaxe des parlers provençaux modernes*. Imprimerie nationale, Paris.