Aesop's Fable "The North Wind and the Sun" Used as a Rosetta Stone to Extract and Map Spoken Words in Under-resourced Languages

Elena Knyazeva, Philippe Boula de Mareüil, Frédéric Vernier

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2072–2079
Marseille, 20-25 June 2022

Aesop's Fable "The North Wind and the Sun" Used as a Rosetta Stone to Extract and Map Spoken Words in Under-resourced Languages

Elena Knyazeva, Philippe Boula de Mareüil, Frédéric Vernier

Université Paris-Saclay, CNRS, LISN, Orsay, France
{knyazeva, mareuil, frederic.vernier}@limsi.fr

Abstract

This paper describes a method of semi-automatic word spotting in minority languages, from one and the same Aesop fable "The North Wind and the Sun" translated in Romance languages/dialects from Hexagonal (i.e. Metropolitan) France and languages from French Polynesia. The first task consisted of finding out how a dozen words such as "wind" and "sun" were translated in over 200 versions collected in the field — taking advantage of orthographic similarity, word position and context. Occurrences of the translations were then extracted from the phone-aligned recordings. The results were judged accurate in 96–97% of cases, both on the development corpus and a test set of unseen data. Corrected alignments were then mapped and basemaps were drawn to make various linguistic phenomena immediately visible. The paper exemplifies how regular expressions may be used for this purpose. The final result, which takes the form of an online speaking atlas (enriching the https://atlas.limsi.fr website), enables us to illustrate lexical, morphological or phonetic variation.

Keywords: semi-automatic spoken word alignment, linguistic atlas, under-resourced languages

1. Introduction

Since the Rosetta Stone, which enabled Champollion to decipher Egyptian hieroglyphs, parallel text translations have made it possible to construct concordances, for the Bible especially. The Bible, with its many translations, has also been used for developing text-to-speech synthesis systems in 700 languages (Black, 2019; Boito et al., 2020). The second most translated book in the world, *Le Petit Prince*, has been translated in over 400 languages or dialects, including Gallo-Romance varieties (Quint, 2022). Another story, "The North Wind and the Sun", one of Aesop's fables, has been used by the International Phonetic Association, for more than a century, to illustrate a number of languages and dialects — the difference between languages and dialects being ill-defined.

On the base of this Aesop fable, a speaking atlas of the regional languages of France has been designed (Boula de Mareüil et al., 2018), before being extended to the French Overseas (Boula de Mareüil et al., 2019a, 2019b) and various European countries. The online linguistic atlas, available at <https://atlas.limsi.fr>, allows visitors to hear and read this 1-munute story in over 800 versions, in minority languages or dialects. Through interactive maps, the aim of this atlas (which received over 800,000 visits) is to enhance the visibility of linguistic diversity among the general public and to provide an open access base for comparative studies.

The automatic processing of such audio data, collected in the field in low-resourced varieties, is not straightforward, as exemplified by earlier projects involving recordings collected in Africa (Adda et al., 2016), Asia (Michaud et al., 2018) or Europe (Knyazeva et al., 2020). In the latter work, the task consisted of extracting verb paradigms from conjugation lists recorded *in situ*, in dialectal varieties spoken in central France — the so-called Linguistic Crescent. In the present study, the aim is to extract translations of words such as 'wind' and 'sun' from versions of Aesop's fable recorded (and transcribed) in minority languages of France, to map them as in traditional linguistic atlases.

Most linguistic atlases, including speaking atlases, are limited to isolated words (Müller et al., 2001; Médélice, 2008; Tisato, 2010; Mutter and Wiatr, 2018). Here, innovative visualisation methods will be developed, centered on the Romance area on the one hand, on French Polynesia on the other hand — Romance and Polynesian languages presenting a certain homogeneity within each family. Since these languages are most often in a critical situation (Moseley, 2010), a pedagogical work is important to enhance their visibility.

With the advent of deep learning approaches in machine translation, interest in word alignment (Och and Ney, 2003) decreased. However, this task has again become a focus of research, more recently, for applications such as lexicon generation (Jalili Sabet et al., 2020; Wu and Dredze, 2020; Dou and Neubig, 2021; Imani et al., 2021). The algorithms developed, unfortunately, require large quantities of training data, which are not available for the languages under consideration.

After a presentation of the corpus and the method, the next sections will show how linguistic variables were selected: a dozen lexical items such as the words/concepts 'wind', 'sun', 'cloak' and 'traveller', for example, taken from over 200 translations of the Aesop fable (Section 2). On the basis of these items, the approach, which used word and phoneme alignment techniques, was evaluated on unseen data (Section 3). The user will be able to readily appreciate phonetic and lexical changes: isoglosses were displayed on maps, with different colour codes associated with different types and groupings of dialects. The paper will finish with these aspects, before concluding remarks and future work (Section 4).

2. Corpus and Method

2.1 Data

The case study considered here, the fable "The North Wind and the Sun", is a 5-sentence text, which totals 120 words

in French. Over 200 speakers of minority languages were recorded in recent years and their translations of this story were orthographically transcribed:

- 193 in Hexagonal France, in *Oïl* dialects (Picard, Gallo, Norman, Mainiot, Angevin, Poitevin-Saintongeais, Berrichon-Bourbonnais, Champenois, Burgundian, Franc-comtois, Lorrain and Walloon), *Oc* dialects (Gascon, Languedocian, Provençal, North-Occitan and *Croissant* 'Crescent'), Franco-provençal, Catalan, Corsican and Ligurian, in addition to non-Romance varieties;
- 14 in the five archipelagos of French Polynesia, in Tahitian (including its Reo Maupiti variety), Pa'umotu (in its Napuka, Tapuhoe, Parata and Maragai varieties), Rurutu, Ra'ivavae, Rapa, Marquesan (in its 'eo 'enana mei Nuku Hiva, 'eo 'enana mei 'Ua Pou, 'eo 'enata varieties), Mangarevan.

The fable begins as follows in French (middle), Languedocian (from Sète, above) and Provençal (from Forcalquier, below).

La cisampa e lo sorelh se fasián au pus fòrt la pelha
La bise et le soleil se disputaient
Se disputavan l'aura e lo soleu

From this excerpt ('The North Wind and the Sun were disputing'), the task basically consists of aligning the words *cisampa* and *aura* with *bise* ('North Wind'), *solelh* and *soleu* with *soleil* ('Sun'), and the idioms *se fasián au pus fòrt la pelha* and *se disputavan* with *se disputaient* ('were disputing').

2.2 Objective and Protocol

Since our objective is to offer a general public-oriented speaking atlas, not all the words are interesting to map. Function words, for instance, were discarded, as well as words exhibiting to little variation (e.g. *fort* 'strong'). Such words, however, could be used so as to improve the detection of other words. After various "trial and error", we decided to focus on a dozen words (hereafter, keywords): 12 tokens (10 lemmas) which, in French, are *bise* 'north wind', *soleil* 'sun', *manteau* 'cloak', *voyageur* 'traveller', *se disputaient* 'were disputing', *ôter* 'to take off', *ôté* 'took off', *souffler* 'to blow', *soufflait* 'blew', *briller* 'to shine', *réchauffé* 'warmed up', *reconnaître* 'to acknowledge'.[1] The first four items are nouns which are repeated in the fable (3 or 4 times); the following ones are verb forms for which it is more likely that no translation is found. The speakers were free to prefer circumlocutions which they found more idiomatic in their varieties, and some of them (only a few ones) produced fairly free translations.

The treatment we applied to Romance and Polynesian languages differed for three reasons:

- French and the other Romance varieties are close to one another, which makes it possible to consider specific strategies;
- verbs are not inflected in Polynesian languages (Lazard and Peltzer, 2000), making it irrelevant to map 'to take off' and 'took off', 'to blow' and 'blew', leaving only 10 items;
- since we started processing our data, one recording in a Polynesian language and 29 in Romance languages have been made. Only the 29 Romance versions were kept, as unseen data, for evaluation purposes.

We will thus focus on Romance languages in the remainder of this article.

Given the similarities among varieties, several clues help match correct translations.

- **orthographic similaritiy** (e.g. French/Occitan *soleil/solelh* 'sun', *se disputaient/se disputavan* 'were disputing', *voyageur/viatjaire* 'traveller', *réchauffé/ rescalfat* 'warmed up', *reconnaître/reconéisser* 'to acknowledge');
- **word position** within the text: although the source (French) text and the target texts (in regional languages) display different word numbers, the position of the words we want to align hardly changes in the translations. For example, 'were disputing' will most likely be present at the beginning of the text. Also, the words repeated several times are more easily identifiable ('north wind', 'cloak' and 'traveller' are repeated 4 times in the original French text, 'to take off' 2 times);
- **word context**: if word order hardly varies across translations, it is possible to detect a difficult keyword by exploring the target text, searching for the neighbouring words in the source text. For example, the phrase 'were disputing', often translated as an idiomatic expression, can be detected thanks to its position between the words 'sun' and 'each', generally easier to detect.

Various phenomena can make the alignment task more complicated. The texts may contain free translations (see above); some words may be absent and/or replaced by non-equivalent or shorter expressions (e.g. 'wrapped in his cloak' translated into *enmantelat* in a few Languedocian survey points); in a text, the same word can be translated in different ways (e.g. 'cloak' into *mantiau*, *hébit* and *gan-nèche* in the central *Oïl* dialect of Lavau). The ordering of the target text may be different from that of the source text (see the example from Forcalquier). The challenge is therefore to extract all the keywords, facing these different types of complications. The next subsection describes the suggested solution to this problem.

2.3 Algorithm

The task we are considering is similar to an alignment task; however, the very limited amount of data does not allow us to use the machine learning approaches cited in the introduction. In our case, it is still possible to perform automatic processing thanks to the properties of the source and target texts listed in the previous subsection. Finding only keyword translations has two advantages over full text alignment. First, fewer problems are expected with missing words: the targeted keywords carry meaning, therefore it is rare for them to be absent from the translations. Second, there are fewer occurrences of reordering: the order of

[1] So as not to weigh down the text, we are talking about words, but compounds of up to 4 terms may be included (see below).

meaningful words is generally defined by the story that is told. Some reordering linked to various grammatical constructions remains possible, but it is generally local.

The algorithm used for keyword alignment may be broken down into three main steps which will be detailed below: selection of candidate alignments, construction of the search lattice and local corrections.

2.3.1 Selection of Candidate Alignments

For each keyword present in the source text, the first step consists in selecting the 50 best alignment candidates among the target vocabulary — the number being chosen on the development set so as not to degrade performance while reducing execution time. For this doing, we defined an alignment score between two words which reflects both their orthographic proximity and the proximity of their occurrences in the source text and the target text. The first component of this score is the normalised Levenshtein distance, that is, the minimum number of edit operations (insertion, deletion, substitution) necessary to switch from a source word to a target word divided by the length of the shorter word.

The second component is calculated by measuring the Euclidean distance between the vectors of positions where words appear in the source and target texts. In order to minimise the impact of the variable text length, normalised positions were used, that is, word indices in the text divided by the text length. When source and target words w_s and w_t do not have the same number of occurrences, it is not possible to directly calculate the Euclidean distance between the two vectors. We then compute the minimum distance between the smaller of the two vectors and all the sub-vectors of the second vector of the same size as the first one. In this case, a penalty proportional to the length difference is added. These two components are combined to form what will be called the local alignment score, following Equation 1.

$$\text{score}(w_s,w_t) = \text{Euclidean}(w_s,w_t) - e^{1-\text{Levenshtein}(w_s,w_t)} \quad (1)$$

In the case of compounds, the score is calculated using a classical technique in the French language processing. For the Levenshtein distance, the terms composing the compound are combined with an underscore, the rest of the calculations being carried out as if it were a single word. For the Euclidean distance, the position of a compound was considered as the mean of the positions of the terms composing it.

2.3.2 Construction of the Search Lattice

The second alignment step consists in building the search lattice. This lattice only includes monotonic alignments between the keywords of the source text and the words of the target text, that is, without possible reordering. This constraint makes it possible to better exploit the context and will be relaxed during the local corrections step. The lattice is constructed as follows: each state represents a partial alignment in which the first *n* keywords of the source text either have been aligned with a word of the target text, or has received a "null alignment". The initial state is the state in which no keyword is aligned yet and the final states are the states in which all the source keywords have been aligned.

A transition between two states is possible if the arrival state represents the monotonic alignment of exactly one keyword more than the departure state, all other keywords being aligned in the same way. The weight associated with each transition is the local alignment score of the additional keyword as defined in Equation 1 — the null alignment is given a high positive score in order to minimise the number of empty alignments. A search for the shortest path in this lattice allows us to obtain an optimal alignment, under the monotonicity constraint.

2.3.3 Local Corrections

A third step relaxes the monotonicity constraint so as to capture richer alignments. For each keyword, a search among the neighbouring words of its current alignment is carried out in order to check whether a better alignment would not be possible. This search considers all of the words in the target text such that an alignment with the source keyword would create an alignment link that would not cross more than two other alignment links. If any of these candidates has a better local score, the monotonic alignment is replaced: in other terms, if the Levenshtein score is very high, we assume that the word has slightly changed position compared to the original text; consequently, we no longer try to optimise its position.

2.4 Forced (Audio-Text) Alignment

As summarised in Figure 1, the last step is to extract the audio segments corresponding to the best candidates selected during the previous steps. To do this, the forced alignment system developed at LIMSI (now LISN) was used, as in Boula de Mareüil et al. (2019b) and the alignment yielded between the text and the audio was evaluated. Given a text file and an audio file, the system provides the time codes of word start and phone start (or end). We also tested the WebMAUS system (Kisler et al., 2017) with French models, which gave very similar results.

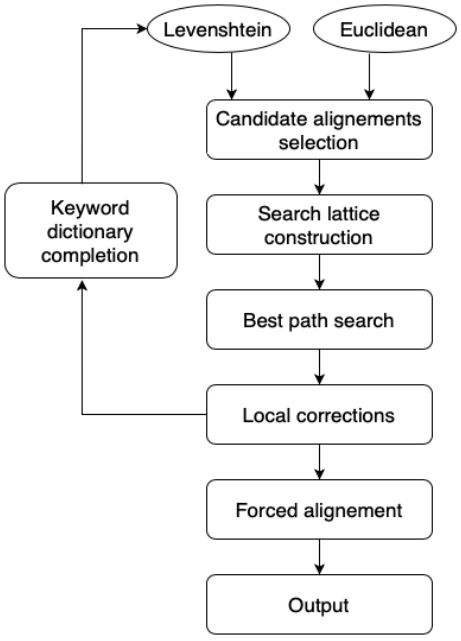

Figure 1: Block diagram of the algorithm.

3. Results

3.1 Enrichment of the Dictionary

The algorithm as described in subsection 2.3 gives good results when the target language is similar to the source language (at least from the point of view of orthography). When the spelling of a searched word is too far away from that of the source word, the position vector alone is unlikely to provide a good alignment, especially in the case where the word appears only once in the text. In contrast, close survey points often share the same spelling (or very similar systems). For example, the word 'were disputing' is translated *litigavanu* in a couple of northern Corsican survey points and *litigavani* in a couple of southern Corsican survey points. Thus, knowing a word translation in one of these survey points could greatly improve the detection in other survey points. To implement this mechanism, a dictionary was built with known translations for each keyword searched. The Levenshtein distance between the source keyword and the target word is then calculated as the minimum distance between this target word and all the variants corresponding to the keyword in the dictionary.

To complete this dictionary, the development set was used. After running the algorithm once with the dictionary containing only French words, we asked a linguist to do a quick pass on the written form of the result, to identify the words that seemed to be correct translations, so as to enrich the dictionary. This method has the drawback of requiring some work (a few hours) on the part of a linguist before the system obtains good performance.

We also explored the possibility of enriching the dictionary automatically with the candidates that have the best scores, but this method gives significantly poorer results — more than twice as many errors. In order to minimise the total time spent by the linguist, as in Knyazeva et al. (2020), we kept the manually checked variants. The results on the test set presented below suggest that new survey points can be processed with the same dictionary as the one created from the development set, without requiring additional intervention.

3.2 Evaluation

The various parameters of the system, such as the score function score(w_s,w_t), the number of candidate alignments per keyword, the crossing constraint of the correction step and a few others, have been optimised. on the development corpus composed of data from 164 survey points from Hexagonal France. The remaining 29 survey points, accounting for a dozen dialects, make up the test corpus used for assessing the system performance. The 12 keywords evaluated in Hexagonal France (by the same linguist as previously who, in case of multiple translations, indicated the "best" candidate on the base of the audio) are presented in Table 1. The survey points in French Polynesia are too few for a formal evaluation. Yet, only 5 errors were found in the development corpus, which suggests that the approach can be applied to unrelated languages.

The first two columns of Table 1 display the results of the keyword detection step. For each word, the evaluation reports the percentage of survey points for which at least one occurrence of this word has been correctly detected in the target text. The (rare) survey points for which the keyword does not appear in the target text are excluded. The next two columns of Table 1 indicate the percentage of accurately aligned keywords, in terms of audio segments, among the correctly detected keywords. Here, "accurately aligned" means that the alignment does not need to be improved by returning to the audio signal: "inaccurately aligned", in general, does not mean that the alignment is incorrect but that a better segmentation of spoken words was felt necessary for the speaking atlas.

%	Detection		Alignment	
	dev	test	dev	test
bise 'north wind'	100	100	84	93
soleil 'sun'	100	100	99	100
manteau 'cloak'	100	100	93	100
voyageur 'traveller'	99	100	98	97
se disputaient 'were disputing'	86	86	85	85
souffler 'to blow'	98	100	90	78
soufflait 'blew'	99	100	92	96
briller 'to shine'	93	93	92	96
ôter 'to take off'	95	96	94	85
ôté 'took off'	96	100	69	39
réchauffé 'warmed up'	97	97	94	100
reconnaître 'acknowledge'	97	100	94	86
Total	**96.7**	**97.7**	**90.3**	**88.1**

Table 1: Evaluation of the results in terms of correct word detection and accurate alignment with the audio signal, on both the development set and the test set (%).

The first four keywords are very well detected and fairly well aligned. As a matter of fact, they are repeated several times in the text, which makes them easier to find —in addition, a single correctly spotted occurrence is sufficient for our purpose. The only error on the development corpus comes from the word 'traveller', which was translated by *baguedenallour* in the Poitevin variety from Cherves: this form does not resemble any other local variant and, moreover, it is repeated only twice — the other two occurrences being synonyms of 'man'. Among the other keywords, two words are also very well detected: 'to blow' and 'blew'. A probable explanation is that the sentence encompassing them has been translated quite literally by the speakers. The lowest results were obtained for the verb 'were disputing', which was frequently translated by idiomatic expressions. For example: *étiant en traen de s'entrprendre* in the Poitevin variety from L'Épine, *étiont en baraille* in the Berrichon-Bourbonnais variety from Limoise.

Standing out from the French language could be a trait shared by some speakers of northern varieties (*Oïl* dialects closer to French). It therefore seemed interesting to compare their productions with those of speakers of southern varieties (Francoprovençal, Occitan, Catalan, Ligurian and Corsican). The results reported in Table 2 show better performance of the system on southern varieties. This outcome suggests that closely related varieties are not necessarily the easiest to process.

'%	Detection	
	dev	test
Northern varieties	93.9	94.1
Southern varieties	98.5	99.2

Table 2: Evaluation of the results in terms of correct word detection in northern and southern varieties, on both the development set and the test set (%).

As for the audio alignments, the relatively high correction rates (10–12%) are due to our desire to achieve the best possible quality, although we could have been satisfied with what the system produced, in a number of cases. Most of the manual corrections we made were to cut out a portion of [a] that was left in forms such as [(a) biz], coming from *la bise* 'the north wind'. Another example, also related with short words, is a portion of [s] that remains in forms such as *ôté son* 'took off his': a (quasi-systematic) problem that is encountered with the WebMAUS aligner as well.

3.3 Mapping

The extracted words were mapped, with their orthographic form, pinned to the coordinates (latitude and longitude) of the survey points of the speaking atlas. By clicking on the corresponding points, the audio versions can be listened to and a window opens, which indicates the name of the locality the recording comes from. A map was generated for each keyword, the base map of which could be modified to make certain linguistic phenomena immediately visible. Lexical variation (e.g. different translations for the word 'cloak' or 'coat'), morphological variation (e.g. verb endings for the imperfect) and phonetic variation (e.g. the evolution of a Latin phoneme) may be illustrated. This can be achieved by drawing Voronoi diagrams associated with different regular expressions. An algorithm was used in order to draw lines passing in the middle of two points of different categories. Areas (Voronoi cells) of the same category are then merged and coloured to form the final zones that appear on the map, associated with different word types.[2]

For instance, the words *bise*, *bize* or *bias* 'north wind' appear in the east and the south of France, whereas variants of *vent* 'wind' (*vint d'amont, vent de galèrne*) prevail in western France.[3] Regular expressions such as *bi-* vs. *v[ei]nt* can easily account for this variation, as shown in Figure 2. On this map, small regions with more specific translations such as *cisampa* (in Languedoc) and *aura* (in Provence), *galèrne* (to the west), *tramuntana* (in Corsica and Monaco) can be seen. These items may be captured by regular expressions which the linguist can easily change.

The regular expression may be very simple, to highlight a phonetic change: the fact that the Latin L of SOLEM (or SOLICULUM) 'sun' gave an intervocalic [l] or an intervocalic [r] in the dialects of France. The so-called rhotacism phenomenon is observed in the four corners of France: in Franc-comtois, Gallo, Gascon and a few varieties of the south-east, as shown in Figure 3. Other examples, such as the reflexes of the Latin morphemes –ELLUM (*-iau*, *-èl*, *-èu*, etc.) and –ATOREM (*-our*, *-aire*, *-ador*, etc.) will be illustrated during the conference.

4. Conclusion

The first goal of this paper was to present a tool developed to save the effort of linguists and relieve their workloads while designing a speaking atlas of minority languages, from recordings collected in the field. Two case studies were examined, giving rise to separate linguistic maps with various colour codes and captions: Romance France and French Polynesia. The principal challenge is the limited amount of data that does not allow the use of mainstream machine learning techniques.

The developed approach requires a limited participation of linguists to set up the system (help in creating a dictionary of keywords). This allowed us to achieve performance of 96–97% correct written words on our data of Romance languages/dialects (3–4% cases where words were not found correctly in the translated texts, implying that a linguist must intervene). The text-to-audio alignment system, on the same data, gave performance of 88–90% accurate segmentations (10–12% cases where the spotted word timecodes needed to be corrected manually by a non-linguist).

Note that the percentages should be interpreted with caution because they were obtained on relatively small samples. On the other hand, the small number of survey points in French Polynesia did not allow a formal evaluation. Yet, the approach may easily be generalised to other texts (we think of the Little Prince translated in a dozen varieties from the centre of France, although not freely accessible) and other languages. Combinations with learning methods to improve word alignment can also be considered. We hope that the resulting maps will find the same success with the general public as the speaking atlas on which this work elaborates.

5. Acknowledgements

This work was funded by the General Delegation for the French Language and the Languages of France (DGLFLF), within the framework of a project with the Cité internationale de Villers-Cotterêts. This work is part of the following projects, both funded by the French National

[2] As described in Boula de Mareüil et al. (2018), intermediate points are automatically added between frontier points of the same area. As they are merged afterword, this only slightly affects the layout when points are well aligned on both sides of a frontier; but when points are not well aligned, intermediate points complete the frontier and reduce the unwanted jigsaw effect along the frontiers.

[3] Using expressions with the word *vent* 'wind' is common in the west of France. To detect such expressions, a heuristic was added to the algorithm: if the word *vent* or *vint* is followed by the same word sequence at least 3 times in the text, it is assumed to be an idiom.

Research Agency: ANR-17-CE27-0001-01 (project "The Linguistic Crescent: A Multidisciplinary Approach to a Contact Area between *Oc* and *Oïl* varieties") & ANR-10-LABX-0083 (program "Investissements d'Avenir", Labex EFL, Strand 3, Workpackage VC2 – "Central Gallo-Romance: linguistics and ecology of a transitional zone"), and the project "*Oc/Oïl*: texts, identity and language contact" funded by the City of Paris. We are grateful to Jacques Vernaudon and Mirose Paia for their linguistic help, to Gilles Adda and Lori Lamel for the forced alignments.

6. 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., Rialland, A., Van de Velde, M., Yvon, F., Zerbian, S. (2016). Breaking the Unwritten Language Barrier: The BULB Project, *Procedia Computer Science*, 81:8–14.

Black, A. W. (2019). CMU Wilderness Multilingual Speech Dataset. *Proceedings of the 44th International Conference on Acoustics, Speech and Signal Processing*, pages 5971–5975, Brighton.

Boito, M. Z., Havard, W. N., Garnerin, M., Le Ferrand, É., Besacier, L. (2020). MaSS: A Large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible. *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 6486–6493, Marseille.

Boula de Mareüil, P., Rilliard, A., Vernier, F. (2018). A Speaking Atlas of the Regional Languages of France. *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 4133–4138, Miyazaki.

Boula de Mareüil, P., Adda, G., Rilliard, A., Vernaudon, J., Vernier, F. (2019a). A speaking atlas of indigenous languages of France and its Overseas. *Proceedings of the International Conference Language Technologies for All*, pages 155–159, Paris.

Boula de Mareüil, P., Adda, G., Lamel, L., Rilliard, A., Vernier, F. (2019b). A speaking atlas of minority languages of France: collection and analyses of dialectal data. *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 1709–1713, Melbourne.

Dou, Z.Y. and Neubig, G. (2021). Word alignment by fine-tuning embeddings on parallel corpora. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2112–2128, online.

Imani, A., Jalili Sabet, M., Senel, L. K., Dufter, P., Yvon, F., Schütze, H. (2021). Graph Algorithms for Multiparallel Word Alignment. *Procedings of the ACL Conference on Empirical Methods in Natural Language Processing,*, pages 8457–8469, online.

Jalili Sabet, M., Dufter, P., Yvon, F., Schütze, H. (2020). SimAlign: High qual-ity word alignments without parallel training data using static and contextualized embeddings. *Procedings of the ACL Conference on Empirical Methods in Natural Language Processing*, pages 1627–1643, online.

Kisler, T., Reichel, U. D. and Schiel, F. (2017). Multilingual processing of speech via web services, *Computer Speech & Language*, 45:326–347.

Knyazeva, E., Adda, G., Boula de Mareüil, P., Guérin, M., Quint, N. (2020). Automatic Extraction of Verb Paradigms in Regional Languages: the case of the Linguistic Crescent varieties. *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 245–249, Marseille.

Lazard, G. and Peltzer, L. (2000). *Structure de la langue tahitienne*. Peeters, Paris.

Médélice, J. E. (2008). Présentation du projet de l'Atlas Linguistique Multimédia de la Région Rhône-Alpes et des zones limitrophes (ALMURA) et commentaires du poster. In Raimondi, G. and Revelli, L., *Dove va la dialettologia?*, pages 199–205, Edizioni dell'Orso, Alessandria.

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*, 12:393–429.

Moseley, C. (2010). *Atlas of the World's Languages in Danger*. UNESCO, Paris.

Müller, M. L., Köhler. C., Kattenbusch, D. (2001). VIVALDI – ein sprechender Sprachatlas im Internet als Beispiel für die automatisierte, computergestützte Sprachatlasgenerierung und presentation. *Dialectologia et Geolinguistica*, 9:55–68.

Mutter, C. and Wiatr, A. (2018). The virtual research environment of VerbaAlpina and its lexicographic function. *Proceedings of the 18th Euralex International Congress*, pages 775–785, Ljubljana.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1): 19–51.

Quint, N. (2022). Les parlers du Croissant : un aperçu des actions actuelles de documentation et de promotion d'un patrimoine linguistique menacé. *Proceedings of the Journée annuelle de la Société de Linguistique de Paris*, to appear, Paris.

Tisato, G. (2010). NavigAIS – AIS Digital Atlas and Navigation Software. In Cutugno, F., Maturi, P., Savy, R., Abete, G., Alfano, I. (a cura di), *Parlare con le macchine, parlare con le persone*, pages 451–461, EDK, Torriana.

Wu, S. and Dredze, M. (2020). Do explicit align-ments robustly improve multilingual encoders? *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*, pages 4471–4482, online.

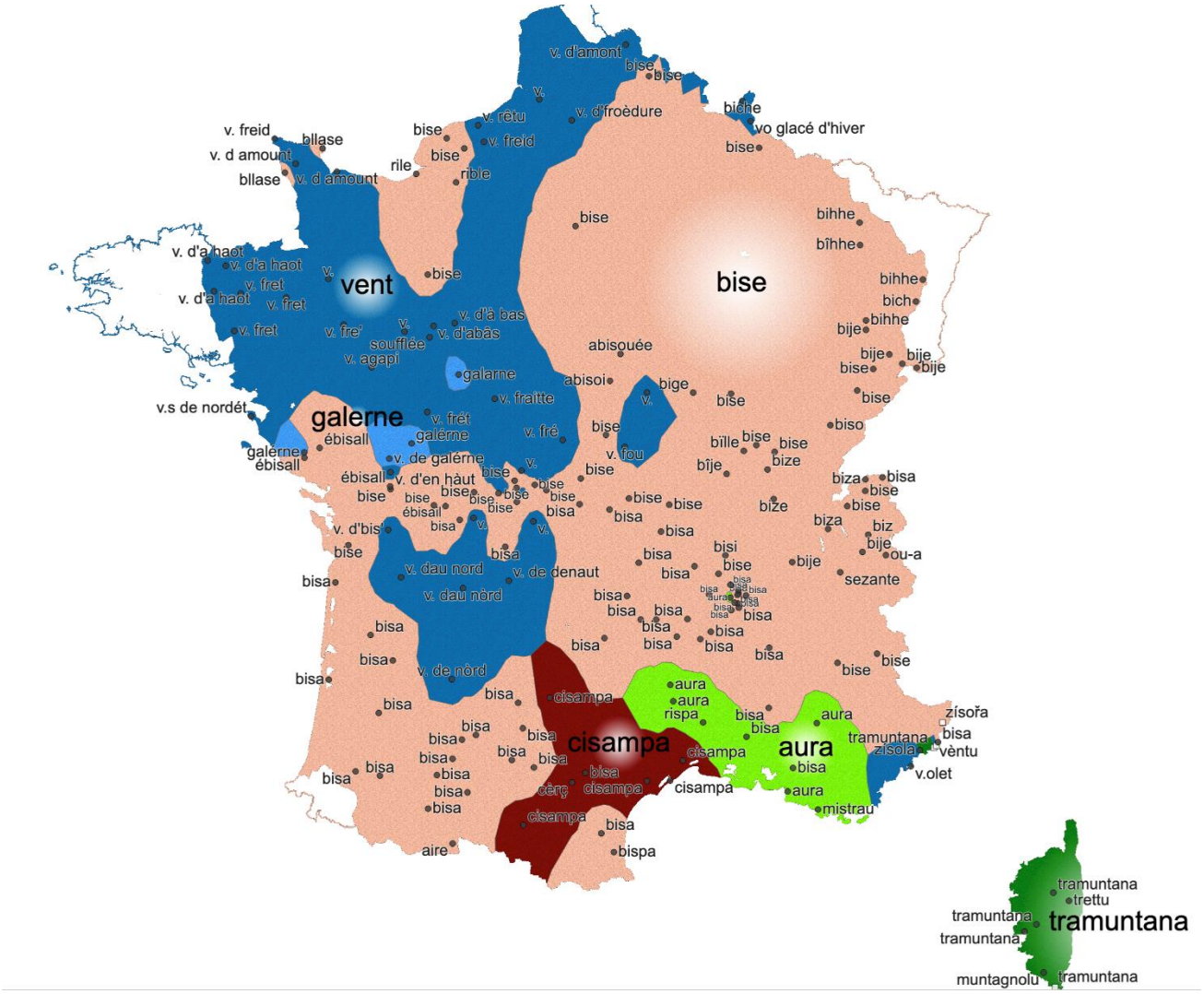

Figure 2: Linguistic map of Romance France for the word *bise* (‘north wind’), illustrating lexical variation. The abbreviation “v.” stands for *vent* or *vint* (‘wind’).

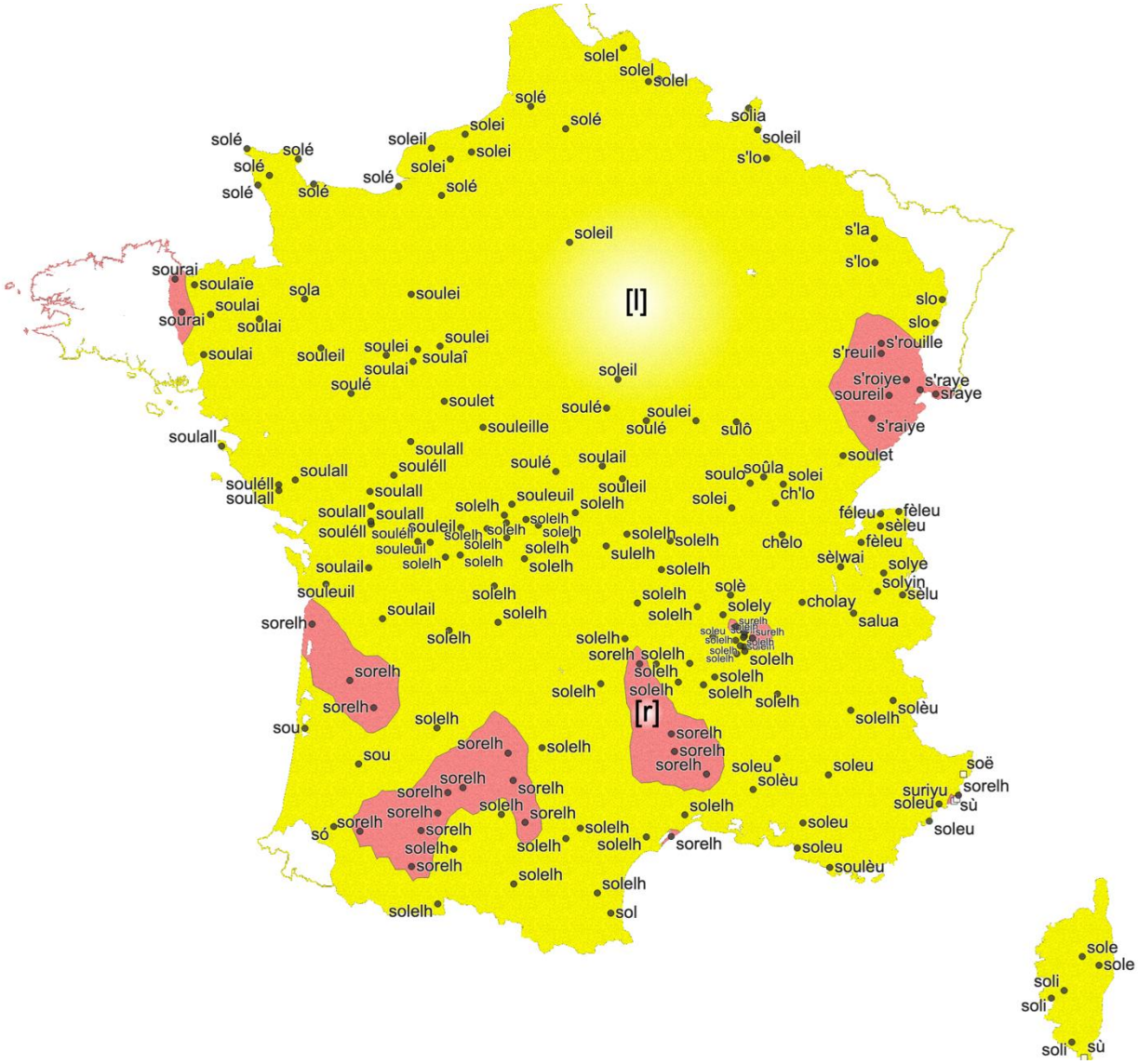

Figure 3: Linguistic map of Romance France for the word *soleil* (‘sun’) illustrating phonetic variation — with an intervocalic [l] in yellow and an intervocalic [r] in pink-red.