University of Dublin

TRINITY COLLEGE

Liaison Marker

A software to assist L2 French speakers improve their oral abilities by identifying liaisons within text.

Conor Evans
B.A. (Mod.) Computer Science, Linguistics and a Language
Final Year Project, April 2020
Supervisor: Elaine Uí Dhonnchadha

Acknowledgements

Thank you to my supervisor, Dr. Elaine Uí Dhonnchadha, for your consistent support and guidance.

Thank you to my parents for just about everything. Thank you to my siblings, too – Marc, Luke, and Maeve – for your attempts to feign interest. You are terrible actors, but it was appreciated nonetheless.

To Amyrose – thank you for proofreading and for your appreciation of the word eponymous. To Gonzalo and Amr – thank you for the advice while setting up the application environment. Moreover, thank you both, as well as Anna, Bogumił, John, Paul, and Robert, for all that I learned working alongside you.

I would like to extend particular recognition to Jules Buffet, French language assitant at Trinity College Dublin. Your contribution to the integrity of the manual marking process cannot be understated. A *grand merci*, if you will, for your insight into the liaison as a native speaker in doing so.

Finally, thank you to all those who have stoked my passion for the French language: Estelle Roullet, Stromae, Université Grenoble-Alpes, Enquête et Reportage, ARTE, Nicolas Mathieu, Haroun, to name but a few.

Abstract

A liaison is the realization of a liaising sound between two words, producing a two-word sequence which has a different phonetic make-up than if the two words were considered in isolation. The liaison is a well-studied French phonological phenomenon. It has been considered under a variety of scopes, from sociolinguistics to language acquisition among children or L2 speakers. However, thus far no tool has been publicly formalized to help this latter group improve their ability to identify liaisons. Liaison Marker takes on this mantle. Custom tokenization processes concerning numbers and punctuation are defined. The tokenized text is then grapheme-to-phoneme (GP) translated using a custom dictionary of regular expression patterns. Once the relevant phonological information is available, the application iterates over a series of liaison-inducing contexts to determine where to mark a liaison. The application draws on a wealth of existing analyses in the definition of its rule base and proposes a novel context. In cases of significant difference between academic observations and empirical observations, the application favours the latter to better emulate the modern French speaker that an L2 speaker likely wishes to embody. The analysis takes novel approaches to the historical-phonological divide with respect to initialisms and the *h muet/h aspire* (silent/aspirated h) dichotomy. The application eschews the traditional etymological approach to the latter and avoids an expensive dictionary look-up in the process. Instead, a series of regular expressions allow for highly accurate GP translation of h-initial words. Classification is measured with respect to accuracy and precision. The latter is championed as we do not want to guide users into realizing erroneous and/or forbidden liaisons. The application was tested on five news articles, each from a different source and each pertaining to a separate news section. A concrete comparison was made to liaison realization by French president Emmanuel Macron in his March 16th presidential address. In both cases, the application was able to identify the correct phonemes and liaisons to a high degree of accuracy and a near perfect degree of precision. Several areas of improvement were identified and a probabilistic interpretation is outlined that would allow the application to better inform users of the likelihood that a facultative liaison is realized.

DECLARATION

I hereby declare that this project is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university

CONOR EVANS _____ 30-04-2020

Name Date

Contents

1 Introduction

The application described and designed herewithin, Liaison Marker, is a French-language software application designed to help L2 speakers improve their oral language abilities. It is a further, more practical, development of previous work in the area (Evans, 2019). Described in further detail in Section 2.3, a liaison is the realization of a liaising sound between two words, producing a two-word sequence which has a different phonetic make-up than if the two words were considered in isolation. It is thus important that any speaker can identify liaisons when reading a text.

The liaison can be a source of difficulty for many L2 students. Many analyses in the area have focused on the perception of the liaison and the difficulty for L2 speakers to accurately parse liaisons in speech. Saunders (1988) proposed that the "variability of spoken French" encumbered L2 students in "breaking down a continuous sequence into its discrete units". These difficulties extend to L2 speech production. In a study undertaken by Alain Thomas (2002), L2 Canadian speakers realized obligatory liaisons at a rate of 91.1%. Their *bourgeois* Parisian counterparts realized obligatory liaisons at a rate of 96.9%. The difference is statistically significant but the rate of liaison is high in both groups. However, the gap is much larger in certain contexts:

Category	N	%	N	%	
Adjective	354	93.2%	348	49.7%	(QUAL.)
			154	54.5%	(petits..)
			28	25%	(grands..)
			311	91%	(NUM.)
Conjunction			64	40.6%	(quand..)
Total obligatory	2667	96.9%	7395	91.1%	

Table 1: Obligatory liaison realization by native speakers (left) and L2 speakers (Thomas, 2002, p. 104).

Liaisons involving qualificative adjectives were realized at 93.2% by native speakers compared to just 49.7% by L2 speakers. This result is particularly significant given the similar sample sizes – 354 vs 348. Although there is no comparison figure, we can see that just 25% of the 28 sampled L2 speakers realized the obligatory liaison following the adjective *grands* 'big', e.g. *les grands/enfants*[1] 'the big children' [le.grɑ̃.**z**ɑ̃.fɑ̃]. This application will seek to propel L2 speakers to a level of obligatory liaison realization similar to

[1]The slash denotes a liaison between the two words.

that of native speakers.

Thomas' findings also identified differences in facultative[2] liaison realization among L2 speakers. Each of the following verbal forms may liaise, but it is not obligatory:

	Native		**L2**	
Facultative	**N**	**%**	**N**	**%**
est	2569	97%	1297	66.2%
sont	279	86%	154	51.2%
suis	139	47%	28	25%
était	364	75%	104	5.8%
ont	381	75%	27	29.6%
Total	3732	76%	1610	35.56%

Table 2: Facultative liaison realization for select verb forms (Thomas, 2002, p. 105).

Logically, the non-realization of facultative liaisons is not ungrammatical and thus one might not consider it an area particularly important to L2 speakers. However, this application aims to serve all levels of speaker and facultative liaisons are of particular interest to intermediate-advanced L2 speakers. Table 2 supports this aim. Native speakers realize a liaison after *est* and *sont* in 97 and 86 per cent of instances respectively. These are levels comparable to and even beyond several of the obligatory liaisons in Thomas' analysis. *Était* was liaised by just 5.8% of L2 speakers compared to 75% of native speakers. Although the sample sizes are not similar, N=104 is a large enough sample size to pique our interest.

Clearly, obligatory liaisons are the starting point for any L2 speaker. However, it is arguable that facultative liaisons really propel an L2 speaker in their quest for oral fluency. The above 'facultative' liaisons realized at a rate approaching 100% are particularly notable as all five are forms of *être* 'to be' or *avoir* 'to have' – the bread-and-butter of most languages. Cases two and five are also notable because they are also the third-person plural form of the auxiliary verb in any *passé composé* (past perfect) clauses:

1. *Ils ont/adopté un chien.*

   [il.zɔ̃.**t**a.dop.te.œ̃.ʃjɛ̃]

   They adopted a dog.

[2]Optional, non-obligatory.

2. *Ils sont/allés au cinéma.*

   [il.sɔ̃.**t**a.le.o.si.ne.ma]

   They went to the cinema.

It is therefore important that the application can identify both obligatory and facultative liaisons. It will be of particular interest to devise some process by which the application can determine the 'facultativeness' of a given liaison, i.e. just how facultative it is – or is not, in the case of *est*.

The particular personal relevance of liaison identification is the primary motivation behind this project. The initial research was completed in the aftermath of a nine-month Erasmus placement in Grenoble, France. Liaison identification was something I worked particularly hard at in my bid to improve my pronunciation, and hopefully Liaison Marker will be a veritable tool for many fellow L2 speakers. Facultative liaisons are noteworthy as there is not always a satisfactory answer behind the realization, or lack thereof, of certain liaisons. Whereas obligatory liaison is determined by grammar, facultative liaison is more independent of grammarians and the rate of realization can be determined by numerous factors such as social setting or socio-economic circumstances.

2 Background

The first subsection outlines the key linguistic concepts relevant to this analysis. The second subsection introduces the technologies through which the application will be developed and explains the motivating factors for using these technologies. Finally, we discuss the metrics by which the application will be evaluated.

2.1 Linguistic Concepts

2.1.1 Syllable

Arguably one of the most accessible linguistic concepts, i.e. understandable to the layperson, syllables are in fact remarkably dense and complex. Syllables can be broken down into two components: onset and rhyme. This latter component breaks down into two further components: nucleus and coda. As one would imagine, the nucleus is at the core of every syllable. The two are in a symmetric relation; there can be no syllable without a nucleus and no nucleus without a syllable. As a basic but not strict rule,[3] the

[3]Syllabic consonants exist (Ridouane, 2008). An example of this in American-English would be the [ɬ] in bottle [bɑ.tɬ].

nucleus is filled by a vowel. By extension of the symmetric relation, a syllable cannot exist without a nucleus which cannot exist without a vowel, therefore a syllable cannot exist without a vowel.

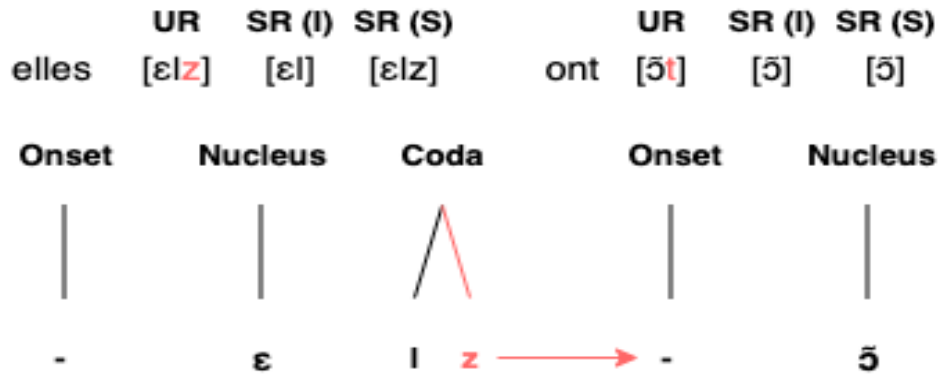

Figure 1: Syllabic breakdown of sequence 'elles/ont' [*el.zɔ̃*].[4]

The terms SR and UR in Figure 1 refer to Surface Representation and Underlying Representation respectively. These are concepts that will underpin the manner in which this application considers the liaison. A word may have many SRs. Édith Piaf's « *Non, je ne regrette rien* » is famous for the rolled Rs in its eponymous line. This differs from the typical uvular R in most SRs of these words, but is no less French nor grammatical for that matter. The UR for these two representations is, however, the same. Generally, a sequence will be realized in the same way as its UR in Standard French. This is not the case, however, for words which may be liaised. In Figure 1, the bracketed I refers to the SR of the word when considered in *isolation*, whereas the bracketed S refers to the word's SR when considered as part of the two-word *sequence*. It is this latter SR that is represented in the syllabic breakdown. We see the first sequence retains the latent [z] of its UR. The second sequence does not retain the latent [t] of its UR. The historical evolution of this SR/UR divide and why certain elements may or may not be realized will be examined in subsequent sections.

2.1.2 Hiatus

A linguistic concept less familiar to the average speaker, a hiatus is defined by the Oxford English Dictionary as "a break between two vowels coming

[4]Note that an empty coda is not denoted in syllabic breakdowns, whereas onsets are represented whether filled or not. This is because the onset is considered to not truly be 'empty'. There is always some onset of breath or glottal excitation that does not necessarily contribute to its first phoneme.

together but not in the same syllable" and derives from the Latin word for 'gap'.[5] Hiatuses arise when two nuclear elements find themselves on either side of a syllabic boundary. In order for this to occur, the first syllable must have a null coda and the second syllable must have a null onset. Hiatuses manifest in two manners:

Inter-word	**Intra-word**
coopérer	*sous un*
[kʊ**.**ɔ.pe.ʁe**]**	**[**sʊ**.**œ̃**]**
cooperate	below a

Liaisons are realized only in the second scenario and thus Liaison Marker will not be seeking to model both cases. This is beneficial as it makes grapheme-to-phoneme (GP) translation a less onerous task. If liaison only occurs at inter-word or *sequence* boundaries, the only phonetic information required to determine the presence of a liaison is the first and final phoneme of each word. Not only does this entail fewer GP translations, it also greatly reduces the phonetic contexts in which the translations must be considered. Consider *secondaire* 'secondary' [sə.gɔ̃.dɛʁ]. In sequence-initial position, *c* is generally realized as [k], e.g. *content* 'content' [kɔ̃.tɑ̃]. At sequence end, *c* may be realized as [k], e.g. *arc* 'ark' [aʁk], or as [], e.g. *escroc* 'crook' [e.skʁo]. In both cases, it may translate to [k]. However, in *secondaire*, the voiceless [k] becomes voiced as [g] as it is both preceded and succeeded by a (voiced) vowel. This is not a context that Liaison Marker needs to model.

2.1.3 Liaison

Liaison comes from the French word *lier* 'to bind'. As such, it could be considered a consonantal sound which binds the two vocalic sounds on either side of the syllabic boundary. It is not unreasonable, thus, to consider this consonant as being 'inserted' at the syllabic boundary to link the vowels and, by extension, see to the disappearance of the hiatus. Published academic literature has even used such terminology (Côté, 2005, p. 79). This application, however, will consider the liaising consonant to be a latent element of the sequence realized when necessary, i.e. when a hiatus occurs, but omitted for fluidity in the flow of speech when superfluous. This is in keeping with Optimality Theory and other Generative Phonologies that contrast SRs and URs. Latent liaising consonants are present in the UR but may or may not be present in the SR – this is determined by

[5] `https://www.lexico.com/definition/hiatus`

whether its "union with the following vowel-initial word" is strong enough to be preserved (Delattre, 1947, p. 148).

2.1.4 Enchaînement

Although a French term, *enchaînement* is a phonological phenomenon that exists in many languages, including English. Much like liaison comes from the verb *lier*, 'to bind', *enchaînement* comes from the verb *enchaîner*, whose meaning is much the same. The Larousse dictionary includes *lier* or a noun equivalent *lien* in three of the four entries for *enchaîner*.[6] However, the two terms must not be confounded, particularly with respect to this application. Both phenomena relate to the flow of speech. Tranel notes that it is generally accepted that CVCV[7] sequences are "ideal" in speech chains (2000, p. 40). These CVCV sequences are indeed favoured by French infants in the babbling stage of linguistic development (Cissé et al., 2014, p. 10). While the liaison is a process whereby a hiatus is avoided – transforming a part of the speech chain from VV to VCV – this is not true for *enchaînement*. This latter can be considered part of the liaison process. Both see a sequence-final consonant brought forward to the onset of the following sequence in accordance with the Maximal Onset Principle (MOP) (Weisler & Milekic, 2000, p. 47). The difference, however, is that this sequence-final consonant is a latent element of the first sequence in a liaison, whereas it is an explicit element of the first sequence in *enchaînement*. Consider the following examples:

			T1	**T2**
3. *sur un*	on a	[sy.ʁœ̃]	sur [syʁ]	un [œ̃]
4. *les amis*	the friends	[le.**z**a.mi]	les [le]	amis [a.mi]
5. *les copains*	the friends	[le.ko.pɛ̃]	les [le]	copains [ko.pɛ̃]

Each T1 and T2 represent the phonetic transcription for that sequence when considered in isolation. In (3), the [ʁ] in *sur* is present in isolation as well as when considered as part of the overall sequence, albeit in a different syllable as a result of the MOP. This is a case of *enchaînement*. In (4), the latent [z] of the UR for *les* is retained and liaises with the following sequence. It has also moved forward in accordance with the MOP. This latent [z] is not needed in (5), in which the word following *les* is consonant-initial, and thus [z] undergoes elision. Overall, distinguishing between cases of liaison and cases of *enchaînement* is rather important. We do not want

[6] `https://larousse.fr/dictionnaires/francais/encha%c3%aener/29149?q=enchainer#29023`

[7] C for consonant, V for vowel.

to mark every case of *enchaînement* even if the end result of linking two sequences may be the same. One aim of this application is to help L2 speakers develop the ability to recognise patterns as to when liaisons must (obligatory) or may (facultative) be realized. If they are presented with information regarding *enchaînement*, this will naturally be of detriment to this pattern recognition.

2.1.5 Grapheme

Graphemes are commonly defined as the written representation of phonemes (Rey et al., 2000, p. 1519). Graphemes and letters can have a one-to-one correspondence. The three-letter word *ami* 'friend' has three graphemes and, by extension, three phonemes: a/[a], m/[m] and i/[i]. However, this does not have to be the case. *Eaux* 'bodies of water' has just one grapheme and one phoneme: eaux/[o]. As such, graphemes provide a "more direct mapping from orthography to phonology" than letters (idem, p. 1520).

2.2 Technology Stack

The application is developed through the Ruby on Rails framework. This section discusses the advantages of this framework. It also explains the relational model used to model the application data and introduces some additional facilitative technologies.

2.2.1 The Ruby on Rails Framework

Ruby on Rails (Rails) is an open-source framework first created by David Heinemeier Hansson in 2003. Its founder describes it as a "web framework that's optimized for programmer happiness and beautiful code".[8] Rails is used by major technology companies such as Netflix,[9] AirBnB,[10] and Github.[11] The programming language through which it is primarily implemented is Ruby, created in the 1990s by Yukihiro Matsumoto as a language that is "simple in appearance, but is very complex inside, just like our human body".[12] Ruby's syntax is very accessible and it allows for very

[8] `https://dhh.dk/`

[9] `https://www.youtube.com/watch?v=YUX2tl3BMpw&feature=emb_title`

[10] `https://www.forbes.com/sites/quora/2018/02/20/what-technology-stack-does-airbnb-use/`

[11] `https://github.blog/2019-09-09-running-github-on-rails-6-0/`

[12] `http://blade.nagaokaut.ac.jp/cgi-bin/scat.rb/ruby/ruby-talk/2773`

concise and clean code. For example, *unless*[13] makes the typical *if NOT* statement more natural.

Rails is a typical Model, View, Controller (MVC) framework:

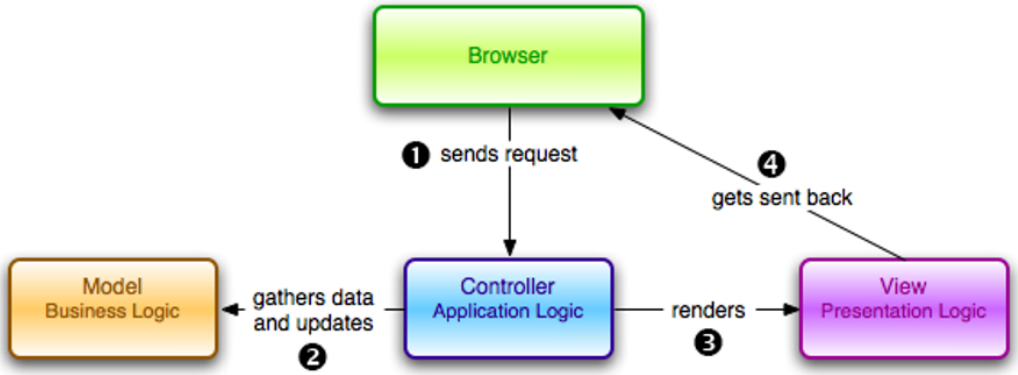

Figure 2: interaction between user and the components of the MVC framework.[14]

The first two components are almost instantly accessible to the layperson. A view is what an application user sees on-screen and that with which they interact. Views can be considered to be the front-end of this full-stack framework.[15] A model is some phenomenon related to the application functionality. The prototypical example of a Rails application, a blog,[16] would have *posts* and *comments* that need to be modelled. The final component of the MVC initialism is concerned with control. The application receives user input via the controller. A controller must control only one model but may interact with other models within its control processes. The core functionality of a controller is typically some iteration of CRUD[17] actions, with the option of extending functionality as desired. These CRUD actions involve database interaction and the controller verifies that the data sent by the user to a given controller action is valid, i.e. whether the changes persist[18] to the database. The full-stack nature of Rails greatly facilitates the

[13] `https://ruby-doc.org/docs/keywords/1.9/Object.html#method-i-unless`

[14] `https://www.sitepoint.com/model-view-controller-mvc-architecture-rails/`

[15] Full-stack frameworks incorporate both front-end (user-side) and back-end (application-side) functionalities.

[16] `https://guides.rubyonrails.org/getting_started.html#creating-the-blog-application`

[17] Create, Read, Update, Destroy.

[18] Persistence refers to whether the actions on a given database item have successfully saved to the database.

transfer of information from the application back-end to the application front-end. Embedded Ruby[19] is a powerful tool that allows application developers to embed Ruby code and Rails functionality (back-end) within CSS, HTML, JavaScript, or JSON files (front-end).

2.2.2 Data Management

Rails supports several Relational Database Management Systems (RDBMS), such as SQLite, MySQL, or PostgreSQL. The benefits of choosing one RDBMS over another depend on the features of the RDBMS itself – each RDBMS is fully integrated with Rails and "regardless of which database system you're using, the [core functionality of Rails] will always be the same".[20] This application is built on a MySQL database as any commonly cited disadvantages – such as scalability[21] [22] – are not pertinent. First proposed by E.F. Codd (1970) of IBM, the relational database model maps relations between the various entities of an application. Each entity is assigned its own table where records are stored. In Rails, these entities are referred to as models. As such, every model has its own table. The relational nature of the entities is represented by *foreign_key*[23] attributes in MySQL, and by 'associations' between models in Rails. These associations do not need to be explicitly defined within a model but make common operations "simpler and easier in your code".[24] They are, in essence, a stored SQL query. Consider the following one-to-one relation:

```
class Student < ApplicationRecord
  has_one :profile
end

class Profile < ApplicationRecord
  belongs_to :Student
end
```

Figure 3: Simple classes representing a Student and their student Profile.

[19] https://apidock.com/ruby/ERB
[20] https://guides.rubyonrails.org/v2.3.11/active_record_querying.html
[21] https://www.datarealm.com/blog/five-advantages-disadvantages-of-mysql/
[22] http://makble.com/the-advantages-and-disadvantages-of-mysql
[23] https://dev.mysql.com/doc/refman/5.6/en/create-table-foreign-keys.html
[24] https://guides.rubyonrails.org/association_basics.html

By defining this one-to-one relation, represented on each model by *has_one* and *belongs_to* respectively, calling *student.profile* will return the profile object associated to a given student by executing the following queries:

```
SELECT `students`.*
FROM `students`
WHERE `students`.`id` = 855 LIMIT 1;

SELECT `profiles`.*
FROM `profiles`
WHERE `profiles`.`student_id` = 855 LIMIT 1;
```

Figure 4: SQL queries equivalent to what Rails executes when ***student.profile*** *is called for a student of ID 855.*

Models in Rails are much more dynamic than their entity equivalent in SQL. Both can implement declarative constraints. These are constraints that can be defined within the database schema on a table or its attributes. These are defined mostly within the schema, but Rails validations[25] can supplement these constraints in areas where SQL is limited. For example, if storing a student's mobile phone number within their profile, it might seem logical to define an INT data type constraint for the number 0851234567. However, SQL removes zero-padding on numbers by default and solving this issue is the topic of many StackOverflow threads.[26] The Rails numericality validation[27] allows users to instead define the column as a string[28] to avoid the loss of leading zeros while asserting the validity of the phone number. The dynamism of Rails models is further shown in the ability to implement application-based constraints. These multi-table constraints are generally referred to as triggers in MySQL. While best practice is to use declarative constraints "wherever possible", they are often "not powerful enough" to support all application logic (Cochrane et al., 1996, p. 4). Rails

[25] `https://guides.rubyonrails.org/active_record_validations.html`

[26] StackOverflow is the main forum for asking programming questions. `https://stackoverflow.com/questions/51345267/entering-phone-number-in-sql-losing-leading-zero`

[27] `https://guides.rubyonrails.org/active_record_validations.html#numericality`

[28] Rails refers to the column value as a string in keeping with its high-level language terminology, but the actual column value is of type *varchar* (SQL terminology).

allows for custom validations which make declarative constraints truly boundless.[29]

2.2.3 Additional Facilitative Software

Docker[30] is a containerization software used in this application. It is beneficial as it allows for developmental consistency across operating systems and applications. Furthermore, it facilitates locally hosting an application and reduces the likelihood of encountering issues particular to a given operation system. It would permit for scalability and consistency if the scope of the application were to change. The aforementioned Rails-using Github is a version control software. Version control allows for the rollback or reversal of any undesired changes and the use of such a system is paramount for quality assurance.

2.3 Evaluation Metrics

The application is ultimately a binary classifier. Such classifiers have two possible outputs: true/false, yes/no, red/blue, or, in this case, liaison/no liaison. The metrics by which we measure our classifier will relate to the confusion matrix:

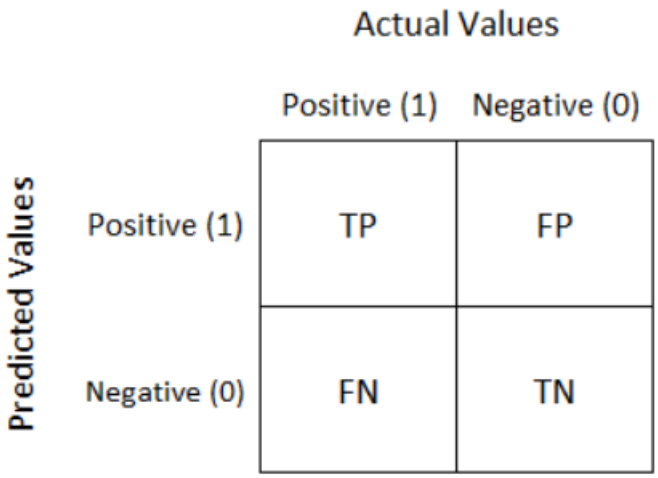

Figure 5: Typical confusion matrix.[31]

[29] https://guides.rubyonrails.org/active_record_validations.html#custom-methods

[30] https://www.docker.com/

[31] https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

The top-left corner represents cases in which the application successfully determines that a liaison has occurred, a.k.a. true positives (TP). The top-right corner represents cases in which the application incorrectly determines that a liaison has occurred, a.k.a. false positives (FP). The bottom-left corner represents cases in which the application incorrectly determines that a liaison did not occur, a.k.a. false negatives (FN). Finally, the bottom-right corner represents cases in which the application successfully determines that a liaison did not occur, a.k.a. true negatives (TN).

Such classification tasks can be considered with respect to many metrics. *Accuracy* measures the number of correct judgments as a proportion of the total number of judgments, i.e. (TP + TN) / all judgments. *Precision* is another important classification metric. It measures the number of correct liaison judgments as a proportion of the total predicted liaisons, i.e. TP / (TP + FP). This application will aim to be as precise as possible. It would be a disservice to L2 speakers to misguide them into realizing erroneous and/or forbidden liaisons.

3 Methodology

Marking liaisons involves the application of a series of phonological rules. The resulting classification is the core functionality of the application. However, the below user-flow shows that several steps must be carried out before this rule application can take place:

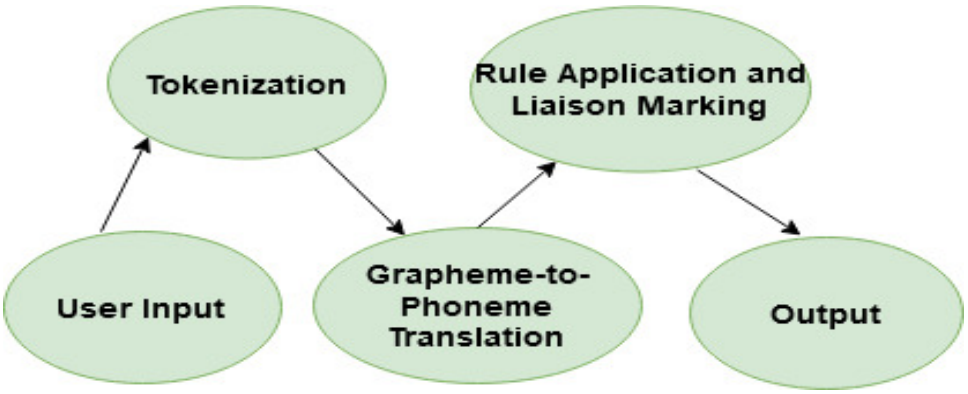

Figure 6: Application user-flow.

3.1 Tokenization

The application is presented with some user input – a chain of text. Parsing this text, known as tokenization, is the "initial phase" of many Natu-

ral Language Processing (NLP) applications (Webster & Kit, 1992). It is generally considered part of data pre-processing – i.e. it is something that must be done to access the 'real' data. In this application, the first step of text parsing consists of breaking the original chain of text down into a sequence of words. Initial considerations were to strip any punctuation as well as to separate any hyphenated words. However, a more rigorous tokenization process is necessary for any phonology-based NLP application as punctuation can be very rich in phonological information. Consider the following sentences:

6. *Mon frère, Luc et moi sommes allés au parc.*

   My brother, Luc,[32] and I went to the park.

7. *Mon frère Luc et moi sommes allés au parc.*

   My brother Luc and I went to the park.

Two issues arise from the indiscriminate punctuation removal between examples (6) and (7). Firstly, the application will ultimately output the user's original, punctuated, text with various 'liaison markings' where appropriate. If the application were to output the text sans comma, the meaning of the sentence would change. Furthermore, it may disregard important phonological information as outlined in the following section.

3.1.1 Punctuation

There are many punctuation markers in French and many of these convey important phonological information. A non-exhaustive list of French punctuation can be found under the punctuation entry in *L'Académie Française*.[33] Although some authors (Thomas, 1916) consider typography such as the apostrophe or the hyphen to be diacritics, this analysis will consider diacritics and punctuation, the label typically attributed to the previous two examples, to be distinct phenomena. Hadáček (2014) defines diacritical marks as "the accent marks used on some characters to denote a specific pronunciation". The diacritic mark only references one specific character and the information brought to the text is solely phonetic. As such, it is more appropriate to model diacritical information in the GP translation step. Contrastingly, punctuation can convey phonological information on a sequential or even sentential level. For example, a question mark at the

[32]English has its own semantic debate here with the Oxford comma, although there would be little ambiguity in this sentence regardless of its usage.

[33]The main council for the matters relating to French.

end of a sentence indicates that the sentence is interrogative and not affirmative. Phonologically, this difference is conveyed by rising intonation toward the end of the sequence:

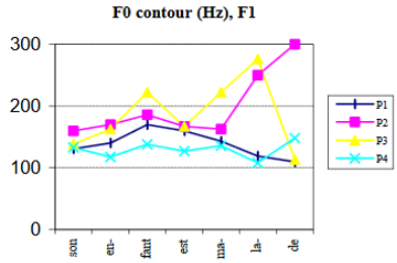

Fig.1. F0 contour (Hz) of F1, the first French speaker

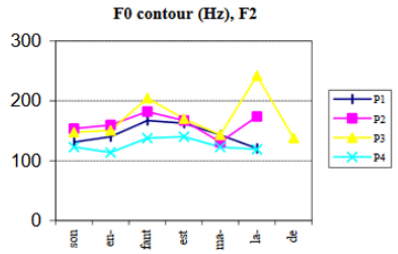

Fig.2. F0 contour (Hz) of F2, the second French speaker

Figure 7: P2 (pink) is the interrogative phrase (Dubost & Su, 1999, p. 1562).

Though the intonative function of the question-mark outlined above is not relevant to liaison marking, the question-mark and other sentential-ending punctuation will be considered as part of tokenization. Full-stops are indicative of the end of a sentence and are generally the first punctuation mark integrated by children (Wilde, 1986). Makkai (1980) defines the full-stop as a 'subsequent semantic pause'[34] and notes that, although optional, pauses are favoured in speech to represent a full-stop boundary.Indeed, a neural-network based phonological analysis conducted by Levy et al. (2012) found the full-stop to be followed by a pause in 96.5% of sentences within their dataset. A pause also generally occurs between the two vowels of a hiatus as the flow of CVCV speech is interrupted. However, while the liaison may serve to restore the CVCV flow and thus lead to the loss of the pause, this may not be the case for sentential-ending punctuation such as the full-stop. Consider the following example:

8. *Je pense donc je suis. Une citation Descartienne, elle ...*

   [ʒə.pɑ̃s.dɔ̃k.ʒə.sɥi||yn.ci.ta.sjɔ̃.de.kaʁ.ti.ɛn.ɛl]

   I think therefore I am. A Descartes quote, it ...

9. *Je pense donc je suis Une citation Descartienne, elle ...*

   [ʒə.pɑ̃s.dɔ̃k.ʒə.sɥi.**z**yn.ci.ta.sjɔ̃.de.kaʁ.ti.ɛn.ɛl]

   I think therefore I am A Descartes quote, she ...

[34]The onset of the entire chain of speech is considered the 'initial semantic pause'.

Although the first phrase is rather arbitrarily formed,[35] it supports the treatment of sentential-ending punctuation as a marker of non-liaison, an anti-rule of sorts. The liaison will never span across a full-stop. Thus, it is important that the tokenization process does not disregard these marks. The hyphen or *trait d'union* (dash of union) is a particularly important punctuation marker with respect to the liaison. It can occur in several different grammatical forms and is noted as one of the major difficulties of French orthography (Mathieu-Colas, 1995). Ultimately, the correct usage or incorrect non-usage is not important to this application. However, when it is used, it must be properly parsed. Consider the following examples:

10. *... une sèche-linge.*

    [yn.sɛʃ.lɛ̃ʒ]

    ... a dryer.

11. *Parle-t-il ?*

    [paʁl.til]

    Does he speak?

12. *Parlent-ils ?*

    [paʁl.til]

    Do they speak?

In (10), the hyphen is used to form a compound noun. Although several examples of words that can be written with-or-without-hyphen are provided by Mathieu-Colas (1995), these are generally words whose pronunciations would not differ as a result of hyphen usage, e.g. *photoroman* 'picture-book' [foto.ro.mɑ̃] versus *photo-roman* [fo.to.ro.mɑ̃]. In (10), there is the potential presence of an *e caduc*. The *e caduc* is a phonological feature of French due to which the schwa may or may not be pronounced in between consonants within a _VCCV_ sequence, e.g. *amener* 'to lead' [am(ə)ne], or at the end of a word, e.g. *sèche* 'dry' [sɛʃ(ə)]. If the hyphen were omitted, it could lead to a different phonetic make-up: *sèche-linge* [seʃ.lɛ̃ʒ] vs. *sèchelinge* [se.ʃə.lɛ̃ʒ]. Thus, proper phonological modelling would be to treat the two components of the compound nouns as separate sequences.

(11) and (12) are both examples of the French inverted interrogative phrase. This is one of the cases in which liaison realization is obligatory

[35]This is a unique poetic form in which the copula is not followed by an attribute of the subject. Furthermore, it would likely be surrounded by quotation marks, which might trigger a separate tokenization process regardless.

(Tseng, 2008, p. 2630). They are an excellent example of the difference between the grammatical obligatory liaison and the more independent facultative liaison. If we look at the phonetic transcription of (11) and (12) the 'liaising' [t] finds itself between the consonantal [l] and the vocalic [i]. These liaisons differ from the hiatus-preventing liaisons we have encountered hitherto. We can imagine that in (11), the *e caduc* present at word-end would have been pronounced in Old French and that it may have since undergone elision. The [t] which would have previously separated the schwa of *parle* and the [i] of *ils* was simply never dropped. Regardless, it is rather unnecessary for us to mark this liaison as the liaising consonant is present as its own morpheme within the text, in the form of *-t-*.

(12) also presents this seemingly futile liaison. Once more, the [t] separates the consonantal [l] from the vocalic [i]. We might again imagine that third-person plural (3PL) *parlent* was pronounced differently in Old French and had [t] as its final phoneme. Like its third-person singular counterpart *parle*, this liaising [t] was never dropped despite the fact that *parlent*, when considered in isolation, has the consonantal [l] as its final phoneme. As this 3PL present-tense case is less obvious 3PS present-tense counterpart, it is one that we may wish to mark.

Previously, we noted that for (10), "proper phonological modelling would be to treat the two components of the compound nouns as separate sequences". However, this does not mean that the hyphen acts as a pause-marker between sequences. Consider the following examples:

	Noun	**Translation**	**Transcription**	**1**	**2**	**3**
13.	*pot-au-feu*	pot-au-feu[36]	[pɔ.**t**o.fø]	pot [po]	au [o]	feu [fø]
14.	*arc-en-ciel*	rainbow	[aʁ.kɑ̃.sjɛl]	arc [aʁk]	en [ɑ̃]	ciel [sjɛl]

We can see that a liaison has occurred after *pot* in (13). Although it does not belong to any of the aforementioned obligatory liaison categories, it is realized to an almost obligatory level. Throughout the entirety of chef Julie Andrieu's television programme *« Le pot-au-feu – Les Carnets de Julie »*,[37] each occurrence is accompanied by a liaison. Nedecky's guide to French diction for singers lists *pot-au-feu* among a number of *familiar expressions* in which a liaison is realized (2011, p. 40). It might thus fall under the remit of this application to mark compound-noun liaisons such as *pot-au-feu*.

However, this poses numerous challenges. In (13) and (14), the equivalent sequences 1 and 2 are identical in their orthographical make-up. Sequence 1 ends in a consonant while sequence 2 begins with a vowel. How-

[36] A culinary dish that retains its name in English.

[37] `https://www.youtube.com/watch?v=8yo6XRTbV_s`

ever, *arc* is realized as [ark] whether isolated or as part of its compound noun, whereas *pot* is a word whose final consonant is latent, realized only when its union with the next word is strong. This is another example of the difference between *enchaînement* and liaison.

In order to determine that a liaison occurs only in (13), we need an identifying feature that can be encoded in the rule-base. However, no solutions are forthcoming. Both *t* and *c* are letters that can be silent in word-final position, e.g. *pot* and *escroc*. Thus, a rule based on the specific consonant itself would not be satisfactory. Alternatively, we would need the entire sequence's phonetic make-up to implement such a constraint. If all graphemes in *arc* were translated to phonemes, we would know that the *c*/[*k*] translation was preceded by the *r*/[ʁ] translation. If this were identified, we could encode the knowledge that because the phoneme preceding the word-final *c* is a consonant, its phoneme could not possibly be a liaising phoneme as there would not be a hiatus between the sequence *arc* and any succeeding sequence.

Like the *t* found at word-final position in pot, word-final *c* must be preceded by a vowel in order to have two surface representations, [k] and []. As noted in Section 2.2, however, the entire sequence will not undergo GP translation. Thus, it would seem that the application cannot model compound-noun cases. This is not a major disappointment. The very nature of hyphenated words means that readers are drawn to the link between individual components. Häikiö et al. (2011) found Finnish children to spend more time reading hyphenated compound nouns than concatenated compound nouns. Similarly, we might expect an L2 French speaker to consider the compound noun more carefully than they might the syllabic gap between two non-compounded constituents. Moreover, representing the hiatus in our output would also be rather unpleasant. It would be necessary to implement rather specific CSS[38] code to allow the underbar[39] to link the *t* of *pot* to the *a* of *au* in *pot-au-feu*, all the while ensuring it navigates under the hyphen itself. Considering the minimal benefit it would provide to application users, this surmountable but unnecessary challenge need not be taken on. With respect to the findings of Häikiö et al. (2011), marking liaisons in this context may even be of detriment as the presence of more typography near to the hyphen of the compound noun might serve to further slow the reader.

[38]Cascading Style Sheets – the language through which most web-app design is done.

[39]See `https://www.internationalphoneticassociation.org/sites/default/files/IPA_Kiel_2015.pdf` under *Suprasegmentals* > *Linking*.

3.1.2 Numbers

Numbers are rather unique with respect to tokenization and, by extension, GP translation. A number can have two representations: numerical (9) or textual (*neuf*). We need to tokenize the former in order to obtain the latter as it is this textual equivalent that GP translation requires.

There are a few caveats that must be considered before this translation is carried out. Unlike words, numbers must be considered in their entirety before any GP translation can be carried out. *34* is not a sequence containing two morphemes – *3* and *4* – that can then be textually translated to *trois* and *quatre* respectively. Rather, the entire sequence *34* must be translated to *trente-quatre*. This is necessary for the correctness of GP translation.

The textual representations of numbers will need to undergo further tokenization to handle punctuation – specifically the hyphen. Many double-digit numbers are textually represented as compound nouns, e.g. *trente-et-un* 'thirty-and-one -> thirty-one' [tʁ.ɑ̃.te.œ̃]. However, although the [t] at the end of *trente* [tʁ.ɑ̃t] does move forward to the coda position of the following syllable, this is not a liaison. Rather, this is a case of enchaînement. It is not a liaison as *trente* has only one surface representation.

There are numbers, however, that have multiple surface representations. Each of *un* 'one' [œ̃], *deux* 'two' [dø], and *trois* 'three' [twa] ends in a vowel and thus may find itself in a hiatus. *Deux* and *trois* see their respective *x* and *s* move from one surface representation, [], to another, [z], if the following sequence begins with a vowel. *Un* is slightly more challenging as translating its numeral representation *1* can depend on its context. It is the French indefinite article equivalent to *a(n)* in English. As French is a gendered language, the succeeding noun can be either masculine or feminine. If the former, 1 is textually represented as *un* [œ̃]. If the latter, 1 is textually represented as *une* [yn]. If the following noun is masculine, *un* will liaise, e.g. *un éléphant* 'an elephant' [œ.**n**e.le.fɑ̃]. However, as *une* ends in a consonontal phoneme, it does not liaise.

Evidently, determining the gender of the succeeding noun is outside of the phonological domain in which this application is based. Thus, if the determiner were to be represented by the numeral 1 rather than *un*/*une*, it is not realistic to determine whether it is a case of liaison (un) or *enchaînement* (une). However, the use of 1 in place of the determiner is very rare and appears only in text-speak.[40] For such a unique case, it would be incredibly costly to determine the gender by some sort of dictionary look-up for the succeeding noun.

[40]Think *CU L8R*, *lol*, et cetera.

There is an open-source Ruby library, *humanize*,[41] that supports French number-to-word translation. However, since the list of numbers we are interested in is so short, it is simpler to implement a trivial check to see if the actual digit itself (1, 3, 20, etc.) is among those that can lead to a liaison. Ultimately, numerical liaisons will represent a very small proportion of the liaisons that we will wish to mark and so excessive time need not be spent on dealing with these cases. Numerical representations in the February 13th front-page article on the website of *LeMonde*, « *Syrie : empêcher un bain de sang à Idlib* », represented just 13 of 644 words (2.01%). When we consider that few numbers have multiple surface representations, it is more pragmatic to deal with the few basic and easily implementable cases than it is to implement full numerical tokenization.

3.1.3 Proper Nouns

Proper nouns are listed among the *forbidden* liaisons in Delattre's tableau (1947). Given the fluidity of graphemes in proper nouns,[42] this presents one fewer challenge with respect to GP translation. However, they present their own challenge. If we do not determine that a given sequence is a proper noun, it will undergo GP translation much like any other sequence. As the correct GP translations may not be in the GP dictionary, there is thus a risk that the GP translation will be incorrect. Furthermore, even if the proper noun were correctly GP translated, there is a risk that the translation leads to a phonological context that triggers one of our rules. Consider the following example:

15. *Le train/italien est arrivé.*

    [lə.tʁɛ.**n**i.ta.ljɛ̃.ɛ.ta.ʁi.ve]

    The Italian train arrived.

16. *Lyon est loin.*

    [li.ɔ̃.ɛ.lwɛ̃]

    Lyon is far away.

(16) is taken directly from Delattre's *forbidden* liaison examples (Delattre, 1947, p. 153). Thus, it should not be marked in any case. However, *Lyon est* is virtually identical to *train est* in its phonological context. *Lyon* and *train* both end in a nasal vowel while *italien* and *est* begin with a vowel. In

[41] `https://github.com/radar/humanize`

[42] Consider Kiera, Ciara, Ciaradh [kɪə.rə] or Cami, Camille [ka.mij].

(15), the liaison leads to the nasalized vowel [ɛ̃] breaking down into the respective vowel and nasal consonant [ɛn]. The phonological context is identical in (16), but the breakdown of the nasal vowel must be prevented as a liaison must not follow a proper noun per Delattre.

Proper nouns do have a typological property that may be of use in tokenization. Unlike normal nouns, proper nouns must be capitalized irrespective of their sentential position. Normal nouns must only be capitalized when they are the first word in a sentence. However, this is not phonological information. Nor is it information that the application would automatically have available. Consider the following example:

17. *X Y Z. Mon/aîné ...*

    [mɔ.**n**ɛ.ne]

    X Y Z. My eldest ...

18. *X Y Z. Lyon est ...*

    [li.ɔ̃.ɛ.lwɛ̃]

    X Y Z. Lyon is ...

Let X Y Z be three arbitrary sequences that make up a sentence. *Mon* [mɔ̃] and *Lyon* [li.ɔ̃] then begin a new sentence. Both sequences begin with an uppercase letter and both end in the same nasal phoneme [ɔ̃]. Both sequences are succeeded by a sequence whose first phoneme is [ɛ]. However, a liaison is present in (17) but not in (18). The aforementioned capitalization feature of proper nouns is not a distinguishing factor here. Thus, we must devise some other procedure by which these two cases can be distinguished. Fortunately, there are only few cases such as (17) that mirror (18). French is a Subject-Verb-Object language and its basic sentence format is Noun Phrase-Verb Phrase. Thus, the first (capitalized) word of the sentence will likely be part of this Noun Phrase. It may be a proper noun *Jean* [ʒɑ̃], a determiner *un* 'a' [œ̃], or a pronoun *il* 'he' [il]. We have already determined that the first cannot result in a liaison. Only one of the nine pronouns in French[43] ends in a nasal vowel and thus it is the only pronoun that can trigger this nasal breakdown.
We are left with the determiner, the grammatical category to which the *mon* in (17) belongs. Unlike proper nouns, this is a rather finite and fixed category determined in its entirety by grammarians such as *L'Académie Française*. Thus, there are fewer members. Fewer yet end in a nasal phoneme.

[43] *Je* 'I' [ʒə], *tu* 'you' [ty], *il* 'he' [il], *elle* 'she' [ɛl], *on* 'one/we' [ɔ̃], *nous* 'we' [nu], *vous* 'you (pl.)' [vu], *ils* 'they (m.)' [il], and *elles* 'they (f.)' [ɛl]

Mon is such a determiner. In order to distinguish cases (15) and (16), the check needs to be two-fold. If it is capitalized and a member of this finite group, it may result in a liaison. If it is capitalized and not a member of this finite group, then it is a proper noun and thus liaison is forbidden.

3.1.4 Initialisms

An initialism is an abbreviated and fully capitalized sequence that represents multiple words. All acronyms are initialisms, but not initialisms are acronyms.[44] A separate GP translation methodology for initialisms could be implemented quite easily by stocking the letters of the alphabet and their phonetic representations in a separate dictionary. As of yet, there have been no rules proposed with respect to initialisms and the liaison.

Initial field research would indicate that liaisons are facultative, if not forbidden. In the M&Ms advertisement « *Les M&Ms Crispy sont là !* », the voice actor does not liaise *les* with *M* in any of the four instances, ergo the hiatus is not avoided. *Les* is a determiner and liaising following a determiner is obligatory (Delattre, 1947, p. 153).

The fact that this liaison is not realized is perhaps indicative that the liaison is moreso a historical phenomenon than a phonological phenomenon. We stated earlier that latent liaising consonants had a "union with the following vowel-initial word" that was strong enough to be preserved (Delattre, 1947, p. 148). M&Ms were first produced in 1941 and first introduced to Europe during the 1980s. The initialism never co-existed with a version of French in which the [z] of *les* was not latent. As such, it is not liaised, despite the resulting hiatus.

3.2 Grapheme-to-Phoneme Translation

3.2.1 Verbs

Present Tense Subject-Verb-Object has been the "established word order" of basic sentences since Old French (Bauer, 1995, p. 110). Although certain exclamative phrases that do not contain a verb are provided by Vinet (1991), these phrases are few and far between. As a general rule, a sentence cannot exist without a verb. French verbs can be split into three groups (Trager, 1944, p. 135) with respect to their infinitive and the following tableau shows typical present-tense conjugations:

[44] An acronym must be pronounceable as a word, e.g. NASA or *SMIC*. An initialism is pronounced one letter at a time, e.g. SNP, DUP, *UE*.

Forme	Groupe 1	Groupe 2	Groupe 3 (exemples)	
Lexème	LAVER	FINIR	SORTIR	BOIRE
Présent 1SG	*lave* /lav/	*finis* /fini/	*sors* /sɔr/	*bois* /bwa/
Présent 2SG	*laves* /lav/	*finis* /fini/	*sors* /sɔr/	*bois* /bwa/
Présent 3SG	*lave* /lav/	*finit* /fini/	*sort* /sɔr/	*boit* /bwa/
Présent 1PL	*lavons* /lavɔ̃/	*finissons* /finisɔ̃/	*sortons* /sɔrtɔ̃/	*buvons* /byvɔ̃/
Présent 2PL	*lavez* /lave/	*finissez* /finise/	*sortez* /sɔrte/	*buvez* /byve/
Présent 3PL	*lavent* /lav/	*finissent* /finis/	*sortent* /sɔrt/	*boivent* /bwav/
Infinitif	*laver* /lave/	*finir* /finir/	*sortir* /sɔrtir/	*boire* /bwar/

Figure 8: Sample present-tense conjugations for three main verbs groups (Bonami, Boyé, Giraudo, & Voga, 2008).

Many verbs in the third group (G3) are morphologically irregular – note that *sortir* does not belong to Group 2 (G2), the so-called '-ir verbs'. This presents a challenge as many of French's most common verbs belong to G3. A cross-corpora study conducted by Greidanus (2014) included the following G3 verbs:[45]

Verb	Translation	Frequency Rank
faire	to do	3
dire	to say	4
tenir	to hold	27
venir	to come	13
savoir	to know	7
voir	to see	7

Table 3: Frequency of selected G3 verbs.

Although Bonami et al. (2008) note that 90% of French verbs belong to the first group (G1), the frequency of these irregular verbs mean that it is imperative that we can account for this group. Furthermore, *être* 'to be' and *avoir* 'to have', of frequency rank 1 and 2 respectively, are the two French auxiliary verbs and are used in any composite verbal tense.
Given their uniformity, G1 and G2 are the logical starting point for GP translation. Similarly, it is logical to consider the present-tense first. Though we saw the endings in Figure 8, let us further break down that analysis to consider the verbs in terms more relevant to liaison marking:

[45]Frequency rank represents the rounded mean for the up-to-five corpora.

Form	Conjugation	Stem		Ending		Transcription
1SG	parle	parl	[paʁl]	e	[]	[paʁl]
2SG	parles	parl	[paʁl]	es	[]	[paʁl]
3SG	parle	parl	[paʁl]	e	[]	[paʁl]
1PL	parlons	parl	[paʁl]	ons	[õ]	[paʁlõ]
2PL	parlez	parl	[paʁl]	ez	[e]	[paʁle]
3PL	parlent	parl	[paʁl]	ent	[]	[paʁl]

Table 4: conjugation of G1 verb parler 'to speak'.

Note that four of the verbal endings, i.e. graphemes, are purely orthographic and translate to the empty phoneme []. This is because of the *e caduc*. Although Novakova's (2012) study on the renowned French opera « *Carmen* » found 95% of potential schwas to be realized, including the 1SG *parle* [paʁ.lə] that was denoted above as [paʁl], singing, and particularly opera, are not always reflective of normal speech. The realization of word-final schwas is a well-known characteristic of southern French speakers (Ranson & Passarello, 2012, p. 1519). This is generally not the case in standard Parisian French, although some linguists claim that it is making a return (Fagyal, 2000). Many educational resources such as Le Goffic's *Conjugated forms of the French verb: oral and written* (1997) transcribe G1 verbs in the same manner as in Table 4. We can quite trivially account for 1SG and 3SG in GP translation. The e will simply be absorbed by a grapheme representing the previous consonantal phoneme. 1PL and 2PL have final graphemes to which only one phoneme can be attributed. The remaining two cases, 2SG and 3PL, are slightly more particular. As the particularity regarding 3PL GP translation applies to verbs in the other two groups, it will be discussed at the end of this section.

The *-les* of *tu parles* that translates to [l] per Table 4 translates to [le] in the word *les* 'the (pl.)'. Similar issues will arise for words such as *mes*, *des* or *ces*. The application could account for this if Part-of-Speech (POS) tagging were implemented. However, as this is not currently the case, the GP translation will translate 2SG verbs using the 'fallback' s/[s] translation. Though this is not the correct translation, it is not of particular concern. The largest sub-class of G1 verbs are those like the above *parler* whose stem ends in a consonant (Trager, 1944, p. 136). We will respect the vowel-consonant (VC) characteristics of these verbs' final phonemes if we translate to the consonantal [s].

Fortunately, G2 verbs pose fewer challenges:

Form	Conjugation	Stem		Ending		Transcription
1SG	finis	fin	[fɛ̃]	is	[i]	[fi.ni]
2SG	finis	fin	[fɛ̃]	is	[i]	[fi.ni]
3SG	finit	fin	[fɛ̃]	it	[i]	[fi.ni]
1PL	finissons	fin	[fɛ̃]	issons	[i.sɔ̃]	[fi.ni.sɔ̃]
2PL	finissez	fin	[fɛ̃]	issez	[i.se]	[fi.ni.se]
3PL	finissent	fin	[fɛ̃]	issent	[is]	[fi.nis]

Table 5: conjugation of G2 verb finir 'to finish'

Each SG ending, *-is* and *-it*, translates to [i]. The final graphemes and therefore phonemes of 1PL and 2PL endings mirror G1 verbs: *-ons* [ɔ̃], *-ez* [e]. Although *ent* only forms part of the 3PL morpheme *-issent*, it does resemble the 3PL of G1 in that the *ent* appears to translate to [], with only *iss* contributing phonetically in this case.

G3 acts as a catch-all of sorts for verbs whose behaviour does not correspond to that of G1 or G2 verbs. The frequency of these verbs and their irregularities mean that GP translation is both important and difficult to get right. Verbal liaison is, however, facultative and it would be onerous to account for each individual irregular verb. We will simply have to rely on the integrity of our GP dictionary and this will be something we may wish to consider when evaluating the effectiveness of our GP translation.

3PL Present Tense endings are uniform for groups 1 and 2 and for most of group 3. The only verbs that do not conform are those whose stem is irregular, such as *avoir* 'to have', *être* 'to be', and *aller* 'to go'. These will be accounted for separately as they are high-frequency, high-liaison verbs – we saw in Table 2 that 3PL *sont* of *être* was liaised in 86% of instances. Verbs with the typical 3PL ending of *-ent*/[] do not liaise to such an extent. Most verbs in Group 1 have a consonant-final stem (Trager, 1944, p. 136). As such, a G1 verb conjugated in the 3PL cannot find itself in a hiatus with a subsequent vowel-initial word.

In example (12), we saw a 3PL verb in its inverted form in which liaison is obligatory. Despite the absence of a hiatus, the [t] of its *-ent* ending was realized as a liaising consonant. We discussed the historical reasons behind this peculiarity. We ultimately decided not to mark any inverted verb liaisons as the hyphen already fills this role. However, 3SG verbs in their regular form can also result in this non-hiatus-preventing liaison:

19. *Ses parents vivent/à Dubaï.*

    [se.pa.ʁɑ̃.viv.**t**a.Du.ba.i]

Her parents live in Dubai.

Unlike the inverted verb form, the hyphen is not present in (19) to fill the role of liaison marker. Although the liaison in (19) is not obligatory, it is a relatively common liaison in a register of higher elocution. The above example was taken from a documentary[46] and other examples found include news reports, comedians, and sports commentators. Nonetheless, it is a context that we may wish to mark.

There are some 3PL present tense verbs that exhibit rates of liaison comparable to obligatory liaisons. These are typically auxiliary verbs, particularly *pouvoir* 'to be able to' and *devoir* 'to be obligated to'. These liaisons seem highly linked to the following lexeme. *Pouvoir* liaises with the infinitive *être* in 81.2% of instances (N=143) (Bybee, 2005, p. 33). *Devoir* liaises with *être* in 94.5% of instances (N=91) (idem). *Vouloir* 'to want' liaises with *être* in 83.3% of instances, although the sample size is low (N=6) (idem). The rate of liaison realization for these three verbs drops to between 45% and 56% of instances before *avoir* and to between 21% and 54% before other infinitives (idem). These rates are for any conjugation of the verb, but the same author notes that the auxiliary in this liaisons will "ordinarily be in the third person" (Bybee, 2001, p. 17). Furthermore, *peut*, the 3PS of *pouvoir*, is unlikely to liaise with *être* because of the fixed form *peut-être* 'maybe'. Thus, we can conclude that 3PL forms of *devoir* and *pouvoir* will liaise facultatively, but to a high degree, before *être* and facultatively, to a lesser degree, before other infinitives. We do not have that infinitive information available until such point that POS tagging has been implemented. As such, we will consider just *être* and *avoir*. Furthermore, we will include the 3PL of *vouloir* despite the low sample size.

Other Tenses French has fourteen verb tenses in total.[47] Though the present tense is the most commonly used, a further thirteen is quite the daunting task with respect to GP translation. Fortunately, however, most of the tenses are far more regular than the present tense, and endings are uniform across all three groups. Furthermore, several of the tenses are incredibly low frequency. The *passé simple* 'simple past' has long since been reserved for narration, while only one of the four subjunctive tenses is used regularly. Additionally, several more are composed tenses and so GP translation remains common:

[46] `https://www.youtube.com/watch?v=OldzOC2LajY&feature=youtu.be&t=15` at approximately 00:17.

[47] See conjugation for *avoir*: `https://www.larousse.fr/conjugaison/francais/avoir/749`

	Imperfect		Conditional		Future	
1SG	-ais	[ɛ]	-ais	[ɛ]	-ai	[ɛ]
2SG	-ais	[ɛ]	-ais	[ɛ]	-as	[a]
3SG	-ait	[ɛ]	-ait	[ɛ]	-a	[a]
1PL	-ons	[ɔ̃]	-ons	[ɔ̃]	-ons	[ɔ̃]
2PL	-ez	[e]	-ez	[e]	-ez	[e]
3PL	-aient	[ɛ]	-aient	[ɛ]	-ont	[ɔ̃]

Table 6: French verb endings in the imperfect, conditional, and future tenses.

The full conjugation for 1PL imperfect/conditional is *-ions* but we are only interested in the final grapheme *-ons*. This simplifies GP translation as the 1PL and 2PL endings are the same as they were for the present tense. There are clearly identifiable patterns for the remaining persons. These three tenses are also the base of several composed tenses. The *plus-que-parfait* 'pluperfect' is composed of an auxiliary verb, either *avoir* or *être*, conjugated in the imperfect, and the past participle of another verb. The past participle in itself is quite regular. The past participle of 90% of French verbs, those in G1, ends in *-é*/[e]. Although the other groups are less regular, many of their past participles are simply accounted for by existing graphemes in our dictionary. Incidentally, the three future singular endings are also identical to the passé simple G1 verb endings for those persons and so GP translation will also account for these forms.

3.2.2 Word-final Consonants

Quémart and Casalis consider that GP translation rules are "highly inconsistent in French" and affirm that "one recurrent manifestation of this inconsistency is the presence of silent letters at the end of words" (2017, p. 85). This is particularly relevant to the liaison, whose origin lies in a "series of processes which reduced syllable codas in Old French" (Morin, 1986, p. 167). We encountered issues in GP translating word-final consonants when considering the tokenization of compound hyphenated nouns.[48] In that case, we were content to acknowledge that marking liaisons within a compound noun was not a priority. However, the concerns are not limited to compound nouns and many consonants can be voiced or silent at word-end, depending on the phonological context. We previously proposed that GP translating the entire sequence would help differentiate the *enchaînement* of *arc-en-ciel* from the liaison of *pot-au-feu*. This was an im-

[48] See examples (13) and (14).

portant consideration as it may be detrimental to performance given that the number of GP translation would greatly increase. It would also dismiss the notion proposed herewithin that solely the first and final phoneme of each sequence is necessary for liaison marking. However, even if the application were to GP translate sequences in their entirety, it still may not be satisfactory for GP translating word-final silent consonants. Consider the following examples:

Sequence	**Translation**	**Transcription**
20. *arc*	bow	[aʁk]
21. *escroc*	crook	[ɛ.skʁo]
22. *art*	art	[aʁ]
23. *complot*	scheme	[kɔm.plo]
24. *brut*	crude	[bʁyt]

We can determine *c* to be silent, i.e. a non-contributive grapheme, when it follows a vocalic phoneme (21) and to be voiced when it follows a consonantal phoneme (20). However, *t* is silent when following a consonantal phoneme (22). Moreover, *t* can be both silent (23) or voiced (24) when following a vocalic phoneme. Even if each sequence were GP translated in its entirety, these inconsistencies would remain. Thus, it is not worth the performance detriment and it will remain the case that the application GP translates only the first and final phoneme of each sequence.

We must then decide how to GP translate these potentially silent word-final consonants. Although a certain number of potential liaisons may be missed, the most consistent manner is to treat them as if they are voiced. This reduces the number of false positives in this context to zero. As a liaison can only occur when the phonemes on either side of the sequence boundary are vowels, no liaison will ever be marked following a sequence whose final phoneme is a consonant.

3.2.3 H-initial Words

H-initial words pose a unique challenge with respect to GP translation. Like many Romantic languages, French does not have [h] in its phonemic inventory (Jakobson & Lotz, 1949, p. 151). Like fellow Latin descendant Spanish, generally considered to have a much more "ortographically transparent system" than French (Serrano & Defior, 2008, p. 82), French GP translation of h-initial words is transparent when considered in isolation. The *h* is simply mapped to an empty phoneme or, as below, a non-contributing element of the grapheme:

Sequence	Transcription	Translation	G-P	G-P	G-P
25. *homo*	[o.mo]	homo	ho-[o]	m-[m]	o-[o]

However, unlike Spanish, which remains consistent in its transparency, h-initial words pose GP dilemmas when considered as part of an overall sequence in French. This is largely due to the liaison. While all h-initial words are treated equally in Spanish, there exist two categories of h-initial words in French: those that begin with a *h muet* 'silent h' and those that begin with a *h aspiré* 'aspirated h'. The latter is "one of the classical problems of French phonology" (Gabriel & Meisenburg, 2009, p. 163). The divide is quite simple and lies in the etymology of a given h-initial word. Those with Latin origins are silent, while those with other origins are aspirated. Gabriel notes that many are of Germanic origin, and Einhorn (1974, p. 3) notes that [h] is sounded in "words of Germanic origin" but aspiration is not limited to those of Germanic origin:

Word	Translation	Origin
26. *héros*	hero	Greek[49]
27. *hausse*	rise	Hebrew[50]

Opinion is divided on the status of [h] within French. We stated above that [h] is not in French's phonemic inventory. Einhorn (1974) and Tranel (1987), however, include it in their analyses. This analysis will not consider [h] to be a French phoneme, but it is worth noting that the *h* of any h-initial word does GP translate to something that would appear to have the status of a phoneme. Scheer segments aspirated *hêtre* 'beech tree' with a non-null onset:

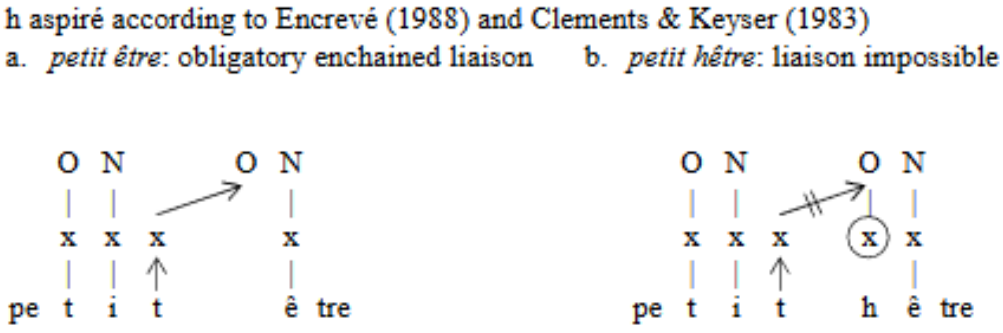

Figure 9: syllabic breakdown of h aspiré (Scheer, Wauquie, & Encrevé, 2015).

[49] https://www.cnrtl.fr/etymologie/h%C3%A9ros

[50] https://www.cnrtl.fr/etymologie/hausse

Realizing a liaison in (b) is impossible, i.e. "forbidden" (Sturm, 2012, p. 159). As such, we could treat *h aspiré* words "as if they began with a consonant" (Selkirk & Vergnaud, 1973, p. 251), thereby ensuring they will not liaise. We could choose either [h] or some arbitrary pseudo-phoneme as its GP translation. However, not only are liaisons allowed following *h muet* words, they can in fact be obligatory (Thitothati, 2013, p. 71).

Although a challenge, the problem of dealing with both *h muet* words and *h aspiré* words is easily resolved. If the word is of Latin origin, we may simply 'ignore' the silent *h* and GP translate the following vowel. If the word is of non-Latin origin, the aspirated *h* may be GP translated accordingly as [h] or as the glottal stop [?]. However, while trivial to solve, it is not trivial to implement. This would require the application to keep a dictionary or to reference a dictionary each time a h-initial word is received in order to determine its etymology. This is simply not feasible. Though CTRL[51] serves as an accurate and extensive etymological dictionary, it has no API. Furthermore, making requests to external APIs is resource intensive and as the information would be necessary prior to GP translation, the request could not be made asynchronously.[52]

If we are therefore unable to distinguish the two cases, one single behaviour must be chosen. Neither option is ideal. If we treat all h-initial words as *h muet*, the first phoneme for the word will always be a vowel and this would almost certainly lead to marking liaisons that are, in fact, forbidden. If we treat all h-initial words as *h aspiré*, the first phoneme for the word will never be a vowel and therefore no liaison will ever be marked. As shown above, however, these liaisons can be obligatory. It is therefore incorrect to not mark a liaison as having occurred.

French is a Romantic language and most of its vocabulary stems from Latin. Therefore, it would not be illogical to assume that most h-initial words also stem from Latin. As such, it is more likely that a given h-initial word is of Latin origin and thus that a liaison may take place. Treating all h-initial words as *h muet* would produce more true positives than false positives. Treating all h-initial words as *h aspiré* would produce more false negatives than true negatives. However, while the former may perform better in the majority of cases, we must not forget that the ultimate goal of this application is to help L2 French speakers improve. While not marking obligatory liaisons may not help them improve, marking forbidden liaisons would actively lead them astray. It is for this reason, failing some process

[51] `https://www.ctrl.fr/`

[52] Asynchronous requests may be made in a different order than they are received to maximize performance.

by which the application could determine the etymology of a h-initial word, that all h-initial words will be treated as *h aspiré*.

3.2.4 Grapheme-to-Phoneme Dictionary

The previous sections analyzed several phenomena that needed to be considered before encoding the application GP dictionary. Most analyses propose that French has between 35 and 37 phonemes (Jakobson & Lotz, 1949; Pierret, 1994). Many consider the [a]/[ɑ] opposition to have largely disappeared, with the former representing 97.6% of occurrences (Léon, 1992, p. 87). Ultimately, unless one of our rules involves a condition based on anteriority or posteriority, it does not matter which is chosen. As such, the more common [a] will be used. More generally, the precision of the GP translation is not pivotal. It does not need to be perfect, it simply needs to produce a phoneme that is of the same category as the actual GP translation. This application makes judgments based on two categories: vowel/consonant (VC) and nasality (N). The judgment made during rule application will be the same provided that these categories are respected.

The following 36 phonemes make up the application dictionary: a, ɑ̃, b, d, e, ə, ɛ, ɛ̃, f, g, ɥ, i, j, k, l, m, n, ŋ, ɲ, o, ɔ, ɔ̃, ø, œ, œ̃, p, ʁ, s, ʃ, t, u, v, w, y, z, ʒ. Furthermore, the pseudo-phoneme [h] will be used until such time that *h aspiré* words can be distinguished from *h muet* words. Note that the [h] pseudo-phoneme we use here is not actually the phoneme [h] as it would be pronounced in English. Rather, it is a placeholder for the glottal aspiration phenomenon. Others use the apostrophe ['] to denote this, but we do not want application behaviour concerned with punctuation to cause an issue as a result of using an apostrophe.

	Regex pattern	
1. [a]	a[hst]?[53]	**a**mont [**a**.mɔ̃]
		ir**a** [i.ʁ**a**]
		c**as** [k**a**]
		b**ah** [b**a**]
		r**at** [ʁ**a**]
	[à|â]	(l)**â** [(l)**a**]
		âge [**a**ʒ]
	ha(b|llu|rm)	**hab**it [**a**.bi]
		hallucinations [**a**.lu.si.na.sjɔ̃]

[53]The question mark means one or fewer occurrence of a given symbol.

2. [ɑ̃]	en(d(s)?|ts)?	**en**terrer [**ɑ̃**.tɛ.ʁe]
		v**end(s)** [v**ɑ̃**]
	an[dst](s)?	gr**and(s)** [gʁ**ɑ̃**]
	[ae]mp(s)?	c**amp(s)** [k**ɑ̃**]
	[em]ment(s)?	abonn**ement(s)** [a.bɔn.m**ɑ̃**]

The first pattern for [ɑ̃] is not perfect. Many words end in an *en* that translates to [ɛ̃] such as *bien* 'well' [bjɛ̃]. Some of these will be correctly translated by patterns for [ɛ̃]. However, some will not, such as *ben* 'eh' [bɛ̃] or *coréen* 'Korean' [ko.ʁe.ɛ̃]. This is not an issue as both VC and nasality characteristics are respected.

The final pattern accounts for all nouns ending in *ment(s)* and all adverbs ending in *(m)ment*. We discussed the issue of translating 3PL present-tense verbs in Section 3.2.1. Verbs such as *semer* 'to sew', whose stem ends in *m*, will end in *ment* when conjugated in the third-person plural. However, these verbs have a unique typological property. The stem *sem* becomes *sèm* in all first persons and the third person plural. This occurs when the stem is consonant-final and serves to emphasize that the *e* therein transforms from a weak schwa vowel to a strong [ɛ] (Blanche-Benveniste, 2002, p. 9). Thus, we can be sure that the *ment* pattern will not mistranslate any verbs conjugated in the 3PL of the present tense.

3. [b]	b(e)?(s)?	**b**ande [**b**ɑ̃d]
		au**be(s)** [o**b**]
4. [d]	d(e)?	**d**on [**d**ɔ̃]
		week-en**d** [wi.kɛ**d**]
		fêtar**de** [fɛ.taʁ**d**]
	[a-hj-np-z]des	f**ades** [fa**d**]

Week-end is a rather unrepresentative example. A loanword from English, the [d] at word-end was retained as part of the anglicism. However, 'd' is generally silent at word-end, typically after a nasal or rhotic phoneme. It will be in these phonemes' grapheme dictionary that these d-final words will be considered. Thus, *week-end* aside, if we map [d] to a 'd' at word-end, it is likely that one of these nasal or rhotic phonemes is lacking a grapheme in its inventory.

The final pattern accounts for *des* at word-end. We exclude *i* and *o* from possible preceding letters as both *ides* and *odes* are valid sequences which would be mistakenly mapped as monophonemic.

5. [e]	é(e)?(s)?	**é**trange [**e**.tʀɑ̃ʒ]
		abîm**ées** [a.bi.m**e**]
	hé	**hé**siter [**e**.zi.te]
	ë	canoë [ka.no.**e**]
	e[rstyz](s)?	jou**er**, jou**et(s)**, jou**ez** [ʒu.**e**]

The final pattern would mistranslate the adjective *fier* 'proud' [fjɛʁ]. This is an exception that could not even be accounted for by POS tagging as other adjectives such as *guerrier* map *-er* to [e]. We may need to consider this exception if we devise a rule based on a context in which the adjective might find itself, but it is not a pressing concern.

6. [ɛ]	ê	**ê**tes [**ɛ**t]
	ai(en)?[stx]?	parl**ai**, parl**ais**, parl**aient** [paʁ.l**ɛ**]
	ès|est	tr**ès** [tr**ɛ**]
	he([bcx]|pt|r[bm])	**hec**tare [**ɛ**k.taʁ]
	ay(s)?	tramw**ay(s)** [tʁam.w**ɛ**]
		ayant [**ɛ**.jɑ̃]

As discussed earlier, the schwa mostly undergoes elision at word-end due to the e caduc. Most words that begin with 'e' in French have [ɛ] as their first phoneme, e.g. *estime* 'esteem' [ɛ.stim]. However, VC characteristics are respected in either case, so it is not important that the phoneme be exact. The e/[ə] pattern in [8] is used as a fallback translation.

7. [ɛ̃]	(ain|ien)(s)?	proch**ain(s)** [pʁɔ.ʃ**ɛ̃**]
	ym	th**ym** [t**ɛ̃**]
	aim	f**aim** [f**ɛ̃**]
	(e)?in(s|t)?	p**in**, p**eins**, p**eint** [p**ɛ̃**]
	oin[gt]?	c**oin**, c**oing** [kw**ɛ̃**]
	hind	**hind**ou [**ɛ̃**.du]
8. [ə]	e	qu**e** [k**ə**]
9. [f]	f(e)?(s)?	**f**ou [**f**u]
		cara**fe(s)** [ka.ʁa**f**]
	ph	**ph**onologie [**f**ɔ.nɔ.lo.ʒi]
10. [g]	g	**g**are [**g**aʁ]
		blo**g** [blɔ**g**]
	xy	**xy**lophone [**g**zi.lo.fɔn]

11. [h]	h(on[dgt])?	**h**anche [**h**ɑ̃ʃ]
		hongrois [**h**ɔ̃.gʁwa]
	hér(o|is)	**hér**o [**h**e.ʁo]
12. [ɥ]	hu[iî]	**hui**t [**ɥ**it]
		huîtres [**ɥ**itʁ]

We have discussed the logic behind pseudo-phoneme [h] already. [ɥ] can only occur as part of a diphthong (Coşciug, 2014, p. 25) – [ɥi] in this case. However, we will translate it as [ɥ] as there are no fewer than 32 diphthong combinations in French (idem). VC characteristics are respected in either case. Additionally, we map 'hong' or 'hond' to [h] in order to mark as *h aspiré* words such as *hongrois* or *hondurien*. This will allow us to then mark words that begin with *hon* but not *hond/hong* as h muet.

13. [i]	(i([isx]|t(s)?)?|[îïy]|ie(s)?)	**is**oler [**i**.zo.le]
		sk**i(s)** [sk**i**]
	hy([bdgm]|per|p(n)?o|st)	**hypno**tisé [**i**p.nɔ.ti.ze]
		hystérie [**i**.stɛ.ʁi]
	hi(l|s[pt]|[bv]e|ppo)	**hist**oire [**i**.stɔ.ʁi]
		hivernal [**i**.vɛʁ.nal]
	in([aeiouy]|é[a-df-rt-z])	**inéd**it [**i**.ne.di]

Each pattern covers quite the range of words. The first accounts for both first and final graphemes, covering the [i] at the start of *y* and of *île*, at the end of *vie(s)*, *lit(s)*, both *Hawaii* and *Hawaï*, among others. The latter two are only concerned with common prefixes, allow of which are indicative of *h muet* words. A number of these prefixes are in fact of Greek origin,[54] dismissing the idea that h-initial words can be dichotomized with respect to etymology.

The final pattern is crucial to prevent false positives triggered by any rule we may define involving nasal vowels. The prefix *in* 'un' [ɛ] will denasalize when the word to which it is prefixed is vowel-initial (Evans, 2019, p. 16). We take extra precaution with the pattern *iné* to prevent false translations for adjectives or past participles ending in *iné(e)(s)*. It may be easier in future to simply separate first-grapheme contexts from final-grapheme contexts. The two processes are executed separately and thus the separation of contexts could be carried out with little issue.

[54] *hyper*, *hydro*, *hygro*, among others.

14. [j]	y[aeiou]	**ya**ourt [**j**a.uʁt]
	ille(s)?	bata**ille(s)** [ba.ta**j**]
	(a|(u)?e(u)?)il(s)?	acc**ueil(s)** [a.kœ**j**]
		trav**ail(s)** [tʁa.va**j**]
15. [k]	k	**k**ilo [**k**i.lo]
	c(k)?(s)?	**c**as [**k**a]
		la**c(s)** [la**k**]
	qu(e)?(nt)(s)?	man**que(s)**, man**quent** [mɑ̃**k**]
16. [l]	l([es]|le(s)?)?	**l**a [**l**a]
		fi**l(e)** [fi**l**]
	[b-zêèâ]les	p**âles** [pa**l**]

As we saw under [j], word-final *l* can contribute to a semi-vowel. Where it does not, l-final words are highly orthographically inconsistent:

33. gentil	[ʒɑ̃.ti]	[l] not realized
34. nombril	[nɔ̃.bri(l)]	[l] may be realized

Although the semi vowels can be encoded thanks to their consistent endings, it is more difficult to account for other examples. Perhaps with POS tagging, we could translate *-il* to [i] for adjectives while for nouns we could avoid this and translate *l* to [l] as in *fil* or *nombril*. However, not even such a distinction based on POS tagging would be consistent. The *l* in the adjective *civil* 'civil' [si.vil] is always pronounced. Resolving this l-final issue would be preferable, as it does not respect VC characteristics: [i] is a vocalic phoneme, while [l] is a consonantal phoneme.

17. [m]	m(e)?	**m**on [**m**ɔ̃]
		tra**m** [tʁa**m**]
	([a-zèù](â)?)(m)?mes	th**ème(s)** [tɛ**m**]
18. [n]	n(e(s)?)?	**n**on [**n**ɔ̃]
		fa**n** [fa**n**]
		fi**ne(s)** [fi**n**]

N-final words will almost always be nasalized – *fan* is a familiar abridged version of *fanatic*, exactly as in English. The integrity of n-final GP translation will fall onto the inventories of our nasal vowels. If these are incomplete, then [n] will act as a fallback grapheme for n-final words. Other than a few exceptions such as *fan*, it will be incorrect if we translate *n* at word

end to [n], so this is something we will strive to minimize.

19. [ŋ]	ng(s)?	campi**ng(s)** [kɑ̃.piŋ]
20. [ɲ]	gn(e)?(s)?	**gn**angnan [**ɲ**ɑ̃.ɲɑ̃]] li**gne(s)** [li**ɲ**]
22. [o]	((h)?ô|o)[ct]?(s)?	**ô** [**o**] **hô**tel [**o**.tɛl] compl**ot** [cɔ̃.pl**o**]
	[e]?au(x|d(s)?)?	**(e)au(x)** [**o**] sal**aud(s)** [sa.l**o**]
	ho(m[éimno]|lo)	**homo**phobe [**o**.mo.fɔb] **holo**graphie [**o**.lo.gra.fi]
	ho(r[ailmort]|lo|n|s[pt])	**horl**oge [**o**ʁ.lɔʒ]
22. [ɔ]	N/A	

French has both [o] and [ɔ]. The former is "generally realized" in closed syllables while the latter is realized in open syllables (Boutin & Turcsan, 2009, p. 141). *Hôtel*, for example, would contain [ɔ] rather than [o] as translated above. As it has no bearing on this application, and as VC characteristics are respected, there is no interest in accurately mapping [ɔ]. We have once more identified many patterns that can be used to identify *h muet* words.

23. [ɔ̃]	on[dst]?	**on(t)** [ɔ̃]
		plaf**ond** [pla.fɔ̃]
	om(b)?	n**om** [nɔ̃]
		apl**omb** [a.plɔ̃]
24. [œ]	œ|oe	**œ**uf, **oe**uf [œf]
	heure	**heure**ux [**oe**.ʁø]
25. [œ̃]	u[mn](s)?	br**un(s)** [bʁœ̃]
	humb	**humb**le [œ̃bl]
26. [ø]	eu[hx]?	p**eu**, p**eux** [pø]

27. [p]	p(e)?(s)?	**p**orc [pɔʁ] ho**p** [ɔ**p**] lou**pe** [lu**p**]
28. [ʁ]	r[det]?(s)?	**r**endre [**ʁ**ɑ̃.dʁ] a**rt(s)** [ʃa**ʁ**] ga**re(s)** [ga**ʁ**]
29. [s]	ç s(se(s)?)? c[eiy] [b-z]ces	**ç**a [**s**a] **s**ur [**s**yʁ] tera**sse(s)** [tɛ.ʁa**s**] **cy**cle [**s**i.kl] comme**rces** [kɔ.mɛʁ**s**]
30. [t]	t [a-zâôû](t)?tes	**t**on [**t**ɔ̃] car**otte(s)** [ka.ʁɔ**t**] fl**ûtes** [fly**t**]
31. [ʃ]	sh(e)?(s)? ch(e)?(s)?	**sh**ampooing [**ʃ**ɑ̃.pu.iŋ] cla**she(s)** [kla**ʃ**] **ch**eval [**ʃ**ə.val] ca**che(s)** [ka**ʃ**]
32. [v]	v(e)?(s)?	**v**ague [**v**ag] sporti**ve(s)** [spɔʁ.ti**v**]
33. [w]	oi[estx]?	**oi**seau [**w**a.zo] f**oi**, f**oie**, f**ois** [f**w**a]

Like both previous semi-vowels, [w] cannot occur in isolation and occurs as part of a diphthong – typically with [a]. As previously, we are not concerned that *oi* is translated to [w] rather than [wa] as VC characteristics are respected.

34. [u]	o[uù][stx]?	**ou**tré [**u**.tʁe] f**ou**, f**ous**, f**out** [f**u**]
35. [y]	[uû](e|es)? hum un[aeiouy]	**u**térus [**y**.teʁys] cr**u(e)(s)** [kʁ**y**] **hum**our [**y**.muʁ] **uni**versité [**y**.ni.vɛʁ.si.te]

The final pattern allows us to distinguish *un* at word-start from *un* at word end. When this pattern precedes a vowel at word-start, its first phoneme is [y]. *Un* at word-end translates to [œ̃]. As a mistranslation would result in violation of nasality characteristics, it is important that we distinguish the two cases. However, it does mean that *une* 'a' [yn] will be translated as monophonemic. It is actually diphonemic. Thus, we will implement a specific procedure for *une* to translate its final phoneme – [n].

36. [z]	z(e)?	**z**oo [**z**o]
	[aeèiouy]se(s)?	th**èse(s)** [tɛz]
37. [ʒ]	j	**j**e [**ʒ**ə]
		te**j** [tɛ**ʒ**]
	g(e|é[a-df-rt-z])	**ge**stion [**ʒ**ɛs.tjɔ̃]
		sa**ge** [sa**ʒ**]

The final pattern for [ʒ] accounts for gé/[ʒ] translations at word-start. We must exclude any cases in which this pattern is followed by *e* or *s*[55] as *é(e)(s)* are common adjectival endings that translate to [e] and not [ʒ].

The above dictionary is evidently non-exhaustive. There are always exceptions to graphemic tendencies. However, it is a strong base for our application. We have stated that it is not crucial that the GP translation is perfect, merely that it respects VC and nasality characteristics to the highest degree possible. Each letter in the French alphabet, barring some that occur only intra-word such as ü, is accounted for in some manner and each of these is represented using a 'fallback' grapheme of sorts as discussed under [n] and [ʁ] among others. As many French letters are orthographically equivalent to their IPA symbol, this fallback should be pretty reliable. The letter *e*, for example, is never going to represent a consonantal phoneme, so if it is mistakenly mapped to [ə] instead of [ɛ] or [e], it is not a major concern and VC integrity is maintained.

NB: each regular expression (regex) pattern is surrounded by \b markers which "matches word boundaries".[56] This means that the sequence or subsequence that we are checking must contain an exact match:

[55]The pattern [a-df-rt-z] means any letter from *a* to *z* excluding *e* and *s*.

[56]`https://ruby-doc.org/core-2.7.0/Regexp.html#class-Regexp-label-Anchors`

Sequence	Subsequence	Matches /ain/	Matches /\bain\b/
train	train	true	false
train	rain	true	false
train	ain	true	true

3.3 Rule Application

As a concept, rule application is rather simple. With respect to a given input, the application iterates through a set of rules and determines whether the input satisfies some series of conditions that lead to the rule being 'fired' or 'triggered'. Within the framework of Optimality Theory (OT), these rules are referred to as constraints. The choice of terminology is of no particular importance and one could argue their preference for either term. This application will use the term 'rule application' for consistency with respect to terms in the field of knowledge representation such as 'rule bases', a.k.a. knowledge bases (KB).

Constraints in OT are defined in a "strict dominance hierarchy" and are said to be "strictly ranked and violable" (Smolensky & Prince, 1993, p. 3). The former term means that once a sample output X has violated a constraint of a certain hierarchy, it can never be a better output than some sample output Y even if this latter violates every single constraint of lower hierarchy. The latter term means that the chosen output does not need to respect every single constraint. This application will respect some of these characteristics but overall, the rule application process will differ from OT's constraint (CON) process. This is simply down to differing application logic in prior processes. In OT, an arbitrarily large series of inputs are determined in the generation (GEN) process. These inputs are in essence possible outputs. The CON process then applies each constraint and determines which of the possible output best respects the strict hierarchy before outputting the evaluation (EVAL). This application does not, however, generate some series of possible outputs. Each sequence is GP translated exactly once and the resulting grapheme-phoneme pairs are ultimately used during rule application.

The notion of rule hierarchy is worth considering. If any rules in this application's KB are triggered, that implies the presence of a liaison. The application could continue to iterate over the KB or it could break the current iteration. The latter may result in several performance improvements. It is impractical to continue the iteration if we have already determined a liaison has taken place. Ultimately, our KB may not incur a heavy performance load but for a large KB, early-exit of an iteration could save thousands of condition evaluations. In doing so, the application relies on the

integrity of the rule hierarchy.

In OT, this hierarchy can only be determined in one manner. As there is a series of potential outputs being processed, the hierarchy must be defined in such a manner that each rule is consecutively less important within the phonology of the language. This is simply the nature of a hierarchy – each element is consecutively 'less important'. As all of this application's rules lead to the same output, i.e. determining that a liaison has occurred, it does not necessarily need to follow this properness hierarchy and the concept of 'important' is not necessarily the same. It could instead hierarchize rules with performance in mind. If a rule X were to feature lower on the properness hierarchy than another rule Y but were more commonly triggered, it may be more pragmatic for this application to adapt its prioritization of rules such that X is 'more important' than Y.

However, all liaisons are not of equal importance: some are obligatory, some are facultative. Logically, it follows that the KB hierarchy should respect this, i.e. that no rule that triggers a facultative liaison would feature higher in the hierarchy than a rule that triggers an obligatory liaison. Delattre's 1947 tableau will serve as inspiration for our rule base. Delattre accepts that the obligatory/facultative/forbidden distinction is not absolute and states that his tableau is based off a 'smart casual' context. This is perhaps the ideal context for this application. An L2 speaker will likely desire to speak eloquently, but not necessarily with the refined properness of a newscaster. However, not all rules in Delattre's tableau will be considered. The inverted verb form, e.g. *ont-ils* 'have-they', will not be marked, as noted in Section 3.1. Furthermore, *formes figées* or composite forms will not be marked. These are multi-word units that form a single nominal phrase such as *Mesdames et Messieurs* 'Ladies and Gentlemen'. This is incompatible with current application behaviour, which splits the overall sequence into individual words or compound words, if the individual units are connected by punctuation. Furthermore, it would require a dictionary of such phrases. Although Delattre provides in the region of 50 such examples that we could encode, there are many more such forms. Although it is "easy to give clear examples", composite forms can be "hard to identify" (Gross, 1993, p. 36). Furthermore, unlike other rules, it is entirely inflexible as it is not based on common characteristics, merely a common statute of 'composite form'.

3.3.1 Rule Base

[1] Personal Pronoun | Obligatory

There are nine personal pronouns in French. The 3PS masculine *il* and 3PS feminine *elle* both end in a consontal phoneme in standard French. 1PS *je* is among the words for which the hiatus is avoided by elision. 2PS *tu* ends in a vowel but has no surface representation in which a latent liaising sound may be realized. Thus, we simply need to check whether our current sequence is among the remaining list of five personal pronouns. An additional check is carried out to ensure that the pronoun is not the noun of a prepositional noun phrase, e.g. *d'entre nous* 'between us', in which case a liaison does not occur.

Examples: ils/ont, on/a, nous/en

[2] Determiner | Obligatory

Delattre's detailed tableau lists twelve 'determinative' forms. Many of them refer to one single word in French, e.g. the definite article *les*. In total, nineteen words are considered for this rule. Many of them can be accounted for using regex patterns as they are orthographically consistent, e.g. *les*/*mes*/*tes*. A few are polysemous, namely *ton* and *son*: the former can be the noun 'tone'; the latter can be the noun 'sound'. As liaising a singular noun is forbidden (Delattre, 1947, p. 152), the preceding sequence is checked against a list of certain singular determinants – *un*, *le*, and *du*. There is one further consideration to be made. Delattre notes a 'special' exception to this rule: the numbers 1, 8, 11 and their derivatives. Only the numbers themselves can be a noun, ergo only the numbers can be preceded by a determiner. However, the textual form of 1 can also represent a pronoun and liaising a determiner with a pronoun is obligatory per Delattre (1947, p. 153). Thus, only the latter two are considered. It is quite rare for either of these numbers to be used as a plural noun ('the eights', 'the elevens') but since the condition is so trivial, these special cases can be accounted for at little computational cost.

Examples: les/amis, quels/horribles, aux/enfants, son/âme

[3] Numeric | Obligatory

As discussed in Section 3.1.2, the list of liaising numbers is short. We sim-

ply check whether the current sequence is the textual or numeral representation of 1, 2, 3, 6, 9, 10, or 20. This functionality is extended to compound numbers, e.g. 42 will liaise. Other than nineteen, numbers in the tens are excluded as numbers 11 through 16 have their own lexemes while numbers 17 and 18 are not vowel-final. Moreover, save *soixante-dix-neuf* (79) and *quatre-vingt-dix-neuf* (99), numbers in the 70s and 90s are excluded as French uses a base-twenty system for these numbers, wherein the number 74 is *soixante-quatorze* 'sixty-fourteen' and the number 88 is *quatre-vingt-dix-sept* 'four-score-seventeen', which in turn uses the 'tens' numbers that we just excluded.

Examples: un/an, deux/heures, 6/enfants, soixante-dix/animaux.

[4] Prenominal Adjective | Obligatory

Delattre's 1947 tableau includes preposed adjectives as an obligatory liaison and he later reaffirmed this: "adjectives always liaise with the following noun" (1956, p. 53). Post (2000), however, found plural prenominal adjectives to liaise in 88% of cases, and this dropped to 61% for certain pairs. Regardless, it is still a high rate for most pairings. This is perhaps the best use-case for POS tagging. If we knew our current sequence was an adjective and that our following sequence was a noun, it would suffice to consult the phonological information of the following sequence, as determined by GP translation, in order to see if a liaison occurs. Regardless, it is relatively trivial to implement. Prenominal adjectives are far fewer in number than postnominal adjectives. As numerical adjectives have already been considered, there are fewer yet. Kelly proposes a non-exhaustive list from which to work (1970, p. 787). A total of fourteen adjectives and twenty-eight forms[57] make up the application dictionary, and regex patterns are used due to commonalities between forms.

Examples: de grands/amis, de bonnes/actions, de nombreux/évènements, les mèmes/auteurs, mes chers/amis

[5] Monosyllabic Preposition | Obligatory

Delattre's tableau includes six monosyllabic prepositions after which liaison

[57] French adjectives are marked for gender and number.

is obligatory, one of which is obselete.[58] Durand and Lyche (2008, p. 17) propose results that "contradict with Delattre" in this respect but note that some monosyllabic prepositions are realized at a rate comparable to obligatory liaisons. *En* is "nearly categorical" at 98% (N = 1124) and generally the rate of liaison realization after prepositions is "overwhelming". This is corroborated by Moisset (2000, p. vi) who found prepositions to demonstrate "high rates of liaison" across fourteen Parisian speakers from different backgrounds and in different contexts.

Seven prepositions including *en* are included in this application. Durand and Lyche (2008, p. 17) observed liaison realization of 95% after *dans* and 88% after *chez*. They proposed that this latter was affected by the category of the following sequence. Liaison realization was "categorical" when followed by a monosyllabic pronoun. If we were to examine facultative tendencies and make our classifications under a probabilistic interpretation with this relevant POS information, we could perhaps better inform application users as to whether a liaison is obligatory. Durand and Lyche could not make conclusions with respect to *sous* and *sans*, but we will respect Delattre's judgment until further information is at hand. Finally, *dès* and *quant* are included because their usage is so limited and within these uses, they are often followed by a vowel-initial word, e.g. *quant à* or *dès à*.

Examples: quant/aux, dans/un, en/effet, sans/encombre

[6] Monosyllabic Adverb | Obligatory

Delattre includes nine monosyllabic adverbs after which liaison is obligatory in his 1947 tableau. However, he later concedes that on closer examination, monosyllabic adverbs have some "facultative tendencies" (1956, p. 51). This is consistent with other findings. Ranson found the 'obligatory' *quand* to liaise in 69% of cases (2008, p. 1673). The 'obligatory' negation marker *pas* was followed by a liaison in just 21.37% of cases across five corpora (idem). *Trop* entailed a liaison in just 15.4% of instances (N=52) (Mallet, 2008, p. 216). These simply cannot be considered obligatory. Some rates of liaison seem highly dependent on context. *Pas + adjective* is liaised in 43% of cases while *pas + preposition* is liaised in just 9% of cases according to Booij and De Jong (1987, p. 1012). The application cannot encode this additional information without POS tagging.

Ultimately, only five monosyllabic adverbs are considered by the application. *Très* and *plus* are frequently liaised – 96% (Ranson, 2008, p. 51)

[58] *dèsà* is written *dès à* and is a) not monosyllabic, b) a composed preposition.

or 63% (Malécot, 1975, p. 164). Additionally, *bien*, *rien* and *quand* are included. It is worth noting that *plus* can be an adverb 'more' or be part of the disjointed morpheme *ne...plus* 'no longer':

28. *Je suis plus/intelligent*

    I am more intelligent

    [ʒə.sɥi.ply.**z**ɛ̃.tɛ.li.ʒɑ̃]

29. *Je ne suis plus intelligent*

    I am no longer intelligent

    [ʒə.nə.sɥi.ply.ɛ̃.tɛ.li.ʒɑ̃]

If the *ne* is present, no liaison takes place. If not, then it is acting as the adverb 'more' and will likely liaise with its succeeding sequence. We can check for this disjoint morpheme. The *ne* can be separated from *plus* by several elements: direct object, indirect object, verb phrase. This latter can be two elements in any composed verbal form. Thus, we can expect a maximum distance of four. The *ne* can also drop to *n'* if the following word is vowel-initial. However, the check will not work if the *ne* is dropped. Ne-deletion is more typical in oral contexts, with *ne* present in just 18% of negated verbal phrases (Ashby, 2001, p. 8). In that case, (28) would be identical to (29) as written, but the two meanings could be distinguished by liaison. This deletion is reflected in casual online communication. *Ne + verb + plus* was reduced to *verb + plus* in 86.1% of non-moderated chat, though *ne + verb + plus* was respected in 100% of moderated chat (Williams, 2009, p. 478-479). Although ne-deletion is becoming more common, we can expect Liaison Marker to be used primarily for practising oration of news articles or of academic assignments. As such, we can expect text adherent to typical written standards, in which the *ne* would not be dropped, and thus we can expect our check for the negation element to be successful.

Examples: bien/entendu, plus/intelligent, très/aimable

[7.] Common Verb | Facultative

Generally, verbal liaison is facultative. Delattre affirms as much in his 1947 tableau. He later expands that liaising is "quite frequent" between verb and adverb and "somewhat frequent" between verb and complement (1956, p. 53-54). Previously, however, we saw *est* to be liaised in 97% of cases and several others to be realized in 75% to 86% of cases (Thomas, 2002,

p. 105). Furthermore, we discussed the quite common liaison of third-person plural *-er* verbs. As the application lacks POS information, the current approach is to simply store a list of common verbal liaisons based on Thomas' findings. These are *est*, *sont*, *était*, and *ont*. Although Thomas found *suis* to liaise in just 47% of instances, I would expect that figure to be higher in any non-casual context. There is a contraction phenomenon in French that may explain this low rate of liaison. *Je suis* [ʒə.sɥi] 'I am' will often reduce to [ʃɥi], transcribed as *chui*, in oral contexts (Tran et al., 2008, p. 1855) (Bertrand et al., 2008, p. 112). This contracted form does not have a latent final consonant, i.e. there is no possibility to realize [ʃɥi**z**], thus it cannot liaise. When it is realized in its uncontracted form, it may liaise. Application users are more likely to be seeking to reproduce a more formal, academic, context and so modelling with the uncontracted form in mind is best practice.

Examples: est/invité, sont/allés, était/invincible, ont/eu

[8.] Auxiliary Verb | Facultative

We discussed auxiliary verbs under *3PL Present Tense* in Section 3. The verbs were *pouvoir*, *devoir* and *vouloir*. It is also possible to liaise these auxiliary verbs in the 3PS/3PL in the imperfect and conditional forms.

Examples: devrait/être, pourrait/avoir, veulent/être

[9.] Third Person Imperfect or Conditional Verb | Facultative

Ågren (1973) found third-person conditional endings to have a "large disposition" to liaise. Delattre also includes several imperfect verbs conjugated in the 3PS in his tableau (1947, p. 154). The two tenses share common endings in the 3PS/PL form, *-ait* and *-aient* respectively. The former needs further consideration. Certain common G3 verbs such as *connaître* 'to know' or *faire* 'to do' conjugated in the third-person singular also end in *-ait* and also may liaise. Rather than prevent these valid liaisons being triggered by this rule, we can simply show the user a different reason, i.e. not *third person imperfect or conditional verb*, if the liaison is triggered by one of these 3PS present-tense forms.

Examples: avaient/un, vivaient/en, seraient/au

[10.] Polysyllabic Preposition | Facultative

Delattre includes this category in his 1947 tableau and later affirms that this liaison is "generally quite frequent" (1956, p. 52). However, he also affirms that polysyllabic adverbs such as *beaucoup* are quite frequently liaised, whereas empirical data found *beaucoup* to liaise in just 6.5% of instances (N=46) (Mallet, 2008, p. 216). This casts doubt upon Delattre's earlier statement. Furthermore, the same author did not observe any polysyllabic prepositions to liaise and tentatively concludes that this may make the polysyllabic preposition an "erratic liaison" (idem, p. 280). However, the author notes that it is completely possible to liaise *après* in a sequence such as *après avoir mangé* 'having eaten' and thus attributes the results to an issue of corpus and style. Certainly, I would reject the initial conclusion. The previous example with *avoir* would be quite a frequent liaison. Others may be rarer and realized generally in a more formal register, but I have included them nonetheless.

Examples: pendant/un, devant/un, après/avoir

[11.] Inverted Nous/Vous | Facultative

We previously discussed not marking the obligatory inverted-verb-form liaison due to the presence of a hyphen that already acts as a marker of union. This form would not liaise if the pronoun were *nous* or *vous*. These two cannot find themselves as the latter sequence of a hiatus as both begin with a consonantal phoneme. They can, however, be the former sequence of a hiatus. In Rule 1, they led to an obligatory liaison as a personal pronoun. When in this form, however, the liaison is facultative. It is easy to distinguish the two liaisons. A regex pattern is used to check whether the sequence ends in hyphen-nous or hyphen-vous.

Examples: Avons-nous/un livre ? Placez-vous/à droite.

[12.] Impersonal Pronoun | Facultative

Seven words are accounted for by this category. There is a lack of empirical data for this category, so the list is based primarily on Delattre's tableau (1947, p. 154).

Examples: dont/on a parlé, plusieurs/ont, eux/aussi, certains/ont, toutes/et

[13.] Pre Invariable | Facultative

Delattre lists five categories after which a pre-invariable liaison can occur: plural noun, plural adjective, pronoun, verb, polysyllabic adverb (1947, p. 155-156). The third category has been accounted for by Rule 12. Delattre lists nine letters after which a liaisons can occur: s, z, x, t, d, n, r, p, and g (sic, p. 150). We do not have POS information available. As such, we will focus on the plural noun and the plural adjective. Of the nine liaising consonants listed by Delattre, only two may denote a plural: *s* and *x*, both of which liaise as [z]. An additional check is carried out in the former context to ensure that the word did not already have [z] as its final phoneme, e.g. *française(s)* 'Frenchwoman(women)' [fʁɑ̃.sɛz]. In this case, the plural marker *s* will not be realized as a second consecutive [z]. Instead, the existing [z] will be linked to the following sequence by *enchaînement*. Additionally, any sequence that is [ʁ]-final is very unlikely to liaise (Delattre, 1955, p. 155). Thus, we ignore such sequences – even if they end in *s* or *x*.

Several verbal forms also may end in *s*: *prends* (1SG/2SG) or *avons* (1PL). It is important that this rule is listed below other obligatory rules that share this context: e.g. the plural determinant *les* or the impersonal pronoun *eux*. The invariables we will consider are *et* 'and', *ou* 'or', and *à* (polysemous). We discussed 3PL verbs in Section 3 and again when defining rules 8 and 9. Rule 8 accounts specifically for auxiliary verbs while rule 9 accounts for conditional/imperfect third-person verbal forms. We can make one final precision in this 'Pre Invariable' rule for 3PL present tense verbs before *à* or its derivatives *au* and *aux*. The condition will also account for t-final polysyllabic adverbs (Delattre, 1947, p. 156).

Examples: violets/ou, plans/à faire, parlons/à, vivent/à

[14.] Nasal Cacophony | Facultative

This is perhaps the most experimental rule. In a previous analysis, I discussed the concept of a "nasal cacophony" with respect to the liaison (Evans, 2019, p. 14). Such a cacophony occurs when both vowels in a hiatus are nasal. I affirmed that such a context would lead to the first nasal vowel (NV) breaking down into its vowel (V) and nasal consonant (NC), i.e. a shift from NVNV to VNCNV. A general nasal rule where only the first vowel in the hiatus needed to be nasal was originally considered. It is quite an effective rule and, in fact, covers many different liaisons. Such cases range from the personal pronoun *mon* of Rule 1 to the adjective *grands* of

Rule 4, from the monosyllabic preposition *dans* of Rule 5 to the 'common verb' *sont* of Rule 7. However, while nasal vowels do seem to have a high tendency to break down, the rule proved to be a little gung-ho and was too general. It had a high rate of false positives. Moreover, it did not distinguish obligatory liaisons as in rules 1 or 4 from rare liaisons like a plural noun followed by a verb, which Delattre found to be "rare in conversation" and "not common, even in reading" (1956, p. 49). It is perhaps incidental that this nasal rule spanned multiple classes of obligatory liaisons. It could be the case that nasal vowels have a higher tendency to liaise, but only when belonging to one of these categories. This may testify once more to the liaison's nature as a historical, grammatical, phenomenon rather than a phonological phenomenon. This is corroborated by Delattre's tableau (1947) in which several rules forbid liaising after nasal vowels. These include *singular noun + x* or *impersonal pronoun + x*, respectively:

30. *Le train est arrivé*

    The train arrived

    [lə.tʁɛ̃||ɛ.ta.ʁi.ve]

31. *Quelqu'un arrive*

    Someone is coming

    [kɛl.qœ̃||aʁiv]

Delattre states that "any liaison [after nasals] that is not obligatory is forbidden" (1947, p. 150). However, he contradicts this statement in his own tableau. Each of the nasal-final *pendant*, *dont*, *sont*, and *souvent* is listed under a different facultative rule. It would appear that a nasal cacophony can lead to a liaison, but only when the liaising consonant is not [n]. Note in the previous four that the liaising consonant is [t]. Moreover, it is possible to liaise when the liaising consonant is [d] or [s]:

32. *Il prend/une pause*

    He takes a break

    [il.pʁɑ̃.**d**yn.poz]

33. *Nous avons/un ami irlandais*

    We have an Irish friend

    [nu.za.vɔ̃.**z**œ̃.a.mi.iʁ.lɑ̃.dɛ]

Thus, we will mark a liaison in a nasal cacophony wherein the first sequence is not n-final. This condition also accounts for contexts such (30) and (31) wherein liaison is forbidden.

Examples: temps/en temps, parlons/anglais

4 Implementation

4.1 Basic Infrastucture

Having modelled each step of the liaison-marking process in Section 3, it comes time to put theory into practice. Before determining specifics and applying our rule-set, however, the general architecture must be encoded. As we are using a relational data model, it is appropriate to envisage how the various phenomena we wish to model relate to each other:

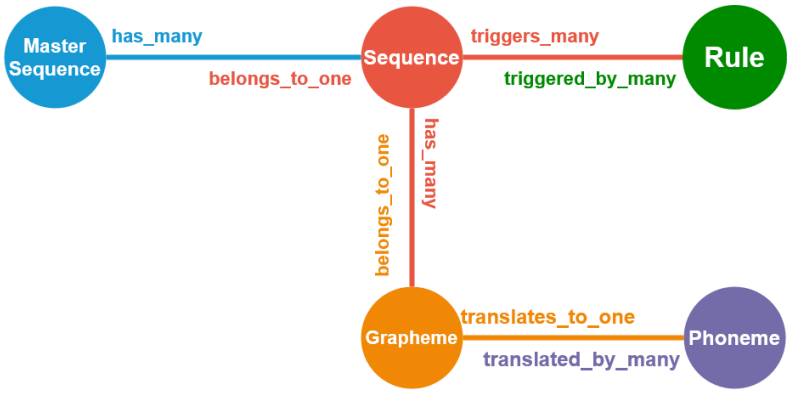

Figure 10: Entity Relationship diagram. It is to be read 'A master sequence has many sequences', and so on.

Very little is actually done with the *Master Sequence*, although its relational dependents such as the *Sequence*, *Grapheme*, and *Phoneme* are all pivotal to marking liaisons. However, it is ultimately the most important entity we are modelling as it is at the base of the application. If the initial chain of user input is not successfully received by the application, no subsequent processes – tokenization, GP translation, and rule application – are carried out. In Ruby on Rails, user-application interaction is determined by the defined *routes* in the application configuration:

```
Rails.application.routes.draw do
  root to: 'master_sequences#new'
  resources :master_sequences, only: [:new, :create, :show]
end
```

Figure 11: Application route configuration.

Ultimately, this application has few routes. The first line determines the *root path* and this is determined by what the "most popular route" is within an application.[59] Any user visiting this application likely wishes to mark the liaisons for some *new* chain of text. The *resources* keyword is a Rails shorthand way to define the seven basic controller actions.[60] Current application behaviour only requires three of these seven: *new*, *create* and *show*. This order captures the user-flow perfectly. A user visits the application and is presented with a form in which they enter the text for a new Master Sequence. If the Master Sequence is successfully created, the user is then shown their marked sequence. Each of the three major steps outlined under Methodology is implemented after the successful creation of a Master Sequence and before the user is shown the updated Master Sequence with any liaisons marked.

4.2 Application Logic

Once a Master Sequence is successfully created, the application can begin to implement its core logic: tokenization, GP translation, and rule application. This is done through a series of *callbacks*. A callback is a very powerful tool in Rails and can be implemented in a variety of manners. For example, a *before_ destroy* callback may verify that the user has permission to destroy an object, or that its destruction will not violate data integrity. In this way, callbacks can be likened to *triggers*[61] in SQL. We are most interested in the *after_ create*[62] callback. This will allow us to implement the core application logic before the user is redirected to the *show* action for the given Master Sequence. Although the Master Sequence is created as part of the controller's response, these callbacks are defined within the model itself and not its controller:

[59] https://guides.rubyonrails.org/routing.html#using-root
[60] https://guides.rubyonrails.org/routing.html#crud-verbs-and-actions
[61] https://dev.mysql.com/doc/refman/8.0/en/trigger-syntax.html
[62] https://guides.rubyonrails.org/active_record_callbacks.html

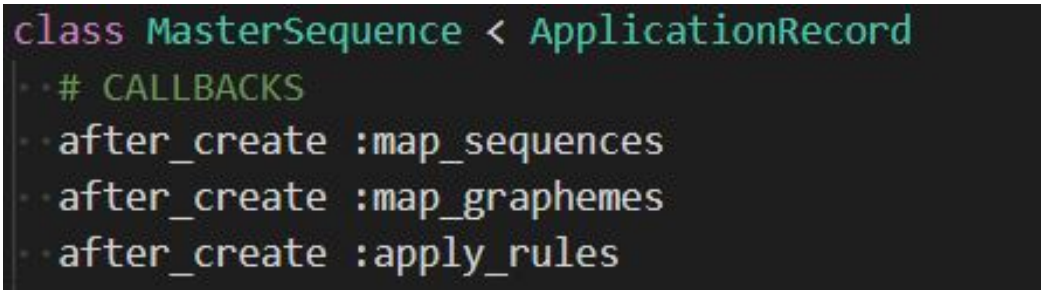

Figure 12: Master Sequence callbacks.

This respects the core Rails practice known as *Fat Model, Skinny Controller*.[63] Laycock states that "any non-response-related logic should go in the model".[64] As none of these processes are relevant to the reception of user data nor to the display of a marked Master Sequence, they should thus be implemented in the model.

Although three callbacks are defined, there is not a one-to-one relationship with the three steps outlined in Methodology. Tokenization and GP translation are done on individual sequences, whereas rules are applied to sequence pairs. Alternatively, a single callback could break down the master sequence into pairs of sequences which are then tokenized, translated, and passed to the rule applicator. However, there is no benefit to this. It is arguably better practice and more transparent to keep different aspects of the application logic separate.

Mapping sequences is a very trivial process. We split the Master Sequence text into words or sequences whenever there is a space in the text. A record is created for each of these sequences and they are associated to the master sequence from which they came. Once each of these sequences are created, the application procedes with the *map_graphemes* callback. This callback implements tokenization. It does not execute GP translation for sequences that correspond solely to punctuation. Several punctuation markers in French require a space beforehand, after, or in both cases (Gedda, 2003, p. 61). The application splits the initial user text on space-boundaries, so it will create individual sequences for these punctuation marks. We thus save resources by not executing a futile dictionary search for corresponding phonemes. It also does not execute GP translation for initialisms:

[63] `https://www.devinterface.com/en/blog/rails-best-practices-1-fat-model-skinny-controller`

[64] `https://www.sitepoint.com/10-ruby-on-rails-best-practices/`

Algorithm 1 map_graphemes

```
loop sequences:
    next if solely_punctuation? || initialism?
    strip_punctuation
    deal_with_hyphenated_words
    map_first_grapheme

    create_schwa_phoneme if monosyllabic_schwa?
    create_n_phoneme if une?
    next if monophonemic? || diphonemic?

    map_last_grapheme
```

Before GP translation occurs, the sequence is stripped of leading or trailing punctuation as punctuation does not carry any graphemic information. This could be an opening or closing guillemet or a comma, among others. A separate process deals with hyphenated words. We discussed in Section 3.1.1 that a grapheme cannot span a hyphen. Given a sequence *peut-être* 'maybe', we will search the first grapheme from the first component *peut* and the last grapheme from the last component *être*. We may note the presence of code that may execute between *map_first_grapheme* (MFG) and *map_last_grapheme* (MLG). This is to account for certain common words that would otherwise be mistranslated. Consider *de* 'of' [də]. Our GP dictionary contains de/[d] for those words that end as such, e.g. *fade* 'dull' [fad]. This would make the word monophonemic and we would thus not map its last grapheme. It is not, however, monophonemic. Additional, *une* requires manual intervention. This prevents *une* from being treated as if *un*/[œ̃] were its first grapheme-phoneme translation. This nasality violation could incorrectly trigger the nasal cacophony rule.

If the word is neither monophonemic nor one of these words for which we have already mapped the second and final grapheme, the algorithm continues to the MLG call. Both MFG and MLG are recursive methods. If no matching grapheme is found in the dictionary, the method calls itself. The next iteration will take a smaller subsequence. For example, if the Sequence content were *éléphant* 'elephant', it might search for *élép* in the dictionary. Were this not to match a dictionary entry, MFG would run again on *élé*. This continues until the empty string, in which case the iteration terminates. Both methods are recursive terminal. This is crucial as it means the program will not become stuck in a recursive loop. Ideally, however, the program would never reach the point at which it must termi-

nate the recursive call. If either MFG or MLG terminate without matching any grapheme, it means that the application has failed to GP translate the Sequence and thus, it cannot accurately determine whether a liaison has occurred. This will be considered this when evaluating the integrity of the GP translation process.

The final callback calls upon a 'RuleApplicator' service. Services are a "holy grail" in Ruby, implemented for "specific business actions" in order to prevent bloating in our Model/Controller classes.[65][66] Moreover, they are particularly useful when implementing an action wherein several models interact. Our rule-base relies on sequential, graphemic, and phonetic information. The service is instantiated for each triple of sequences {i,j,k} in a Master Sequence of {1..ijk...n} sequences, wherein j is the current sequence. The one rule that requires the use of information relating to the previous sequence i is the obligatory 'Personal Pronoun' (PP) rule, wherein the liaison does not occur if the PP is part of a prepositional noun phrase.

A modicum of experience with established functional programming languages such as Java or C++ might trigger someone to consider out-of-bounds exceptions in the triple {i,j,k}. When j is the first sequence, there is no previous i. Likewise, when j is the final sequence, there is no next k. Fortunately, Ruby is more dynamic in this respect. In the event an array-like structure[67] does not contain a certain index, a null object is returned. We are aware of the fact that we may access a null index and thus may implement appropriate procedures for these cases. The iteration over the rule base is remarkably concise:

```ruby
def call
  return if next_sequence.nil?
  return if onset.consonant?
  return if current_sequence.ends_in_punctuation?

  Rule.active.prioritized.each do |rule|
    trigger(rule) && return if send("#{rule.code}")
  end
end
```

Figure 13: Rule are applied by 'calling' the RuleApplicator service.

[65] https://medium.com/selleo/essential-rubyonrails-patterns-part-1-service-objects-1af9f9573ca1

[66] https://www.toptal.com/ruby-on-rails/rails-service-objects-tutorial

[67] In this case, the sequences are a Rails 'CollectionProxy'.

The first three lines are known as 'guard clauses' in Ruby, used to prevent "the rest of your code from executing if not necessary".[68] The first clause accounts for null-indexing. A liaison cannot occur if the onset of the following sequence is consonantal. We discussed in Tokenization[69] that liaison does not span punctuation such as full-stops, commas, or exclamation marks. The rule application algorithm embodies the dynamism of Ruby on Rails. From our database, we access any rules that are *active*. If at any point a rule proved to be inaccurate and needed reconsideration, we could simply deactivate it from within the application or the command-line with the click of a button. No logic would need to be changed and the code would remain untouched. If we were able to solve the issue with the rule, we could then simply reactivate it and the relevant code would be executed.

We discussed the idea of rule hierarchy in Section 3.3. We apply the rules in a *prioritized* manner with respect to the order they were defined in Section 3.3.1. Each rule has an associated code. The RuleApplicator calls the method that corresponds to this code, i.e. the method in which various conditions are evaluated to determine whether a liaison associated to a given context occurs. If the conditions are satisfied, we *trigger* the rule and *return*. The latter allows for early-exit of the iteration. Given that we have identified a liaison, we do not need to continue searching for another liaison-inducing context among the remaining rules of lower priority. The *trigger* operation creates an association between the rule and the current sequence. It is this association that we will use to determine which sequences will be marked in the output.

4.3 Application Interface

The application has a simple interface. The screen is bisected; the left-hand side provides a textbox for user input; the right-hand side displays the marked output:

[68] `https://anthonygharvey.com/ruby/guard_clauses_vs_nested_if_statements`
[69] See Section 3.1.1 – Punctuation.

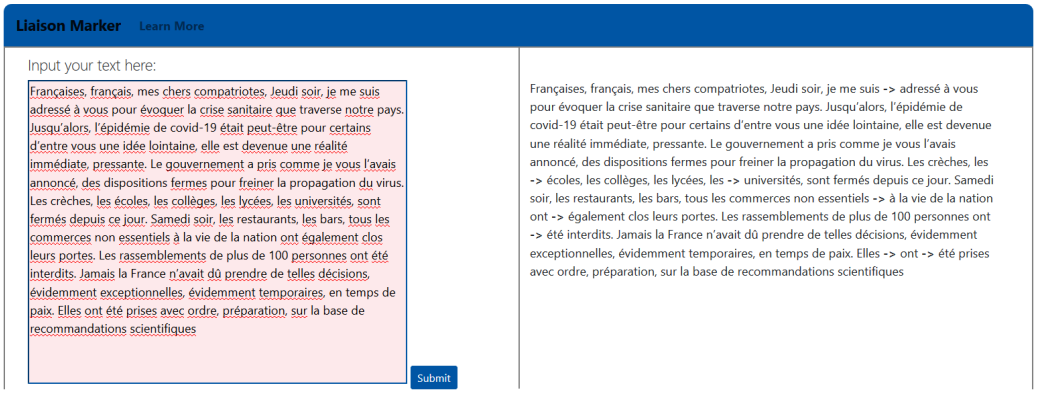

Figure 14: Part of Macron's address marked (right) by Liaison Marker.

It may be difficult to discern without zooming in, but the text on the right-hand side of Figure 14 has been marked by the application. An emboldened arrow[70] was chosen to represent the idea that one word leads in to the next. On hover, the cursor changes to show a click-prompt, inciting the user to click the arrow. The user is then presented with information regarding the liaison – whether it is obligatory/facultative, the rule that triggered the liaison, a short explanation of the liaison and pertinent examples. This information is represented using a modal, commonly referred to as a pop-up:

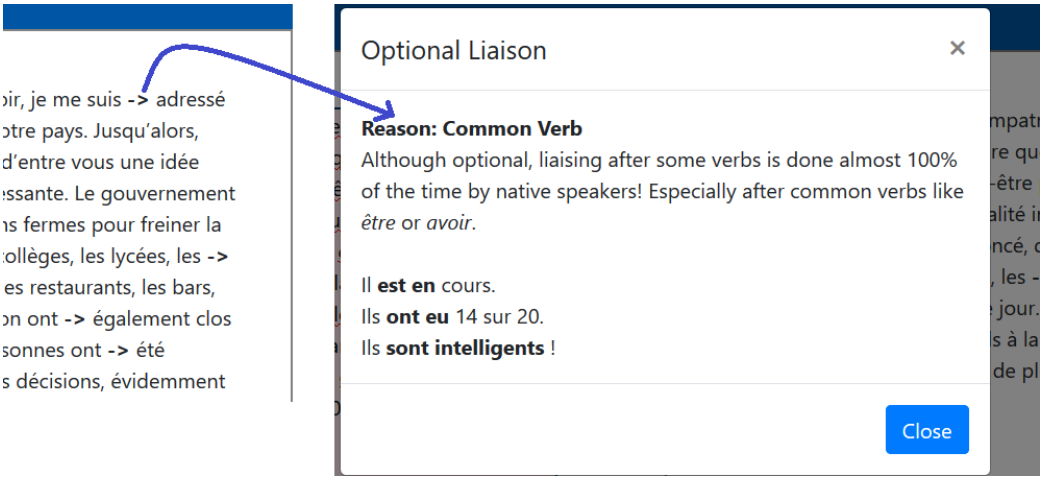

Figure 15: Marked liaison and the informative modal that shows on-click.

[70] ->

5 Results

While Delattre's 1947 tableau has had a strong bearing on the application rule-base, the inconsistencies between his use of *obligatory* and its use in other literature will lead to inconsistent results if the examples given by Delattre are marked by this application. Delattre accepts that the obligatory/facultative/forbidden distinction is "not absolute" (1947, p. 152). Furthermore, the tableau was composed over seventy years ago and thus may not reflect the current state of French. Finally, the tableau contains inconsistencies: *rien à faire* is obligatory, *divin à voir* is forbidden. The former is obligatory because it is a monosyllabic adverb, the latter is forbidden because it is an invariable form. Not only have we observed monosyllabic adverbs to display facultative tendencies, *rien à faire* is an invariable form, ergo forbidden. If examined within the framework of Optimality Theory, perhaps the adverbial constraint would outrank the invariable constraint. However, no such approach is outlined by Delattre.

Five news articles, one from each of five major French newspapers,[71] will be used to test our application. The articles were chosen in the final days of March 2020. Each consists of between 330 and 700 words and each is taken from the front page of a different newspaper section, e.g. Sport or Business. This methodology seeks to provide a diverse range of writing and vocabulary for the testing of the application. Each article has been manually marked by myself and by a native French teaching assistant at Trinity College Dublin. The liaisons were assigned to one of three categories: obligatory, common facultative (CF), and rare facultative (RF). The distinction between the latter two category is based on empirical data where possible but is ultimately subjective. The three categories are of descending importance and the application is primarily concerned with the first two categories.

We will also consider the first five minutes of Emmanuel Macron's « *Adresse aux Français* » on March 16th 2020. The presidential address is accompanied by a teleprompter transcription of Macron's speech, but the application uses a redacted version to reflect exactly what Macron says.[72] The liaisons were assigned to the three categories outlined above.

[71] Le Monde, Le Parisien, Le Figaro, L'Express, and Libération.

[72] For example, in the list *les crèches, les écoles, les lycées, les universités*, the teleprompter drops some of the *les* to transcribe quicker, even if Macron says each of them.

5.1 Grapheme-to-Phoneme Translation

5.1.1 Integrity

In Section 4.2, we discussed the recursive-terminal nature of the GP translation algorithm. This means that the application will not infinitely recurse if it cannot find any applicable grapheme-phoneme pair. Thus far, out of 1749 unique sequences, none have required this unsuccessful termination of the GP translation process. Rails' extensive functionality allows verification of this information in just one command. If the scope of the application were expanded, this could be run as a simply daily or weekly report to refine the GP translation process and to ensure its integrity.

Algorithm 2 check_for_unfinished_translations

```
loop sequences: |s|
    # second condition excludes (correctly translated) monosyllabic sequences
    if s.graphemes.length == 1 & s.graphemes.first.content ≠ s.content
    # second condition excludes punctuation/initialisms
    or s.graphemes.length == 0 & s.content.match?(/[a-z]/)
      s.report_unfinished
```

5.1.2 Accuracy

The first 74[73] words of Macron's address and of each of the five articles were verified for GP accuracy, for a total of 444 words and up to 888 GP translations. The correct phoneme was determined in **98.67% (817/828)** of cases. Four of eleven misclassifications pertained to a proper noun. We stated in Section 3.1.3 that proper nouns are an unpredictable category with respect to GP translation. Many proper nouns have retained their historic pronunciation. Imported foreign nouns also contravene typical French phonetic tendencies. Other mistranslations are simply difficult, if not impossible, to model. *Ecouter*, for example, is normally written as *écouter* wherein the *é*/[e] translation is nigh on infallible. However, a characteristic of modern French journalism is to eschew punctuation atop capital letters. Without this diacritic, the application loses critical phonological information.

Nasality characteristics were respected in all but one instance. VC characteristics were respected in just under half of the misclassifications.

[73] If this seems arbitrary, it is because I chose 75 and then mistakenly formatted all of the appended tables to fit 74. Plus 75 is just as arbitrary anyway, eh?

Nonetheless, these are encouraging results. It is a robust and diverse corpus, yet the application translates to the correct phoneme in roughly 99% of cases, with a range of 1.4% and a relatively low standard deviation of 0.66%. With respect to current data, we can say to a 99.9% confidence level that the application will translate at worst 97.8% of phonemes correctly. It is worth noting that some words feature in several if not all of the articles, e.g. prepositions (à, de), determiners (les, des), et cetera. This is not necessarily a fault of the corpus. It is simply the nature of language. These units are the common building blocks for sentences and thus we would expect much of the user input for our application to reflect this. It would nonetheless be beneficial to see if these results were to extrapolate over a larger corpus.

5.2 Liason Marking

5.2.1 News Articles

209 possible liaisons were manually identified across the five articles – 78 obligatory, 55 CF, and 75 RF. 162 of these (**77.51%**) were marked by the application. One further context was incorrectly marked as liaison. The precision of the classification is thus **99.39% (162/163)**. The false positive was a case of *enchaînement* and not liaison. Given that *enchaînement* and liaison ultimately lead to similar results in that both processes involve the 'linking' two words via the Maximal Onset Principle, this false positive is at least preferable to identifying a forbidden liaison.

97.44% (76/78) of obligatory liaisons were marked. A review of Table 1 tells us that even native speakers occasionally err – obligatory liaison realization was 96.9% among native speakers (N=2667) (Malécot, 1975). If our data is conducive to extrapolation, then the application will be of great benefit to speakers at a beginner-intermediate level. Both false negative classifications could be rectified if POS information were known to the application. The list of approximately thirty prenominal adjectives known to the application does not include *riches* 'rich (pl.)', nor does it include *merveilleux* 'marvellous'. Both of these can be either preposed or postposed adjectives. Generally prenominal adjectives are monosyllabic, unlike *merveilleux*, but there is a certain stylistic licence to prepose an adjective and sometimes the position is semantically motivated (Abeillé & Godard, 1999, p. 13). Though Rule 4 does account for the most common preposed adjectives, e.g. *petit* 'small', *grand* 'big', or *bon* 'good', it is unsatisfactory for both prenominal adjectives in our corpus.

85.45% (47/55) of common facultative liaisons were marked. There

are some identifiable patterns among the misclassifications. Five of these liaisons are in a context of *plural noun + adjective*, listed among the rules in Delattre's tableau (1947, p. 153). Delattre later went on to affirm that this liaison in this context is "quite rare in everyday conversation" but is "highly sensitive to style", becoming more frequent in a more refined register and very frequent when reading (1956, p. 49). It would be rather trivial to define a rule to account for this context if the application implemented POS tagging. Other CF misclassifications would benefit from a contextual analysis. Of the four possible liaisons concerning the negation marker *pas*, one was judged to CF while three were judged to be RF:

34. [*ne*] *soit* [*pas encore*]
    are [not yet]

35. [*ne*] *soit* [*pas assez*]
    are [not quite]

36. [*n'*] *était* [*pas*] *emballé*
    were [not] thrilled

37. [*ne*] *riment* [*pas*] *avec*
    are [not] synonymous with

The *pas* in examples (34) and (35) is qualified by the adverb that follows, i.e. this latter modifies *pas*. Syntactically, *pas* is said to 'dominate' this modifier. In examples (36) and (37), however, it dominates the verb phrase as a whole: NOT [were thrilled], NOT [are synonymous with]. This notion of dominance can be related to the idea of *union* that we introduced when first defining the liaison.[74] In that case, we were discussing a phonological union wherein the liaising consonant prevented a hiatus. Here, the union is syntactic. Although the lack of union can be a distinguishing factor in non-liaison in (36) and (37), the union between sequences in (34) and (35) differs despite each *pas* dominating the following element. I would tentatively hypothesize that the syntactic union in (35) is weakened by its phonological context. *Pas* is [a]-final while *assez* is [a]-initial. The cacophonous [a.a] hiatus is not avoided by the realization of a liaising [z]. Instead, the 'double' [a] reduces to a sequence resembling an elongated [a:] that is split by a modicum of glottal obstruction.[75] This differentiates (38) from (39):

[74] See Section 2.1.3.

[75] This will be represented by a superscript glottal stop [ʔ]. [aˀ] conveys the notion of a glottal stop that is less distinct than in [aʔa].

38. *Ils sont pas assez...*[76]
    They are not quite...
    [il.sɔ̃.pa$^{?}$.se]

39. *Ils sont passés.*
    They came over.
    [il.sɔ̃.pa.se]

5.2.2 Macron

Macron did not realize any of the twenty RF liaisons. Five of these were marked by Rule 13, wherein a plural noun, plural adjective, or verb liaised with an invariable. Of the twenty-three liaisons realized by Macron, fourteen were judged to be obligatory and nine were judged to be CF. Thirteen of the fourteen obligatory liaisons were marked while eight of the nine CF liaisons were marked. If we were to consider Macron the gold standard of liaison realization, this would give us an accuracy of **83.72% (36/43)** and a precision of **80.77% (21/26)**. With respect to obligatory/CF liaisons, the accuracy is **91.30% (21/23)** and the precision is **100.00% (21/21)**. Ultimately, no one person is the gold standard and, critically, no forbidden liaisons were marked. Let us consider the two obligatory/CF judgments in which the application differed from Macron:

40. non/essentiel

41. mieux/accueillir

(40) is the sole obligatory liaison that was not marked by the application. It is not accounted for by any of Delattre's 1947 rules. We discussed hyphenated words in Section 3.1 under punctuation and (40) is an example of a sequence that can be written with-or-without-hyphen. Indeed, in an article on April 15th 2020,[77] Belgian journalist Jennifer Mertens alternates between the hyphenated *non-essentiel* and the unhyphenated *non essentiels* from one sentence to the next. The transcription of Macron's speech uses the unhyphenated form and the application is unable to identify that there is a union between the two words – *non* is a modifier of *essentiel*. As *non* can be used as a nominal phrase or as an exclamative, it cannot be presumed that it is a modifier of the following sequence. The POS tagger

[76] Ne-deletion as in (38) is common in casual oral contexts, but the non-deletion of *ne* could also distinguish (38) from (39).

[77] `https://geeko.lesoir.be/2020/04/15/la-france-force-amazon-a-cesser-le-commerce-non-essentiel/`

TreeTagger[78] is able to determine that the *non* in (40) is indeed an adverb, i.e. that it may modify the following sequence. The application could account for the liaison if that POS information were known. (41) is a facultative liaison realized by Macron but unmarked by the application. *Mieux* is not listed by Delattre (1947) among the monosyllabic adverbs that liaise. It figured in a category that liaised in just 7.4% of instances (N=27) per Ranson (2008, p. 1676). It may thus be harsh to consider this an 'incorrect' judgment, even if Macron realizes this liaison.

5.2.3 H muet vs. H aspiré

105 h-initial words[79] were tested in the context *des + plural noun*. This is an obligatory liaison for *h muet* words. It is forbidden to liaise *h aspiré* words in any context. The words were selected from 2,057 h-initial words listed on the *L'internaute* dictionary.[80] The words were selected using a random number generator and were ignored if they could not be pluralized as a noun. As *L'internaute* is a web-based dictionary that admits some informal and otherwise unrecognized neologisms, the validity of each word was cross-checked against the *Larousse* dictionary.[81] The data items were spread as evenly as possible across the six vocalic letters that can follow the letter h in French. 23 were *Ha*, 18 were *He/Hé*, 11 were *Hi*, 28 were *Ho/Hô*, 10 were *Hu*, while 14 were *Hy*. The counts are proportional to the total number of words listed under the Hx words. Furthermore, some of these Hx words are more likely to repeat patterns as they contain prefixes: 29.9% of Hy words are prefixed by *hydr(o)*.

The application made the correct judgment in **103 of the 105** cases, producing an accuracy of **98.1%**. The two misclassifications were false negatives, i.e. the precision of the classifier was perfect. In no case would the application misguide a user into producing a forbidden liaison. The two misclassifications are infrequently used words. *Hameçon* 'hook' has just two further derivatives: one nominal, one verbal. *Helléniste* 'hellenist' is one of the various entries that uses the Greek stem for the word Greek itself. Much like in English, it is rare to see this form used outside of literature. Both of these cases could be accounted for very easily. The patterns *hameç*/[a] and *hellé*/[ɛ] would result in correct judgments and would not produce any false positives. While they are uncommon words, performing two extra regex pattern matches would not be computationally expensive.

[78] `https://cental.uclouvain.be/treetagger/`

[79] See Appendix 7.3.2 for breakdown.

[80] `https://www.linternaute.fr/dictionnaire/fr/abecedaire/h/1/`

[81] `https://www.larousse.fr/`

However, I decided not to correct these misclassifications as there will always be edge cases in GP translation and the 98.07% accuracy is more representative of what we might expect to achieve if the testing were extrapolated to account for all h-initial words.

6 Conclusion

Several recurrent themes have manifested themselves throughout this analysis and the implementation of the Liaison Marker application. Firstly, although closely linked, the hiatus and the liaison are not mutually inclusive. Neither the inverted verbal form nor the auxiliary verb rule defined herewithin avoid a hiatus, yet liaison is obligatory in the former case and is very common in the latter case. In both cases, the liaison was attributed to historical French phonology rather than modern French phonology. The standard modern French surface representation for *peuvent* is [pœv], a CVC sequence. This consonant-final sequence is well set for *enchaînement*, wherein the sequence will become CVCV, considered ideal in French (Tranel, 2000), should the following sequence be vowel-initial. Despite this, historical phonological phenomena cause *peuvent* to manifest itself as [pœvt] before certain vowel-initial sequences. Not only does this create a CVCCV sequence, [vt] is otherwise impermissible as a coda sequence in modern French. These multiple violations of modern French phonology suggest that the historical union between two sequences is the most important factor in liaison realization. This was corroborated by our novel analysis of initialisms, wherein we concluded that modern initialisms do not liaise because they never existed in a version of French where the latent final consonant of the preceding sequence was realized, i.e. there is no historical union between the two sequences. Not one of the four *Les M&Ms* sequences were liaised by the voice actor in a 2018 advertisement.[82] The non-realization of the liaison results in a hiatus. If the liaison were a phonological rather than a historical phenomenon, it would seem logical for the sequences to liaise in order to avoid this hiatus. This is compounded by the fact that *les* is a determiner, one of the few liaisons categorically considered obligatory.

Our analysis of the *h aspiré* presented similar findings. Once more, the non-realization of these liaisons results in a hiatus. It is true that aspiration can sometimes discern a minimal pair. If there were no aspiration, *le hêtre* 'the beech tree' [lə.'ɛtʁ̥] would be pronounced in the same way as *l'être* 'the being' [lɛtʁ̥]. However, French is no stranger to homophones and

[82] `https://www.youtube.com/watch?v=Mptb7A6z2Z0`

I would be hesitant to accept this as the prevailing reason for the survival of the aspiration phenomenon. Indeed, Delattre cites the *h aspiré* as the "most obvious" case of historically-based liaison, stating that "neither style nor union" can explain why these words do not liaise (1955, p. 49). In the case of *les haricots*, a growing cohort of modern French speakers eschew aspiration and liaise. In doing so, they create the 'ideal' CVCV sequence and avoid a hiatus. However, *L'Académie Française* continue to take the position that it is "indisputably incorrect".[83] Aspiration is indeed phonological information, but it is a historical phenomenon based on a word's etymology rather than the phonological characteristics of the following vowel. From open-mid front unrounded [ɛ], e.g. *être*/*hêtre*, to mid-back rounded [õ], e.g. *honorer*/*Hongrois*, French vowels can figure in both *h muet* and *h aspiré* words. The historical nature of aspiration is not as applicable to neologisms. The more formal anglicism *holding* 'holding company' conforms to the aspiration associated with words of Germanic origin. However, certain neologisms of younger generations such as *hipster*, also of Germanic origin, are *h muet* and thus liaise. It would be interesting to conduct a more detailed analysis but it may indicate that these speakers make the judgment with respect to the modern French phonological system rather than the historical etymology-based system.

We dismissed the notion that h-initial words can be dichotomized with respect to etymology and resigned to treat all words as aspirated in order to prevent false positives. Nonetheless, we were able to define a GP translation procedure that allowed the application to make the correct judgment for 98% of h-initial words. Moreover, no *h aspiré* words were liaised, which adheres to the application objective of being very precise in classification, i.e. minimizing false positives.

In the introduction, facultative liaisons were highlighted as an area of particular interest as we sought to devise some process whereby the user could be informed as to the likelihood of a given facultative liaison being realized. Ultimately, the application can only discern inter-category facultativeness. Rule 7 related to 'Common Verbs' while Rule 13 related to 'Pre Invariables'. In his address, President Macron realized each of the seven liaisons attributed to the former but none of the six liaisons attributed to the latter. Application users are informed of the facultativeness of a given category within the information modal.[84] However, the application is not as descriptive with respect to intra-category facultativeness. For example, Rule 9 is unable to say that the liaison in *doivent être* is more likely

[83] `http://www.academie-francaise.fr/questions-de-langue` – discussion (43).
[84] See Figure 15.

to be realized than the liaison in *peuvent être*.[85] The notion of facultativeness was discussed in the introduction and we found liaison realization to be highly context dependent at times.[86] Ultimately there is no process whereby the application can convey that notion to the user. This is something that would require closer examination of such highly contextual liaisons. This data analysis may be conducive to the use of machine learning (ML). Indeed, the application is a binary classifier and classification is a quintessential ML problem. Furthermore, the notion of facultativeness might benefit from a probabilistic interpretation. This is a process wherein the trained ML model classifies a certain data point and then uses known data to determine the probability that the classification is correct.

The application would also benefit from further consideration of the rare facultative (RF) liaison. The application was able to identify 41 of 76 RF liaisons (53.95%). If the goal were to identify every single context in which a liaison can happen, this 54% would mean the application is performing 'better' than most native speakers.[87] However, this is not necessarily the case. Liaison Marker wants to help L2 speakers improve their spoken French and exhibiting significant differences from native speakers may well contravene that very goal. 8 of 35 (22.85%) unmarked RF liaisons were in the context *infinitive -er verb + X*. Delattre found liaison to be "very rare" (1956, p. 51) in this context and had previously stated that liaising with [ʁ] is rare for both "psychological and phonetic" reasons (1955, p. 47). Indeed, I must admit I was previously unaware that it was possible to liaise in this context – I have simply never heard this liaison despite spending a year in France and years of regular consumption of French media, wherein enunciation and eloquence are generally more apparent than in everyday speech. The Frenchman who verified the transcriptions for this analysis stated that "[you can] absolutely liaise after the infinitive ... it is indeed very formal, and increasingly rare" and that "it can even serve to a humoristic effect" – the latter a veritable corroboration of the psychological aspect noted by Delattre.

Indeed, if the application were publicized, I would default the marking to 'obligatory/common'. If the user were a very advanced learner and wished to see exactly where they could liaise, they could then request RF liaisons be marked by simply checking a box. This separation of behaviour would require the application to formalize common/rare intra-rule distinctions. For example, the 'Pre Invariable' rule marked thirty-nine liaisons,

[85]These liaisons were discussed under Section 3.2.1 *3PL Present Tense* and the rates of liaison for these two examples can be found there.

[86]See *pas* under Rule 6 in Section 3.3.1 for an example of this.

[87]I dare say that one in two is not rare.

just eight of which were common (20.51%). The eight reveal very clear patterns. They either form a fixed unit, e.g. *mis/en place* 'put in place' or they prevent a cacophony: *français/et étrangers* 'french and foreign' [fʁɑ̃.s**ɛ.(z)ɛ.e**.tʁɑ̃.ʒe]. They liaising [z] here prevents a sequence of [ɛɛe]. Delattre noted that liaison was more likely when the vowels on either side of the hiatus were of the same timbre (1955, p. 47). This information could be easily encoded – the application knows exactly which phonemes are on each side of the hiatus thanks to GP translation.

Finally, POS tagging was a recurring cast member throughout all three main processes – tokenization, GP translation, and rule application. Although the application is able to classify to a high level of accuracy, POS information would be beneficial. Not only would it allow for the implementation of many further rules, it would also allow the existing rules to be better formalized. The only false positive arose from a naive implementation of pluralization, wherein any s-final or x-final word might liaise with an invariable.[88] It is worth noting this false positive represents just 3.33% of the thirty liaisons marked by this naive implementation. Nonetheless, POS information would better serve the application. Furthermore, several obligatory liaisons went unmarked as a result of non-inclusion in the application's list of known prenominal adjectives. POS information would allow for a simple POS tag check in lieu of the list-based check.

Generally, it is best to avoid list iteration whenever possible. However, the lists in this application are all insignificant in size with respect to computational cost. Each contains fewer than twenty-five elements. Moreover, the application instantiates these lists as type *Set* rather than type *Array*. List iteration, in this case a look-up, is generally much faster for the former datatype.[89] [90] Set also has a worst case performance of $\mathcal{O}(1)$, whereas Array has a worst case performance of $\mathcal{O}(n)$. Although POS tagging might not decrease computational cost in this case, its functional benefits are evident. We referenced TreeTagger earlier and it is possible to access this online via UC Louvain.[91] Almost all POS taggers are reliant on the Stanford Parser.[92] Originally written in Java, host-language wrappers and interfaces exist for many major languages including Ruby.[93] Indeed, I was able to run the POS tagger locally but including the wrapper within the Docker container proved to be very challenging and was made more difficult by clash-

[88] See Rule 13.

[89] `http://seamusabshere.github.io/2014/02/25/ruby-array-include-is-slow/`

[90] `https://www.rubyguides.com/2018/08/ruby-set-class/`

[91] `https://cental.uclouvain.be/treetagger/`

[92] `https://nlp.stanford.edu/software/lex-parser.shtml`

[93] `https://rubygems.org/gems/treetagger-ruby`

ing behaviour between Windows and Docker. Nonetheless, it is possible to implement this POS tagger and as I look towards publicizing this application to provide concrete benefit to L2 students, this will be my first port of call.

7 Appendices

7.1 Source Materials

Le Monde – Théâtre

https://www.lemonde.fr/culture/article/2020/03/27/au-theatre-chez-soi-en-temps-de-confinement_6034656_3246.html

LeParisien – Commerce

http://www.leparisien.fr/economie/consommation/petits-commerces-le-gouvernement-nous-a-oublies-27-03-2020-8288898.php

LeFigaro – Sport

https://sport24.lefigaro.fr/football/etranger/italie/actualites/baisse-des-salaires-qui-suit-l-exemple-de-la-juve-998042

L'express – Mode (Fashion)

https://www.lexpress.fr/styles/mode/comment-american-vintage-transforme-des-robes-en-tapis_2122095.html

Libération – Environnement

https://www.liberation.fr/terre/2020/03/26/du-blanc-au-vert-la-difficile-conversion-des-stations-de-ski_1781549

Macron – Adresse aux Français

https://www.youtube.com/watch?v=MEV6BHQaTnw&

7.2 Grapheme-to-Phoneme Translation

Each of the following tables consists of 74 words. The table is to be read as 'Word | First Grapheme Last Grapheme|' and the words are ordered from top to bottom, left to right. Incorrect translations are marked with an asterisk, then VC/N respect, whereby ++ means 'VC respected, N respected'. VC/N respect figures are calculated with respect to incorrect judgments.

1. Le Monde – Théâtre

Ecouter	E* + +	er	ne	ne	e
le	le	e	manquent	m	quent
Décaméron	D	on	pas,	p	as
de	de	e	venant	ve	ant
Boccace	B	ce	de	de	e
lu	l	u	théâtres	t	res
par	p	r	ou	ou	
des	de	es	de	de	e
comédiennes	c	nes	compagnies	co	ies
et	et		qui	qu	i
des	de	es	permettent	pe	t
comédiens	c	iens	de	de	e
français	f	ais	réfléchir	r	r
et	et		et	e	
étrangers.	é	ers	de	de	e
Plonger	P	er	rêver	r	er
dans	d	ans	en	en	
les	le	es	s'inventant	s'inv	ant
riches	r	ches	un	un	
heures	heure	res	fauteuil	f	euil
du	d	u	de	de	e
Théâtre	T	re	spectateur	s	r
du	d	y	chez	che	ez
Soleil.	S	eil	soi,	s	oi
Rire	R	re	et	et	
avec	a	c	en	en	
Phèdre	Ph	re	laissant	l	ant
racontée	r	ée	l'imagination	l	on
par	p	r	parcourir	p	r
François	F	ois	la	l	a
Gremaud.	G	aud	scène	s	ne
Participer	P	er	grand	g	and
à	à		théâtre	t	re
un	un		du	d	u
voyage	v	ge	monde	m	de
immobile	i	le		**Correct**	**Incorrect**	**Accuracy**
inédit.	iné	it	**Phoneme**	137	1	**99.28%**
Les	le	es	**VC Respect**	1	0	100.00%
propositions	p	ons	**Nasality Respect**	1	0	100.00%

2. LeParisien – Commerce

À	À		Joël	J	l	
côté	c	e	Mauvigney	m	ey	
des	de	es	déplore	d	re	
magasins	m	ins	que	que	e	
de	de	e	le	le	e	
la	l	a	rôle	r	le	
grande	g	nde	de	de	e	
distribution	d	on	ces	ce	es	
quelque	que	que	commerçants	c	ants	
petits	pe	its	ne	ne	e	
commerces	c	rces	soit	s	oit	
de	d	e	pas	p	as	
bouche	b	che	assez	as	ez	
les	le	es	mis	m	is	
boulangeries,	b	ies	en	en		
boucheries,	b	ies	valeur	v	r	
charcuterie,	ch	ie	par	p	r	
fromagers	f	ers	le	le	e	
sont	s	ont	gouvernement.	g	ement	
aussi	au	i	Il	i	l	
autorisés	au	és	est	est		
à	à		positif	p	f	
rester	res	er	dans	d	ans	
ouverts	ou	rts	la	l	a	
pendant	p	ant	mesure	me	re	
cette	ce	te	où	où		
période	p	ode	nos	n	os	
de	de	e	entreprises	en	ises	
confinement.	c	ement	ont	ont		
Le	Le	e	été	é	é	
président	p	t* – –	authorisées	au	ées	
de	de	e	à	à		
la	l	a	poursuivre	p	re	
Confédération	C	on	leur	l	r	
générale	gén	le	activité	a	é	
de	de	e		**Correct**	**Incorrect**	**Accuracy**
l'alimentation	l	on	**Phoneme**	140	1	**99.29%**
de	de	e	**VC Respect**	0	1	0.00%
détail	d	ail	**Nasality Respect**	0	1	0.00%

3. LeFigaro – Sport

A	A		Coronavirus	Co	s	
l'image	l	ge	et	et		
des	de	es	les	le	es	
joueurs	j	rs	conséquences	k	nces	
de	de	e	économiques	é	ques	
la	l	a	désastreuses	d	uses	
Juventus	J	s	qu'elle	qu	le	
ayant	ay	ant	pourrait	p	ait	
proposé	p	é	avoir	a	r	
de	de	e	sur	s	r	
baisser	b	er	leur	le	r	
leur	le	r	club,	c	b	
salaire	s	re	les	le	es	
pour	p	r	joueurs	j	rs	
aider	ai	er	de	de	e	
leurs	le	rs	la	l	a	
clubs,	c	bs	Juventus	J	s	
d'autres	d	res	ont	ont		
ont	ont		décidé	d	é	
également	é	ement	de	de	e	
consenti	co	i	consentir	c	r	
à	à		à	à		
un	un		une	une	ne	
effort	e* + +	rt	baisse	b	sse	
financier.	f	er	drastique	d	que	
Et	Et		de	de	e	
cela	ce	a	leur	l	r	
pourrait	p	ait	salaire	s	re	
continuer.	c	er	durant	d	ant	
Voilà	V	à	les	le	es	
ce	ce	e	quatre	qu	re	
qui	qu	i	prochains	p	ains	
s'appelle	s	lle	mois.	m	ois	
montrer	m	er	Une	Une	ne	
l'exemple.	l	ple	initiative	ini	ve	
Face	F	ce		**Correct**	**Incorrect**	**Accuracy**
à	à		**Phoneme**	138	1	**99.28%**
l'épidémie	l	ie	**VC Respect**	1	0	100%
de	de	e	**Nasality Respect**	1	0	100%

4. L'Express – Mode (Fashion)

American	A	n	marque	m	que		
Vintage	V	ge	de	de	e		
s'appuie	s	ie	mode	m	ode		
sur	s	r	Reportage.	Re	ge		
la	l	a	Aïn	A	n		
coopérative	c	ve	Leuh.	L	euh		
de	de	e	Moyen-Atlas.	M	as* – +		
tissu	t	u	Fès	F	ès – +		
d'Aïn	d	n	est	est			
Leuh,	L	euh	à	à			
au	au		deux	d	eux		
Maroc,	M	oc* – +	heures	heure	res		
pour	p	r	de	de	e		
offrir	o	r	route.	r	te		
une	une	ne	En	En			
seconde	se	de	cette	ce	te		
vie	v	ie	fin	f	in		
à	à		de	de	e		
ses	s	es	matinée	m	ée		
stocks	s	cks	de	de	e		
dormants	d	ants	fin	f	in		
et	et		janvier,	j	er		
transformer	t	er	la	l	a		
des	de	es	bourgade	b	de		
robes	r	bes	marocaine	m	ne		
en	en		semble	s	le		
tapis	t	is	bien	b	ien		
boucherouites.	b	ites	vide.	v	de		
Un	Un		Malgré	M	é		
projet	p	et	la	l	a		
d'upcycling	d	ing	lumière	l	re		
en	en		crue.	c	ue		
phase	ph	ase	Les	L	es		
avec	a	c	cerises	ce	ises		
la	l	a	reines	r	nes		
politique	p	que		**Correct**	**Incorrect**	**Accuracy**	
RSE			**Phoneme**	134	3	**97.81%**	
de	de	e	**VC Respect**	0	3	0.00%	
la	l	a	**Nasality Respect**	3	0	100.00%	

5. Libération – Environnement

A	A		Bourdeau,	B	eau
quoi	qu	oi	professeur	p	r
ressemblera	res	a	à	à	
l'avenir	l	r	l'Institut	l	t* – +
?			d'urbanisme	d	me
A	A		et	et	
une	une	ne	de	de	e
«sanctuarisation	s	on	géographie	géo	ie
du	d	u	alpine	a	ne
ski	s	i	à	à	
alpin	a	in	l'Université	l	é
dans	d	ans	Grenoble-Alpes	G	pes
les	le	es	dans	d	ans
très	t	ès	la	l	a
grandes	g	ndes	revue	re	ue
stations	s	ons	spécialisée	s	ée
capables	c	bles	Urbanisme.	U	me
d'investissements	d	ents	Plombées	P	ées
en	en		par	p	r
communication	c	on	le	le	e
et	et		changement	ch	ement
en	en		climatique,	c	que
infrastructures	in	res	les	le	es
:			autres	au	res
neige	ne	ge	stations	s	ons
de	de	e	seraient	se	aient
culture	c	re	donc	d	c
généralisée,	gén	ée	vouées	v	ées
remontées	re	ées	à	à	
mécaniques,	m	ques	fermer	fe	er
équipements	é	ents	si	s	i
ludiques,	l	ques	elles	e* + +	lles
commerciaux	c	aux	ne	ne	e
et	et		se	se	e
festifs	fes	fs	réinventent	r	t
de	de	e		**Correct**	**Incorrect**	**Accuracy**
prestige»,	p	ge	**Phoneme**	132	2	**98.51%**
écrit	é	it	**VC Respect**	1	1	50.00%
Philippe	Ph	pe	**Nasality Respect**	2	0	100.00%

Macron

Françaises,	f	ises	réalité	r	é
français,	f	ais	immédiate	i	te
mes	me	es	pressante.	p	te
chers	che	ers* − +	Le	le	e
compatriotes,	c	otes	gouvernement	g	ement
Jeudi	j	i	a	a	
soir,	s	r	pris	p	is
je	j	e	comme	c	mme
me	me	e	je	j	e
suis	s	is	vous	v	ous
adressé	a	é	l'avais	l	ais
à	à		annoncé,	a	é
vous	v	ous	des	de	es
pour	p	r	dispositions	d	ons
évoquer	é	er	fermes	f	rmes
la	l	a	pour	p	r
crise	c	ise	freiner	f	er
sanitaire	s	re	la	l	a
que	que	e	propagation	p	on
traverse	t	se	du	d	u
notre	n	re	virus.	v	us
pays.	p	ays* + +	Les	le	es
Jusqu'alors,	j	rs	crèches	c	ches
l'épidemie	l	ie	les	le	es
de	de	e	écoles	é	oles
covid-19	co	19	les	le	es
était	é	ait	collèges,	c	ges
peut-être	pe	re	les	le	es
pour	p	r	lycées,	l	ées
certains	ce	ains	les	l	es
d'entre	d	re	universités,	uni	és
vous	v	ous	sont	s	ont
une	une	ne	fermés	fe	és
idée	i	ée	depuis	de	is
lointaine,	l	ne	ce	ce	e
elle	e*	lle		**Correct**	**Incorrect**	**Accuracy**
est	est		**Phoneme**	136	3	**97.84%**
devenue	de	ue	**VC Respect**	2	1	66.67%
une	une	ne	**Nasality Respect**	3	0	100.00%

7.3 Liaison Marking

Each of the following five subsections is to be read with respect to the following key:
When the text is bold, the application has marked a liaison.
Obligatory
Facultative, common, formal to a degree
Facultative, rare, very formal
Marked by application, but not by either speaker. Requires consideration.
Consecutive liaisons, e.g. x y z, are to be read from left to right. X and Y are green, Y and Z are blue. Despite the green parent layer extending past Y, X and Z are not associated – this is simply a limitation of using colorboxes in LaTeX.

To prevent confusion in sequences where successive liaisons take place, any word in this context that is followed by '||' signifies a judgment of 'no liaison' by the application.

Note that some liaisons are deliberately unmarked. We have discussed that the hyphen acts as a liaison marker (see Section 3.1.1). Therefore *quelques-uns* is unmarked, for example, even though it is obligatory.

1. LeMonde – Théâtre

Ecouter le Décaméron de Boccace lu par des **comédiennes et** des comédiens **français et** étrangers. Plonger dans les riches heures du Théâtre du Soleil. Rire avec Phèdre racontée par François Gremaud. Participer à un voyage immobile inédit... Les propositions ne manquent pas, venant de théâtres ou de compagnies, qui permettent de réfléchir et de rêver, en **s'inventant un** fauteuil de spectateur, chez soi, et en laissant l'imagination parcourir la scène du grand théâtre du monde, de Florence en 1348 à aujourd'hui. Depuis mardi 24 mars, tous les matins, à 8 heures, on peut suivre un feuilleton extraordinaire, imaginé par Sylvain Creuzevault. Parce qu'il aime tirer les fils entre **les arts** et **les époques**, le metteur en scène a choisi le Décaméron, de Boccace, qui, en ces temps de confinement planétaire, renvoie à **un autre** temps de confinement : à Florence, lors de l'épidémie de peste de 1348, sept **dames et** trois chevaliers quittent la ville pour la campagne où ils **vivent au** rythme **des histoires** que chacun raconte, chaque jour. Pour faire entendre **ces histoires**, Sylvain Creuzevault a battu le rappel de **ses amis**, **comédiens et** comédiennes, en France, Allemagne, Autriche, Suisse, Italie... **Ils ont** répondu présent et, chaque jour,

l'un ou l'une d'entre eux lit. Dans le prologue, que l'on peut réécouter en podcast, des voix multiples se mêlent, Jean-Luc Godard, Laurence Chable, Louis Garrel, Frédéric Leidgens, Maya Bösch… puis Nicolas Bouchaud lit l'introduction de la première journée. Suivent, le 25 mars, Julien Gosselin, avec Le Reproche ingénieux, et, le 26 mars, Dominique Valadié, pour Le Mari **jaloux et** cruel. De **grandes et** belles voix, un son et un mixage travaillés au mieux dans le contexte : ce feuilleton **est un** bonheur, appelé à durer cent **jours et** à tresser une couronne de mots à travers l'Europe, et plus loin encore. Vous n'avez pas vu Molière, 1789, Les Naufragés du Fol Espoir, Le Dernier Caravansérail ? Vous pouvez les voir, via le site du Réseau Canopé. **C'est une** belle occasion d'entrer dans la longue et exemplaire histoire de la troupe d'Ariane Mnouchkine, dont les films d'autres créations seront prochainement offerts par ce même site. Ces films s'accompagnent, sur le site du Soleil, de nombreuses archives **visuelles et** sonores, qui parcourent un très large champ, esthétique et politique : la guerre du Vietnam et le génocide cambodgien à travers L'Histoire terrible mais inachevée de Norodom Sihanouk ; la décolonisation du continent indien à travers L'Indiade ; la tragédie antique à travers **Les Atrides** ; la tragédie et l'histoire à travers Shakespeare… pour ne citer que quelques-uns des thèmes abordés par le Théâtre du Soleil, qui a toujours porté un grand soin à la pédagogie et à la transmission. Il le prouve une nouvelle fois, **en offrant || aux élèves** aussi bien qu'aux spectateurs avertis une immersion au cœur de l'art et de la puissance inaltérable du théâtre. A l'origine, le point d'exclamation **était un** point d'admiration. Il retrouve son sens premier dans le titre de Phèdre ! proposé par François Gremaud. Quand le Théâtre Vidy-Lausanne lui a demandé de faire découvrir d'une manière moderne un classique **aux élèves**, cet inclassable artiste suisse, né en 1975, a aussitôt pensé à la tragédie qu'il préfère, et opté pour une pratique **dont il est un as** : la conférence décalée. Créé dans **les écoles**, ce Phèdre ! a été réécrit pour la scène, et présenté en 2019 à Avignon, où il a triomphé. Le Français Romain Daroles fait merveille en conférencier transi d'admiration pour son sujet. Il revendique « une joie de l'étonnement » qui le mène à emprunter tous les chemins, dont celui d'une inénarrable naïveté. Mais cette naïveté n'est qu'apparence. Elle masque une connaissance magnifique de Phèdre, de **ses enjeux**, de sa composition, et de **ses alexandrins**. Ce spectacle, qui s'adresse à **tous et rend un** merveilleux hommage à la tragédie de Racine, on peut le revoir ou le découvrir, du lundi 30 mars au dimanche 5 avril, sur le site du Théâtre Vidy-Lausanne. Puisque le temps **est au** voyage immobile, la Comédie de Valence, dirigée par Marc Lainé, propose à ceux qui en rêvent de s'offrir une échappée. Il leur suffit de s'inscrire à l'adresse notregrandeevasion@comediedevalence.com. Le dessinateur Stephan Zimmerli les contactera, et

leur demandera de répondre à la question : « Si vous pouviez à l'instant précis vous téléporter vers un lieu idéal, réel ou imaginaire, à quoi ressemblerait-il ? » Stephan Zimmerli dessinera ce lieu, en fonction des réponses. Et les dessins, ajoutés les uns aux autres, formeront le Carnet d'un voyage immobile en période de confinement, publié sur les comptes Facebook et Instagram de la Comédie de Valence.

2. LeParisien – Commerce

À côté des magasins de la grande distribution, quelque 402 000 petits commerces de bouche – les boulangeries, boucheries, charcuterie, fromagers, primeurs... — sont aussi autorisés à rester ouverts pendant cette période de confinement. Le président de la Confédération générale de l'alimentation de détail, Joël Mauvigney, déplore que le rôle de ces commerçants ne soit pas assez mis en valeur par le gouvernement. Il est positif dans la mesure où nos entreprises ont été autorisées à poursuivre leur activité et ainsi donner « à manger » aux Français. Évidemment, on doit s'adapter et se réinventer face à l'ampleur de cette épidémie. Certains d'entre nous parviennent à maintenir – plus ou moins – leur chiffre d'affaires : c'est le cas notamment du boulanger, du boucher, du charcutier, du fruit et légumes. Pour d'autres en revanche, la situation est plus compliquée. Les pâtissiers par exemple ne réalisent que 20 % de leurs ventes habituelles. C'est pire encore pour d'autres types de commerces, comme les glaciers et les chocolatiers, qui ont été obligés de fermer. Mais dans ce contexte, ce qui nous met hors de nous, c'est surtout de voir qu'au gouvernement, on nous a oubliés ! Certains ministres ne parlent jamais de nous. Comme Bruno Le Maire qui ne cite que la grande distribution. On dit aux consommateurs : « Aller dans les grandes surfaces, vous y trouverez tout ce qu'il vous faut ». Mais chez nous aussi, vous trouverez tout ce qu'il faut. Et ça, personne ne le dit ! Il n'y en a que pour les supermarchés ! Or, nous aussi, nous maintenons nos activités. Nous aussi, nous avons des salariés valeureux, heureux de répondre présents, qui ont des contacts avec les clients et qui prennent donc des risques. Nous appliquons évidemment toutes les règles sanitaires : le marquage au sol pour qu'il y ait un mètre de distance; nous faisons rentrer nos clients au compte-gouttes. Le lavage des mains faisait déjà partie de notre quotidien, ainsi que les gants. Nous avons aussi cherché à nous procurer du gel. La grande difficulté, ce sont les masques. Mais nos clients, les habitués comme les autres, voient bien ce que nous avons mis en place. Ils sont reconnaissants qu'on soit ouverts. D'autant plus que la différence avec une grande surface, c'est que chez nous, il y a des discussions, un lien. On prend

des nouvelles **des uns** et **des autres**, notamment de notre clientèle âgée. On **les appelle** si on ne les voit pas pendant plusieurs jours, certains d'entre nous se **sont aussi** organisés pour livrer des produits, des repas aux plus démunis. Il y a **un élan** de solidarité. Et **on aimerait** que ce soit souligné. Juste de ne pas **nous oublier**, de parler aussi de notre rôle, **des efforts** consentis. Il y a d'ailleurs beaucoup d'hypocrisie dans les discours quand je vois le PDG de Carrefour ou encore Auchan offrir une prime de 1 000 euros aux caissières, aux gens qui remplissent les rayons. Vous savez pourquoi ils font ça? Parce que ces gens-là qui sont payés au smic **toucheraient autant** s'ils se **mettaient en** retrait. Or, **ils ont** besoin d'eux. Nous, nous n'allons rien faire car nos salariés, eux, ne sont pas payés au smic. C'est toute la différence! Si nous subissons une baisse d'activité, beaucoup d'entre nous vont mettre la main à la poche plutôt que de réduire **les heures** de travail de **nos employés**. Il y aura peut-être du chômage partiel dans nos commerces les **plus impactés**, il y a peut-être quelques droits de retraits. Mais ce sera loin d'être la majorité, croyez-moi.

3. LeFigaro – Sport

A l'image des joueurs de la Juventus ayant proposé de baisser leur salaire pour aider leurs clubs, d'autres clubs **ont également** consenti à **un effort** financier. Et cela pourrait continuer. Voilà ce qui s'appelle montrer l'exemple. Face à l'épidémie de coronavirus, et les conséquences économiques désastreuses qu'elle **pourrait avoir** sur leur club, les joueurs de la Juventus ont décidé de consentir à une baisse drastique de leur salaire durant les quatre prochains mois. Une initiative dont le défenseur Giorgio Chiellini **est à** l'origine, et soutenue par des cadres du vestiaire dont Gianluigi Buffon ou encore Cristiano Ronaldo. Le Portugais renonce ainsi à presque 4 millions d'euros sur la période allant de mars à juin, alors que la Juventus va économiser au total près de 90 millions d'euros. Un beau geste qui pourrait faire **des émules** dans le monde du sport. Car la Juventus n'est pas le seul club susceptible de rencontrer des difficultés financières dans les **mois à** venir. Auparavant, l'Atlético Madrid avait lui aussi par exemple annoncé vendredi que ses joueurs **avaient accepté** de se plier à une baisse de leurs salaires le temps que la crise perdure, afin de soutenir le club. **En Allemagne**, Mönchengladbach, Brême ou encore Schalke 04 **ont eux aussi** **mis en** place un dispositif similaire, alors que les joueurs du Bayern Munich **ont accepté** de revoir à la baisse de 20% **leurs émoluments**. Et en France ? Si rien n'a encore été annoncé, certains joueurs se sont déjà dits **prêts à** baisser leurs revenus, à l'image d'Ander Herrera. Le milieu de terrain du PSG, placé au chômage partiel comme l'ensemble de ses coéquipiers, a con-

fié à la radio Cadena Ser être favorable à une baisse de son salaire «si les choses sont raisonnables». De là à imaginer une décision unanime des joueurs du PSG à l'image de ceux de la Juventus ? Il est encore trop tôt pour le dire. Ce qui est certain en revanche, c'est que Herrera ne réagira pas de la même manière qu'une partie des joueurs de Barcelone si son club lui demande de consentir à un effort financier. Les Blaugranas ont en effet eux aussi annoncé que l'équipe avait accepté une baisse de salaire, mais cela a été au prix de plusieurs journées de négociations avec un vestiaire qui n'était pas emballé par l'idée. Une prise de position qui n'a d'ailleurs pas du tout plu en Espagne. Reste maintenant à savoir ce que vont faire les derniers clubs à n'avoir rien communiqué, à l'image de l'autre grand d'Espagne, le Real Madrid.

4. L'express – Mode (Fashion)

American Vintage s'appuie sur la coopérative de tissu d'Aïn Leuh, au Maroc, pour offrir une seconde vie à ses stocks dormants et transformer des robes en tapis boucherouites. Un projet d'upcycling en phase avec la politique RSE de la marque de mode. Reportage. Aïn Leuh. Moyen-Atlas. Fès est à deux heures de route. En cette fin de matinée de fin janvier, la bourgade marocaine semble bien vide. Malgré la lumière crue. Les cerises, reines de la région, ne sont pas encore de saison. Seuls quelques habitués sont installés au café du coin. Des enfants s'échappent furtivement dans la rue... Tout semble très calme. Derrière la bâtisse, très sobre, de la coopérative des tisseuses de tapis du village, aussi. Si l'heure est au répit – ou plutôt à la préparation du déjeuner – les dernières semaines ne riment pas avec oisiveté. Lhasmia, Khadija et leurs amies – ou cousines, nièces, collègues avec qui elles ont lié des relations bien plus fortes que celles du labeur – sont actives. Parmi les quinze tisseuses, dix sont mobilisées sur la dernière commande d'American Vintage. Au programme : 60 tapis boucherouites confectionnés à partir de 2450 robes. Livraison prévue : le 26 mars. Né dans les années 1960 – historiquement pour isoler les habitations berbères du froid – le tapis "boucherouite" est composé de morceaux de tissus usés noués les uns aux autres. Comme leurs "homologues" d'Azrou, les tisseuses d'Aïn Leuh auraient pu refuser le projet... Car "dans l'esprit des gens, un tapis fait avec des chutes de tissu est un tapis qui n'a pas de valeur", nous confie Abdel-Ilah Neghrassi (appelez-le Abdou), marchand de tapis à Azrou, ville réputée pour son savoir-faire traditionnel en la matière. Pourtant, la technique s'est offert une place de choix en Occident. Au sol ou accroché au mur, telle une oeuvre d'art, le tapis boucherouite séduit designers, chineurs... et autres collectionneurs. D'où l'idée, bienvenue, d'American Vintage

: transformer une partie de son stock dormant ou **défectueux en** collection déco.

5. Libération – Environnement

A quoi ressemblera l'avenir ? A une «sanctuarisation du ski alpin dans les très grandes stations d'altitude capables **d'investissements en** communication et **en infrastructures** : neige de culture généralisée, remontées mécaniques, équipements ludiques, **commerciaux et** festifs de prestige», écrit Philippe Bourdeau, professeur à l'Institut d'urbanisme et de géographie alpine à l'université Grenoble-Alpes, dans la revue spécialisée Urbanisme. Plombées par le changement climatique, **les autres** stations seraient donc **vouées à** fermer si elles ne se réinventent pas durablement. «Dans trente ans, c'est cuit» pour le ski alpin à faible et moyenne altitude, insiste Vincent Vlès, professeur émérite **des universités en aménagement** et urbanisme. «La neige de culture produite grâce aux canons permettra de tenir **quelques années** mais c'est d'une certaine manière un prolongement **dans un** cul-de-sac», poursuit le chercheur CNRS au laboratoire Certop. Pourtant nombre de stations **s'accrochent à** cette solution. Le département de l'Isère estime par exemple que 42 % de la surface de ses domaines skiables seront équipés pour la neige de culture d'ici à 2025, avec l'espoir de «maintenir un niveau **d'enneigement en** 2050 similaire à celui d'aujourd'hui». D'autres **stations investissent** pour grappiller quelques mètres d'altitude. Depuis cet hiver, Valloire (Savoie) s'étend sur une dizaine d'hectares supplémentaires, ce qui a permis de réhausser son point culminant. Un nouveau télésiège a été installé pour accéder aux **pistes en** hauteur. En complément, la station renforce sa capacité à produire de la neige artificielle et complète par du ski de randonnée et d'activités estivales . Ça, c'est pour la diversification. Dans cette veine, quelques stations misent désormais sur le concept «Quatre saisons». C'est le cas de Puigmal, dans les Pyrénées-Orientales. **Après une** faillite en 2013, elle a été reconvertie grâce aux fonds de l'équipementier Rossignol, qui mise sur ce type de nouveaux projets. Fini le ski alpin, les remontées mécaniques **sont à** l'arrêt. Depuis fin 2019, **on y** pratique le VTT, le ski de randonnée, la marche nordique et la randonnée. Toujours dans les Pyrénées, le groupe N'Py, qui gère huit stations, veut prendre la même orientation pour pérenniser **son activité** . Dans l'Isère, Chamrousse veut même devenir une «smart station 4 saisons» à l'horizon 2030 : une restructuration pour ouvrir toute l'année. Elle prévoit la construction d'un centre aquatique, d'une piste de luge d'été et d'une salle **multi-usages en** plus **des activités** VTT,

randonnées et trail. Le village veut reconquérir **des habitants à** l'année, être mieux connecté à la vallée grenobloise et garantir plus d'hébergements pour les touristes. Pour le côté écolo, on vise 100 % d'énergies **renouvelables et** on valorise la proximité du Parc naturel régional de Belledonne. «Même par effet indirect, il y a toujours cette propension à vouloir conquérir, domestiquer, coloniser la nature. Au-delà de l'artificialisation des sols, attention aux déséquilibres **des espaces** naturels que peut générer une trop grande fréquentation», avertit Vincent Vlès. Il doute aussi de la viabilité économique : «Le ski est très pourvoyeur d'emplois, le tourisme 4 saisons l'est moins, du moins sur **les activités** actuelles.» **Ces offres sont en effet** souvent les mêmes d'une station à l'autre. Et dans beaucoup de cas, elles **visent à** compléter l'offre de ski alpin pour rester compétitif plutôt que de basculer **dans un** nouveau modèle adapté au changement climatique. «On n'est pas du tout sur des choix radicaux», confirme Véronique Peyrache-Gadeau, maître de conférences **en économie** territoriale à l'université de Savoie-Mont-Blanc, pour qui la sortie du «tout neige» massif **est encore** loin. «Il faudrait quitter les néons strictement touristiques de la montagne, réfléchir pas **seulement au** public du tourisme saisonnier mais aussi aux populations locales, à la ressource **en eau**, forestière, de l'agriculture de proximité.» Les projets de demain doivent donc être **pensés à** une échelle spatiale et temporelle plus large que celle des stations. Il faut aussi prendre en compte le désintérêt de la jeune génération pour le ski, perçu comme coûteux. C'est plus largement la relation à la montagne qui doit évoluer. Privilégier la contemplation au grand frisson ? Un certain retour du «climatisme» **pourrait en** tout cas s'effectuer avec les fortes chaleurs **attendues en** plaine. Les **personnes en** quête d'air **frais et** de bien-être peuvent réinvestir les **logements inoccupés** des stations. Une façon de valoriser **les infrastructures** déjà existantes.

Macron

This subsection is to be read with respect to the following key:
When the text is bold, the application has marked a liaison.
Obligatory, realized by Macron
Facultative, common, formal to a degree, realized by Macron
Facultative, rare, very formal, not realized by Macron
Consecutive liaisons, e.g. x y z, are to be read from left to right. X and Y are green, Y and Z are yellow. Despite the green parent layer extending past Y, X and Z are not associated – this is simply a limitation of using colorboxes in LaTeX.

Françaises, français, mes chers compatriotes, Jeudi soir, je me **suis addressé** à vous pour évoquer la crise sanitaire que traverse notre pays. Jusqu'alors, l'épidémie de covid-19 était peut-être pour certains d'entre vous une idée lointaine, elle est devenue une réalité immédiate, pressante. Le gouvernement a pris comme je vous l'avais annoncé, des dispositions fermes pour freiner la propagation du virus. Les crèches, **les écoles**, les collèges, les lycées, **les universités**, sont fermés depuis ce jour. Samedi soir, les restaurants, les bars, tous les commerces non **essentiels à** la vie de la nation **ont également** clos leurs portes. Les rassemblements de plus de 100 personnes **ont été** interdits. Jamais la France n'avait dû prendre de telles décisions, évidemment exceptionnelles, évidemment temporaires, en temps de paix. **Elles ont été** prises avec ordre, préparation, sur la base de recommandations scientifiques avec un seul objectif : nous protéger face à la propagation du virus. Dans la journée de jeudi, un consensus scientifique et politique s'est formé pour maintenir le premier tour **des élections** municipales, et j'ai pris avec le Premier ministre la décision de maintenir le scrutin. Hier, dimanche, **les opérations** de vote ont donc pu se tenir. Je veux, ce soir, remercier les services de l'Etat, les maires, l'ensemble des services des mairies, tous ceux qui ont tenu les bureaux de vote et qui ont donc permis l'organisation de ce scrutin. Je veux aussi saluer chaleureusement les françaises et les français qui malgré le contexte se sont rendus **aux urnes**, dans le strict respect des consignes sanitaires, des gestes barrières contre le virus. Je veux aussi ce soir adresser mes félicitations républicaines aux candidats élus au premier tour, environ 30 000 communes sur 35 0000 **ont après** ce premier tour un conseil municipal. Mais dans le même temps, alors même que les personnels soignants, les services de réanimation alertaient sur la gravité de la situation, **nous avons** aussi vu du monde se rassembler dans des parcs, des marchés bondés, des restaurants, des bars, qui n'ont pas respecté la consigne de fermeture, comme si au fond la vie n'avait pas changé. A tous ceux qui, adoptant ces comportements, ont bravé les consignes, je veux dire ce soir très clairement : non seulement vous ne vous protégez pas vous, et l'évolution récente a montré que personne **n'est invulnérable**, y compris les plus jeunes, mais vous ne protégez pas **les autres**, même si vous ne présentez aucun symptôme, vous pouvez transmettre le virus. Même si vous ne présentez aucun symptôme, vous risquez de contaminer **vos amis**, vos parents, vos grands-parents, de mettre en danger la santé de ceux qui vous sont chers. Dans le **Grand Est**, dans les Hauts-de-France, **en Ile-de-France**, nos soignants se battent pour sauver des vies, avec dévouement, avec force, au moment où la situation sanitaire se dégrade fortement, où la pression sur **nos hôpitaux et** nos soignants s'accentue, tout notre

engagement, toute notre énergie, toute notre force, doivent se concentrer sur un seul objectif : ralentir la progression du virus. Je vous le redis avec force ce soir : respectons les gestes barrières, les consignes sanitaires, c'est le seul moyen de protéger les personnes vulnérables, d'avoir moins de concitoyens **infectés et** ainsi de réduire la pression sur les services de réanimation pour qu'ils puissent mieux accueillir et mieux soigner. Sans signe grave, contactons notre médecin traitant, n'appelons le Samu et ne nous **rendons à** l'hôpital qu'en cas de forte fièvre, de difficultés à respirer, sans quoi ils ne pourront faire face à la vague de cas graves qui déjà se profile dans certaines régions.. Faisons preuve au fond d'esprit solidaire et de sens des responsabilités. Chacun d'entre nous doit à tout prix limiter le nombre de personnes avec qui il **est en** contact chaque jour. Les scientifiques le disent, c'est la priorité absolue. C'est pourquoi **après avoir** consulté, écouté **les experts**, le terrain, et en conscience, j'ai décidé de renforcer encore les mesures pour réduire nos **déplacements et** nos contacts au strict nécessaire.

7.3.1 H-initial Words

Each sequence was tested in the context *des + word*, wherein *h muet* words must liaise and *h aspiré* words must not liaise.

	Actual Liaison	**Actual No Liaison**
Predicted Liaison	habitations, habillages, hallucinations, harmonies, hebdomadaires hébergements, hectares, hécatombes hédonistes, héliotropes, hélicoptères hémisphères, heptathlons, herbes, héritages, hétérosexuels, heures, hexagones, hibernations, hilarités, hindous, hippodromes, hispanophones, histoires, historiens, hivernants, homéopathes, homicides, hommages, hommes, homosexuels, honneurs, hôpitaux, horizons, horloges, horoscopes, horreurs, horticulteurs, hospices, hostilités, hôtels, huiles, humains, humeurs, hybrides, hydravions, hydrocarbures, hygiénistes, hygroscopes, hymens, hymnes, hypertendus, hypocrites, hypnotiseurs, hystérectomies	
	Count: 55	**Count: 0**
Predicted No Liaison	hameçons, hellénistes	hâbleurs, haches, hackers, hadrons, haies, haïkus, haines, halls, halos, haltes, hameaus, hampes, handballeur, hanches, harangues, harcèlements, harpons, hasards, hausses, hennissements, héros, hérissons, hersages, heurtoirs, hiatus, hippies, Hittites, hochets, hockeyeuses, holdings, homards, Hongrois, hontes, hoquets, hordes, hottes, houblons, houlettes, houpettes, housses, houx, huards, huches, huitaines, huppes, hurlements, huskys, huttes
	Count: 2	**Count: 48**

References

Abeillé, A., & Godard, D. (1999). La position de l'adjectif épithète en français: le poids des mots. *Recherches linguistiques de Vincennes*(28), 9–32.

Agren, J. (1973). *Etude sur quelques liaisons facultatives dans le français de conversation radiophonique: Fréquences et facteurs*. Uppsala: Acta Universitatis Upsaliensis.

Ashby, W. J. (2001). Un nouveau regard sur la chute du ne en français parlé tourangeau: s' agit-il d'un changement en cours? *Journal of French Language Studies*, *11*(1), 1–22.

Bauer, B. L. (1995). *The emergence and development of svo patterning in latin and french: diachronic and psycholinguistic perspectives*. Oxford University Press.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le cid-corpus of interactional data-annotation et exploitation multimodale de parole conversationnelle. *ATALA*, *49*, 105–134.

Blanche-Benveniste, C. (2002). Structure et exploitation de la conjugaison des verbes en français contemporain. *Le français aujourd'hui*(4), 11–22.

Bonami, O., Boyé, G., Giraudo, H., & Voga, M. (2008). Quels verbes sont réguliers en français? In *Congrès mondial de linguistique française* (p. 141).

Booij, G., & De Jong, D. (1987). The domain of liaison: theories and data. *Linguistics*, *25*(5), 1005–1025.

Boutin, B. A., & Turcsan, G. (2009). La prononciation du français en afrique: la côte d'ivoire. *Journal of French Language Studies*, *21*, 131-152.

Bybee, J. (2001). Frequency effects on french liaison. *Typological studies in language*, *45*, 337–360.

Bybee, J. (2005). La liaison: effets de fréquence et constructions. *Langages*(2), 24–37.

Cissé, I. A. H., et al. (2014). *Développement phonético-phonologique en fulfulde et bambara d'enfants monolingues et bilingues: étude du babillage et des premiers mots*. LOT, Utrecht.

Cochrane, R., Pirahesh, H., & Mattos, N. (1996). Integrating triggers and declarative constraints in SQL database systems. In *Vldb* (Vol. 96, pp. 3–6).

Codd, E. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, *13*(6), 377–387.

Coşciug, A. (2014). De la phonetique française a sa didacticite aux apprenants roumains: Cas de la methode dite progressive. *Glotodidactica*, 21.

Côté, M.-H. (2005). Le statut lexical des consonnes de liaison. *Langages*(2), 66–78.

Delattre, P. (1947). La liaison en français, tendances et classification. *French review*, 148–157.

Delattre, P. (1955). Les facteurs de la liaison facultative en français. *French review*, 42–49.

Delattre, P. (1956). La fréquence des liaisons facultatives en français. *French review*, 48–54.

Dubost, J.-M., & Su, T.-t. (1999). Prosodic differences and similarities between mandarin and french in declarative, interrogative, surprise and doubt expressions. In *Proceedings of the 14th international congress of phonetic sciences* (pp. 1561–1564).

Durand, J., & Lyche, C. (2008). French liaison in the light of corpus data. *Journal of French Language Studies*, *18*(1), 33–66.

Einhorn, E., & Einhorn, E. (1974). *Old french: A concise handbook*. Cambridge University Press.

Evans, C. (2019). *The Syllable and Optimality Theory: An Insight into the viability of Optimality Theory in Speech Synthesis and its effectiveness in modelling the French language.* (Available at `https://github.com/conorevans/essays/blob/master/optimality-theory-and-the-hiatus.pdf`. Accessed on 2020-01-05.)

Fagyal, Z. (2000). Le retour du e final en français parisien: changement phonétique conditionné par la prosodie. In *Actes du xxiie congrès international de linguistique et de philologie romanes: Vivacité et diversité de la variation linguistique* (pp. 151–160). Max Neimeyer.

Gabriel, C., & Meisenburg, T. (2009). Silent onsets? an optimality-theoretic approach to french h aspiré words. *Variation and gradience in phonetics and phonology*, 163–184.

Gedda, M. (2003). Quelques règles de composition typographique. *Kinésithér (annales)*(20-21), 60–62.

Greidanus, T. (2014). *Les constructions verbales en français parlé: étude quantitative et descriptive de la syntaxe des 250 verbes les plus fréquents* (Vol. 243). Walter de Gruyter GmbH & Co KG.

Gross, M. (1993). Les phrases figées en français. *L'information grammaticale*, *59*(1), 36–41.

Hadáček, J. (2014). Detection and recognition of diacritical and punctuation marks in real-world images. *Bachelor thesis. Czech Technical University in Prague*.

Häikiö, T., Bertram, R., & Hyönä, J. (2011). The development of whole-word representations in compound word processing: Evidence from eye fixation patterns of elementary school children. *Applied Psycholinguistics*, *32*(3), 533–551.

Jakobson, R., & Lotz, J. (1949). Notes on the french phonemic pattern. *Word*, *5*(2), 151–158.

Kelly, R. C. (1970). The order of preposed adjectives in french. *The French Review*, *43*(5), 783–794.

Le Goffic, P. (1997). *Les formes conjuguées du verbe français: oral et écrit.* Editions OPHRYS.

Léon, P. R. (1992). *Phonétisme et prononciations du français: avec des travaux pratiques d'application et leurs corrigés.* Nathan.

Levy, T., Silber-Varod, V., & Moyal, A. (2012). The effect of pitch, intensity and pause duration in punctuation detection. In *2012 ieee 27th convention of electrical and electronics engineers in israel* (pp. 1–4).

Makkai, A. (1980). Period of mystery: Or syntax and the semantic pause. *Rice Institute Pamphlet-Rice University Studies*, *66*(2).

Malécot, A. (1975). French liaison as a function of grammatical, phonetic and paralinguistic variables. *Phonetica*, *32*(3), 161–179.

Mallet, G. (2008). La liaison en français: descriptions et analyses dans le corpus pfc. *Unpublished PhD dissertation. Université Paris Ouest, France*.

Mathieu-Colas, M. (1995). Un dictionnaire électronique des mots à trait d'union. *Langue française*, 76–85.

Moisset, C. (2000). Variable liaison in parisian french. (Dissertations available from ProQuest. AAI9965531. `https://repository.upenn.edu/dissertations/AAI9965531`)

Morin, Y.-C. (1986). On the morphologization of word-final consonant deletion in french. *Sandhi phenomena in the languages of Europe*, *33*, 167.

Nedecky, J. (2011). *French diction for singers: A handbook of pronunciation for french opera and mélodie.* Self published.

Nováková, S. (2012). La production et la perception du schwa (e caduc) en français et en tchèque. étude comparée et applications pédagogiques.

Pierret, J.-M. (1994). *Phonétique historique du français et notions de phonétique générale* (Vol. 19). Peeters Publishers.

Post, B. (2000). Pitch accents, liaison and the phonological phrase in french. *Probus*, *12*(1), 127–164.

Quémart, P., & Casalis, S. (2017). Morphology and spelling in french students with dyslexia: the case of silent final letters. *Annals of dyslexia*, *67*(1), 85–98.

Ranson, D. (2008). La liaison variable dans un corpus du français méridional: L importance relative de la fonction grammaticale. In *Congrès mondial de linguistique française* (p. 150).

Ranson, D., & Passarello, M. (2012). L'élision variable du schwa en fin de mot chez des hommes méridionaux: L'effet des consonnes environnantes et de la fréquence de la lexie. In *Shs web of conferences* (Vol. 1, pp. 1519–1535).

Rey, A., Ziegler, J. C., & Jacobs, A. M. (2000). Graphemes are perceptual reading units. *Cognition*, *75*(1), B1–B12.

Ridouane, R. (2008). Syllables without vowels: phonetic and phonological evidence from Tashlhiyt Berber. *Phonology*, *25*(2), 321–359.

Saunders, G. (1988). The structure of errors in the perception of French speech. *Revue de Phonétique Appliquée*(86), 43–99.

Scheer, T., Wauquie, S., & Encrevé, P. (2015). Autosegmental news from h aspiré and liaison without enchaînement. *13èmes Rencontres du Réseau Français de Phonologie (RFP), Bordeaux Selkirk, EO (1974). French Liaison and the X'Notation. Linguistic Inquiry*, *5*, 573–590.

Selkirk, E. O., & Vergnaud, J.-R. (1973). How abstract is french phonology? *Foundations of Language*, *10*(2), 249–254.

Serrano, F., & Defior, S. (2008). Dyslexia speed problems in a transparent orthography. *Annals of dyslexia*, *58*(1), 81.

Smolensky, P., & Prince, A. (1993). Optimality theory: Constraint interaction in generative grammar. *Optimality Theory in phonology*, 3.

Sturm, J. L. (2012). Liaison in l2 french: The effects of instruction. In *Proceedings from the 4th pronunciation in second language learning and teaching conference* (pp. 157–166).

Thitothati, K. (2013). Syllable pronunciation by adding consonants, error analysis and pronunciation development of 1st year students. *MANUTSAT PARITAT: Journal of Humanities*, *30*(2).

Thomas, A. (1916). Une tentative de réforme de l'orthographe française sous philippe le bel. *Journal des Savants*, *14*(11), 508–513.

Thomas, A. (2002). La variation phonétique en français langue seconde au niveau universitaire avancé. *Acquisition et interaction en langue étrangère*(17), 101–121.

Trager, G. L. (1944). The verb morphology of spoken french. *Language*, 131–141.

Tran, T., Trancart, M., & Servent, D. (2008). Littéracie, sms et troubles spécifiques du langage écrit. In *Congrès mondial de linguistique française* (p. 168).

Tranel, B. (2000). Aspects de la phonologie du français et la théorie de l'optimalité. *Langue française*(126), 39–72.

Tranel, B., & Bernard, T. (1987). *The sounds of french: An introduction.* Cambridge university press.

Tseng, J. (2008). L inversion pronominale: histoire et analyse. In *Congrès mondial de linguistique française* (Vol. 226, p. 2629-2644).

Vinet, M.-T. (1991). French non-verbal exclamative constructions. *Probus*, *3*(1), 77–100.

Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in nlp. In *Coling 1992 volume 4: The 15th international conference on computational linguistics.*

Weisler, S., & Milekic, S. P. (2000). *Theory of language.* MIT Press.

Wilde, S. J. (1986). *An analysis of the development of spelling and punctuation in selected third and fourth grade children (orthography, papago, o'odham).* The University of Arizona.

Williams, L. (2009). Sociolinguistic variation in french computer-mediated communication: A variable rule analysis of the negative particle ne. *International journal of corpus linguistics*, *14*(4), 467–491.