Linguistic Resources for Corpus Processing: the ATISHS project

Max Silberztein

Title:

Linguistic Resources for Corpus Processing: the ATISHS project

Author(s):

Max Silberztein

Université de Franche-Comté

Besançon

France

Declarations:

Funding: ATISH Project, funded by the Région de Franche-Comté, France

Conflicts of interest/Competing interests: None

Availability of data and material: free linguistic resources

Code availability: ATISHS free software; NooJ is free and open source

Authors' contributions: Author of the ATISHS and NooJ software

Abstract:

Nowadays, most corpus processor software applications analyze texts as if they were constituted by sequences of graphical wordforms. However, in users of these software who are working in the social sciences and the humanities are looking for information in the corpora they study expressed by linguistic units of meaning. It is therefore necessary to establish a link between the graphical wordforms that constitute the text files and the units of meaning that are of interest to the users of these software. We show that creating this link requires precise linguistic methods and resources. We present the linguistic resources developed with the NooJ linguistic platform and demonstrate how they are used by the new ATISHS corpus processor to provide more reliable statistical analyses for the digital humanities.

Linguistic Resources for Corpus Processing: the ATISHS project[1]

Max Silberztein

Université de Franche-Comté

Abstract: Nowadays, most corpus processor software applications used in the Digital Humanities analyze texts as if they were constituted by sequences of graphical wordforms. However, users of these software who work in the social sciences and the humanities are looking for information expressed by units of meaning. It is therefore necessary to establish a link between the graphical wordforms that constitute the text files and the units of meaning that are of interest to the users of these software. We show that creating this link requires precise linguistic methods and resources. We present the linguistic resources developed with the NooJ linguistic platform and demonstrate how they are used by the new ATISHS corpus processor to provide more reliable statistical analyses for the digital humanities.

Keywords: Corpus Processing, Digital Humanities, Linguistic resources.

1 Introduction

Nowadays, most researchers in the humanities and the social sciences explore and study their corpora with corpus processor software applications such as Alceste[2], Hyperbase[3], IRaMuTeQ[4], Lexico[5], Sketch Engine[6], TXM[7], Word2Vec[8]. All these software applications process text files as sequences of graphical wordforms, on the assumption that these graphical wordforms represent the pieces of information useful to their users:[9] their concordances display wordforms in context; their search engines compile indices of wordforms in order to answer their users' queries (themselves segmented as sequences of wordforms); the word clouds they provide display wordforms; their statistical analyses detect interesting frequencies of wordforms (or collocations of wordforms), etc.

However, wordforms almost never correspond to the units of meaning — concepts, entities, predicates and relations — that are sought by the researchers in social sciences and the humanities to study their corpus.

1 This research was supported by the project "Analyseur de Textes Innovant pour les Sciences Humaines et Sociales" [A new Text Analyzer for the Social Sciences and the Humanities] from the Region of Franche-Comté, see: http://www.nooj-association.org/atishs.html.

2 Cf. (Reinert, 1999).

3 Cf. (Brunet, 2011).

4 Cf. (Loubère et alii, 2014).

5 Cf. (Laballe et alii, 2002).

6 Cf. (Kilgarriff et alii, 2014).

7 Cf. (Heiden, 2010).

8 Cf. (Church, 2017).

9 By (graphical) wordforms, we mean contiguous sequences of letter characters delimited by non-letters (a.k.a. delimiters). For instance, "tomorrow" and "stepfather" are two wordforms, whereas the similar sequences "next day" and "father-in-law" contain respectively two and three wordforms.

In the NooJ framework,[10] units of meanings are elements of a language's vocabulary and are named "Atomic Linguistic Units" (ALUs).[11] By definition, ALUs are *atomic*, *i.e.* not to be analyzed: locutors have learned them by heart, learners of second-language cannot invent them, and linguistic computer software can only find their properties by consulting dictionaries.

Although linguists have not always agreed on how to characterize ALUs, applications such as Machine Translation or Second-Language teaching make it easier to understand what vocabulary elements are. For example, no English speaker would try to guess how to say "tank top" in French by translating "tank" [réservoir] and "top" [dessus]: *tank top* must be translated as one single French noun (*débardeur*): *tank top* is an ALU. Reciprocally, no French person should try to translate "pomme de terre" [potato] in English by translating "pomme" [apple], "de" [of] and "terre" [earth]: *pomme de terre* is an ALU, and MT software should directly link the sequence "pomme de terre" to "potato", by just looking up a bilingual dictionary.

Conversely, people who learn French do not need to learn the meaning of the verbs "redormir", "remanger", "recréer", etc. They just need to learn the base verbs *dormir* [to sleep], *manger* [to eat] and *créer* [to create], and know that the prefix "re-" can be added in front of any French verb to express repetition: "re-" is a French ALU.

Therefore, there has to be an intermediary step during which corpus processors link the graphical wordforms that occur in the text files to the ALUs that are actually of interest to their users. In the following, we show why linguistic methods and resources are needed, and how they are used in the ATISHS software.

2 Spelling

Obviously, graphical wordforms are defined by their spelling. However, spelling is not always fixed. For example, in the Meditext XIVth century corpus of French texts, the term *seigneur* [lord] has over fifty spelling variants. In order to offer historians and medievalists the means to explore and analyze their corpus, a corpus processor must therefore identify the term *seigneur* with all its spelling variants. In the PALM exploration software,[12] this mapping is performed by using the following NooJ grammar.

10 In the following, we will use the NooJ notation to display dictionary, grammars and grammar rules, see (Silberztein 2016). NooJ is a free and open-source linguistic development environment used to formalize linguistic phenomena (from orthography to semantics) for over 30 languages, as a corpus processor used in the Digital Humanities, as well as as a linguistic engine to develop NLP software applications. To download NooJ, its manual, its linguistic resources as well as access tutorials and references, see: http://www.nooj-association.org.

11 In NooJ's linguistic engine, ALUs are represented as annotations such as: <eaten, eat, VERB +Transitive +N0Anim +N1Conc +Pastparticiple>, where graphical wordforms (e.g. "eaten") are associated with a lexical entry (e.g. "eat"), a category (e.g. "VERB") and a list of properties (e.g. +Transitive, +N0Anim, +N1Conc and +Pastparticiple).

12 Cf. (Aouini 2018).

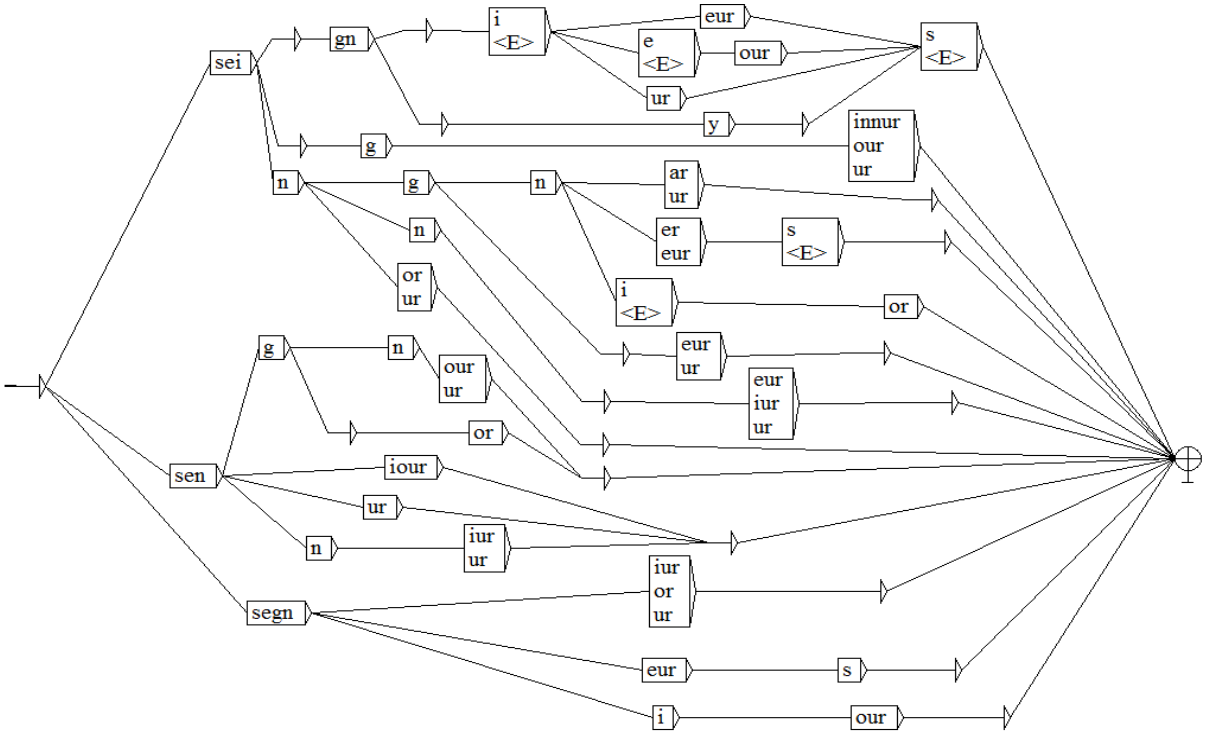

Figure 1: Grammar that represents all spelling variants of the term *seigneur* [lord]

The grammar in figure 1 recognizes any wordform that can be spelled by following a path that starts at the initial node (the horizontal arrow on the left of the graph), follows any connection, and ultimately reaches the terminal node (the target on the right of the graph). Using this grammar, the PALM corpus processor finds all the occurrences of the term *seigneur* [lord] in the corpus — whatever their spelling — and can therefore produce reliable statistical analyses of this term: characterize its frequency in sub-corpora, find interesting collocations with other terms such as *vassal*, *liege*, etc., study the evolution of its use in different regions, etc.

Today, spelling is not much more reliable in the texts typically used by researchers in the social sciences: tweets, blogs, text messages, emails etc. But even in better spelled texts such as newspaper articles or works of literature, spelling variants are very frequent.[13] There is systematic hesitation in the spelling of words of foreign origin (e.g. *yoghurt* vs. *yogurt*; *oenology* vs. *enology*), in the use of space and hyphenation (e.g. *audiovisual, audio-visual, audio visual*), capitalization (e.g., Middle Ages vs. middle ages),[14] mute letters (e.g., *judgement* vs. *judgment*), double letters (e.g. *traveled* vs. *travelled*), word endings (e.g. *program* vs *programme*)[15], regional spelling variants (e.g., *harbour* vs. *harbor*), etc. Acronyms, synonyms and abbreviations add further complexity to the mere recognition of terms and proper names (e.g. *US Republican Party vs. Grand Old Party* vs. *GOP*, the Conservative Party), etc.

13 (Mathieu-Colas 1990) demonstrated that over 5% of the spelling of the entries of the three French reference dictionaries (Grand Dictionnaire Encyclopédique Larousse, Grand Robert and Trésor de la Langue Française) were inconsistent.

14 In order to fix the problem of massive case variation, most computer software are case-insensitive. However, ignoring case makes these software produce a number of false positives, such as finding March (the month) when looking for the verb to march, Polish (from Poland) for the verb to polish, etc. More systematically, corpus processors that use case-insensitive tokenizers cannot distinguish acronyms from proper names and common nouns, e.g. RAM (Random Access Memory) vs. Ram (variant of Rāma) vs. a ram (the sheep).

15 In French, plural endings (-s) are mute. As a consequence, there is a systematic variation between singular and plural components for tens of thousands of terms, such as a flight simulator (un simulateur de vol or un simulateur de vols?), a citrus juicer (un presse-agrume or un presse-agrumes?), etc.

By ignoring variants of a term they process (find, count, analyze, compare etc.), corpus processors systematically underestimate its frequency and thus its relevance in texts. For example, a Google search only finds 70 million occurrences for the query "audio visual," whereas it finds over 700 million occurrences for the query "audiovisual." With only 10% of the results found, we are far from the 95% recall rate typically claimed by most Natural Language Processing applications used in the Digital Humanities.

To offer users of corpora truly reliable analyses, a corpus processor must therefore take terms' variations into account. Variations can be described either by grammars such as the one in figure 1 or by dictionaries. For example, figure 2 shows a sample of a NooJ dictionary that describes the term *tsar* and its spelling variants:

```
tsar,N
csar,tsar,N
czar,tsar,N
tzar,tsar,N
```

Figure 2: The lexical entry *tsar* and its three variants

This dictionary formalizes the fact that the wordforms "csar", "czar" and "tzar" are variants of the lexical entry "tsar." By accessing it, a corpus processor can collect all variants of the term *tsar* before counting its occurrences: it will thus get a much better recall rate than a corpus processor that would only count the graphical wordform *tsar*.

3 Inflectional Morphology

A few software tools used in the Digital Humanities do not perform any morphological analysis of wordforms before performing statistical computations on their frequencies: they treat wordforms such as the following ones as independent units: *march, marches, marching, marched*. The authors of these software justify this policy. For example, for Lexico, (Lamalle et al., 2002) note that in left-wing texts of the 1970s, one uses the plural term *libertés* in the context of defending social rights to housing, education, etc., whereas the singular term *liberté* occurs in right-wing texts to designate the concept of entrepreneurial freedom. These are two different concepts, thus their occurrences should not be unified nor their frequencies aggregated.

This argument is relevant: there are indeed abstract nouns that do not take a plural form in a semantically transparent way. However, over 95% of abstract, concrete, animated and human nouns can be put in the plural with no difference in meaning other than their number:

(Abstract noun): *One demonstration ⟶ two demonstrations*
(Concrete noun): *One table ⟶ two tables*
(Animated noun): *One giraffe ⟶ two giraffes*
(Human noun): *One baker ⟶ two bakers*

Treating occurrences of the two wordforms "demonstration" and "demonstrations" as unrelated units, for instance, would be problematic for a historian or a sociologist who seeks to evaluate the number of conflicts in a

country by exploring a newspaper archive. By not adding the frequencies of these two wordforms, a statistical analyzer would underestimate the overall frequency of this term, hence it might miss its importance and relevance.

A better solution is to lemmatize wordforms that can be put in the plural in a transparent way, but still treat as independent entries those whose meaning is different in the singular and in the plural form. That can be implemented in the platform NooJ by a dictionary such as the following:

```
liberty,N+Singular
liberties,N+Plural
demonstration,N+FLX=TABLE

TABLE = <E>/Singular | s/Plural;
```

Figure 3: Three lexical entries and one inflectional paradigm

The dictionary in figure 3 formalizes the fact that *liberty* (= freedom) and *liberties* (= breach of social convention) are two independent terms, whereas the term *demonstration* is associated with two wordforms, generated by the inflectional paradigm rule TABLE (*i.e.* singular, takes an "s" in the plural).

A more general argument used to not lemmatize wordforms is that the process of lemmatization intrinsically loses information that might be relevant. For example, for a literary study, a psychological study or a semantic analysis, the difference in meaning carried by the wordforms *ate* and *eaten* could be significant, as these two wordforms have different narrative and logical values: *Joe has eaten already* implies that he has finished eating (he is probably not hungry anymore), which cannot be inferred by *Joe ate an apple yesterday*. However, there are also crucial narrative and logical differences in the three following sentences:

Eat your soup! I only eat in kosher restaurants. Next week, we'll eat in an Italian restaurant.

In the first sentence, the wordform *eat* represents an order, to be followed in the near future; in the second, it represents a past and present habit of the narrator; in the third, it represents an activity that will happen in the future. In these three sentences, the same graphical wordform *eat* has three different narrative values: it does not make more sense to aggregate them together while separating the wordforms *ate* and *eaten.*[16] Distinguishing the wordforms *eaten* and *ate* while aggregating the different values of the wordform *eat* is therefore not consistent.

A more consistent approach would be to process not the graphical wordforms, but the ALUs themselves: <ate=eat,VERB+Preterit>, <eaten=eat,VERB+Pastparticiple>, <eat=eat,VERB+Infinitive>, <eat=eat,VERB+Imperative> and <eat=eat,VERB+Present>. That way, the user can choose to perform statistical analyses on the graphical wordforms (before the equal sign) or on the whole ALU, depending on their needs.

4 Stochastic vs. Linguistic Lemmatization

16 Reciprocally, note that the two following sentences are synonymous: "Have you eaten yet?" = "Did you eat yet?".

Most software tools used in the digital humanities do offer some lemmatization functionalities, provided by stochastic lemmatizers such as *Gate*,[17] *NLTK*[18] or *TreeTagger*.[19] Stochastic lemmatizers tag the user's text by comparing the contexts of its wordforms with the ones found in a reference training corpus.[20]

Unfortunately, reference corpora contain a large number of mistakes. For example, (Taylor 2003) shows that the Penn Treebank contains many mistakes such as "Battle-tested" and "Japanese" tagged as proper names; (Silberztein 2018) shows that the Open American National Corpus (OANC), tagged with *Annie*,[21] contains over 11,019 impossible tags such as *abbreviate* tagged as a noun, *about* as an adjective, *anomaly* as an adverb, etc. In the COCA corpus, many typos such as "accrossthe" are tagged as nouns, the wordform "that" is incorrectly tagged as a conjunction in "The one that is most concerned," as a determiner in "My contention, broadly stated, is that we ought to concern ourselves...", as an adverb in "Ellen Ashdown pointed out that many humanities...", etc. [22] We should not expect corpus processing software that use unreliable references to produce reliable results.

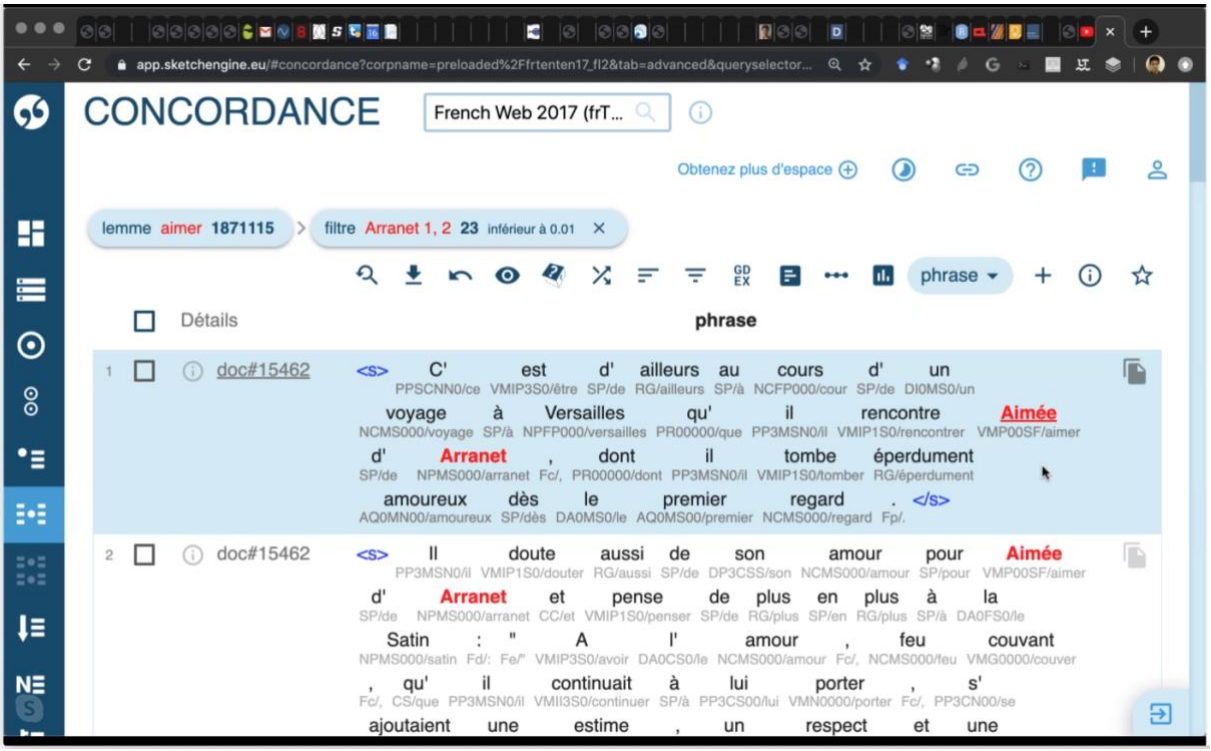

Figure 5: Tagging mistakes

For instance, here is a French sentence[23] tagged by Sketch Engine:

17 Cf. (Cunningham, 2002).
18 Cf. (Loper et alii, 2002).
19 Cf. (Schmid 1994).
20 Such as, for English: the Penn Treebank, the Corpus of Contemporary American English (COCA), the Open American National Corpus (OANC), etc.
21 Annie is a Gate plugin.
22 In figure 4, I have replaced Penn Treebank's tag codes (e.g. "NNP") with a more readable form (e.g. "SingularProperName"). The fact that these training corpora contain large number of mistakes is well known and has been the subject of many publications, e.g. (Dickinson 2012), (Green & Manning 2010), (Volokh & Neumann 2011), etc.
23 From the fictional character Nicolas Le Floch's biography, a series of detective novels that take place in the XVIIIth century, by Jean-François Parot (1946-2018).

C'est d'ailleurs au cours d'un voyage à Versailles qu'il rencontre Aimée d'Arranet, dont il tombe éperdument amoureux dès le premier regard.

[By the way, it is during a trip to Versailles that he meets Aimée d'Arranet whom he instantly falls madly in love with].

In this sentence, *cours* has been incorrectly tagged as "NCFP000/cour", i.e. a plural form of the feminine noun *cour* [courtyard], *Aimée* as "VMP00SF/aimer", i.e. a past participle form of the verb *aimer* [loved], *rencontre* as "VMIP1S0/rencontrer", i.e. a first person singular form of the verb *rencontrer*, *tombe* as "VMIP1S0/tomber", i.e. a first person singular form of the verb *tomber*, *doute* as "VMIP1S0+/douter", i.e. a first person singular form of the verb *douter* and *pense* as "VMIP1S0+/penser", i.e. a first person singular form of the verb *penser*. Moreover, the tagged text does not represent the fact that "d'ailleurs" [by the way] is an adverb rather than a sequence of a preposition and a locative adverb, "au cours de" [during] a preposition (no relation whatsoever with *courtyards*), "tomber amoureux" [fall in love] a frozen expression and "dès le premier regard" [instantly] an adverb. Producing 6 incorrect tags and ignoring 4 multiword units in a 26-wordforms-long sentence cannot be good enough for any NLP application, especially in the digital humanities. For instance, if a linguist wants to use a corpus to study locative nouns, they will get many false positives such as *cour* [courtyard], that will produce higher frequencies and incorrect collocations. If a researcher in literature performs a stylistic analysis of this text, they will grossly underestimate the number of an author' use of concessive adverbs such as *d'ailleurs* [by the way]. If they perform a narrative analysis of the text, they will greatly overestimate the author's use of verbs conjugated in the first person, etc.

For instance, accessing a dictionary that contains the multiword unit "au cours de = PREPOSITION" would prevent the software from producing the mistake *cours* = Noun, Feminine Plural. Accessing local grammars such as the following: would prevent the software from producing incorrect analyses for the four words *rencontre*, *tombe*, *doute* and *pense*:

il <V>/<V+3+s>

(if the wordform "il" is followed by a verb, then this very is conjugated in the third person singular).

The very principle of using training corpora to analyze texts, while standard today's in most NLP applications, is questionable:[24]

— Natural languages contain an infinite set of potential sentences. Therefore, however large a training corpus is,[25] it will never be big enough to represent the infinite number of sentences one might encounter when analyzing texts as varied as the ones used by historians (e.g. XIVth century corpus), psychologists (e.g. children interviews), researchers in literature studies (e.g. novels, theater plays, poems), sociologists (e.g. polls), political scientists (e.g. political debates), etc.;

— Analyzers that use training corpora to analyze a text compare the context of the wordforms in the text with the contexts of the same wordforms in the training corpus. These contexts are typically defined as sequences of a

[24] (Silberztein 2016) lists ten shortcomings of stochastic approaches to Natural Language Processing.

[25] Several projects aim at constructing larger and larger training corpora, such as LDC's English gigaword dataset, cf. https://catalog.ldc.upenn.edu/LDC2003T05 and https://deepai.org/dataset/gigaword (4 million articles).

few wordforms. However, texts in natural languages do not have a linear structure: most wordforms may occur before of after any sequence of word categories. For instance, the wordform “sees” might be followed by an adjective (e.g. *Joe sees pretty flowers*), a determiner (e.g. *Jane sees some flowers*), a noun (*He sees flowers*), a conjunction (e.g. *Nobody sees but you*), a preposition (e.g. *She sees behind herself*), a pronoun (*Everyone sees it*), an adverb (e.g. *She sees very well*). It is thus not surprising that analyzers that use wordforms’ context produce mistakes.

A linguistic approach would constitute a more reliable approach: carefully construct handcrafted dictionaries so that they do not contain incorrect tags such as “accrossthe = Noun”, label each lexical entry to avoid incorrect analyses such as “anomaly = Adverb,” do not ignore multiword units such as “a boatload of = Determiner,” and describe the inflectional paradigm of each lexical entry properly:

```
abbreviate,VERB+FLX=LIVE+Transitive
about,PREPOSITION
anomaly,NOUN+FLX=ABILITY+Abstract
a boatload of,DETERMINER
eat,VERB+FLX=EAT+Transitive
many,DETERMINER+Plural

LIVE = <E>/INF | s/3+s | <B>ing/Gerund | d/Preterit | d/PastParticiple
ABILITY = <E>/Singular | <B>ies/Plural
EAT = <E>/INF | s/3+s | ing/Gerund | <B3>ate/Preterit | en/PastParticiple
```

Figure 6: Lexical entries and inflectional paradigms in an error-free NooJ dictionary

By looking up a dictionary that is complete and does not contain any mistake, a corpus processor would be able to tag ALUs with a perfect recall rate and a perfect precision:
— the recall rate would be perfect, as if a wordform were to be missing its proper analysis, a lexicographer could just add the corresponding ALU it to the dictionary instantly;
— the precision would be perfect, as if an incorrect analysis were to be found, a lexicographer could correct the dictionary instantly.[26]

Such a dictionary should contain all the elements of the standard vocabulary, which for English amounts to less than 100,000 simple words and 250,000 multiword units.[27] Proper names could be recognized by accessing specialized dictionaries such as JRC-Names.[28] Technical and specialized terms could be recognized by accessing dictionaries adapted to the corpus domain such as Medline for medical texts.[29]

5 Derivation

As in the aforementioned case study of a political scientist who seeks to evaluate the state of social peace by analyzing archives of a daily newspaper, simply lemmatizing the wordform *demonstration* would not be enough:

26 One must not confuse tagging mistakes (such as tagging “anomaly” as an adverb) and ambiguities (involved when a wordform corresponds to multiple potential tags). We discuss how to represent homography and solve wordform ambiguities in section 7.
27 Cf. the size of the DELAS and the DELAC dictionaries.
28 See https://ec.europa.eu/jrc/en/language-technologies/jrc-names.
29 See https://www.nibib.nih.gov/content/medical-dictionary-medlineplus.

the corpus processor should take into account all its derived forms to find mentions of this term in the following sentences:

... ***Demonstrators*** *shouted against police brutality and racism and some held signs against discrimination …*
https://www.aa.com.tr/en/europe/demonstrators-in-france-protest-george-floyd-s-death/1867169

... *Wearing face masks, waving black flags and keeping two yards apart, thousands of Israelis* ***demonstrated*** *against prime minister Benjamin Netanyahu ...*
https://www.reuters.com/article/us-health-coronavirus-israel-protest/anti-netanyahu-rally-draws-thousands-under-coronavirus-curbs-idUSKBN2210S4

... *Parents to* ***redemonstrate*** *against air pollution, demanding officials to take immediate and efficient actions and for the well-being of children...*
https://mongolia.gogo.mn/r/156657

Morphological families can be described either using dictionaries associated with morphological rules such as the ones in figures 3 and 6, or by grammars such as the following:

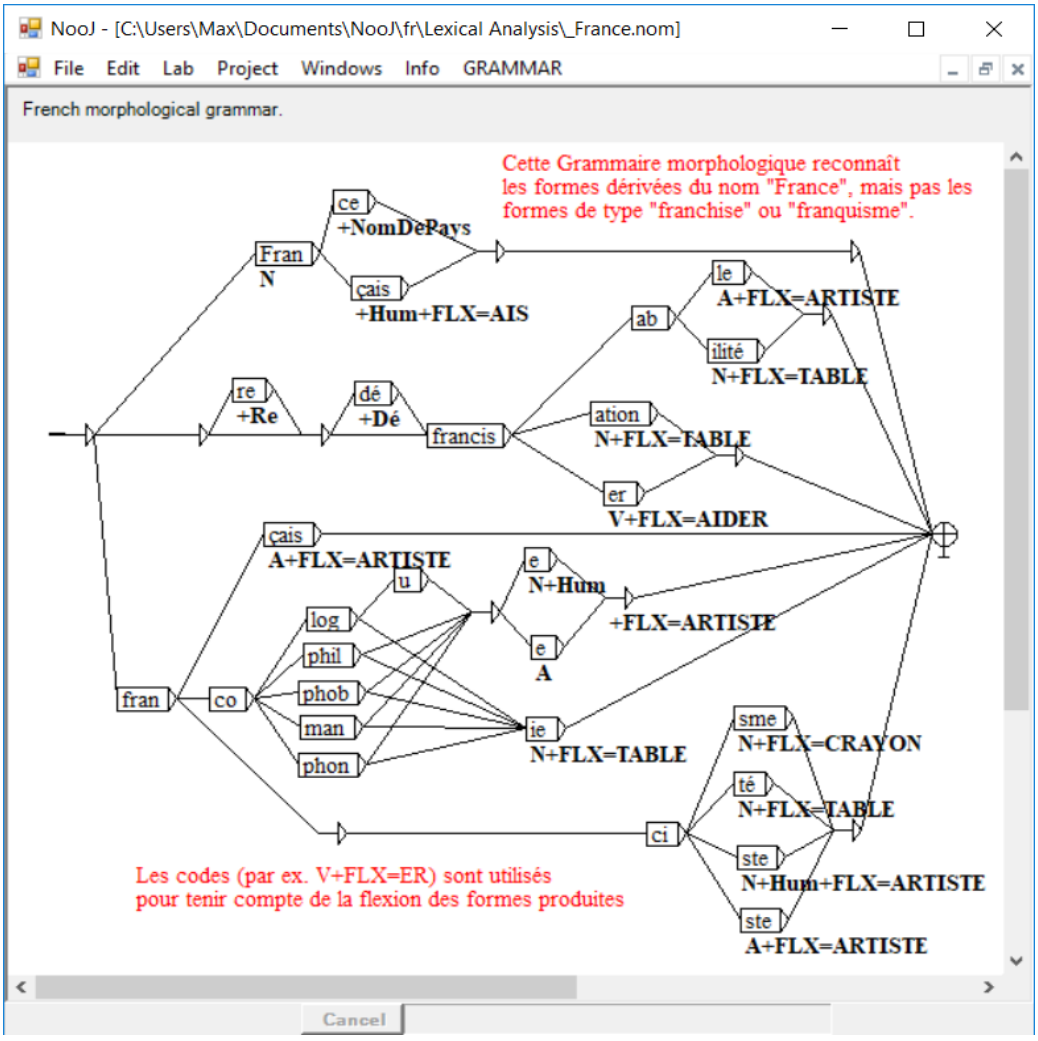

Figure 7: A NooJ grammar that represents the set of wordforms derived from the ALU *France*

The grammar in figure 7 recognizes the wordforms *Français* [Frenchman] *franciser* [to frenchify], *francophone*, *francophobe* [Francophobic], etc., as well as all their inflected forms (e.g. *Françaises*, *franciserons*), which amount to more than three hundred wordforms.

Corpus processors that do not have access to this description will underestimate the frequency of France-related topics in a corpus. For example, in the archives of the newspaper *Le Monde*, 2006, the wordform *France* occurs 38,000 times, whereas its derived forms occur more than 60,000 times.

Users of a corpus processor that does have access to this grammar (or equivalent) will be able to analyze the importance of France's cultural or political impact in newspapers such as *Jeune Afrique* or *Le matin d'Algérie* and analyze its use in relation to other concepts, detect whether it occurs in positive or negative sentences, find collocations with interesting terms, find out if it is more present in left-wing or right-wing political articles, characterize the speeches of certain political figures, etc.

Some corpus processors use root-based orthographical parsers to overcome the fact that they do not access any linguistic description. For example, Alceste[30] and Hyperbase[31] use orthographic rules such as "am.*" to compute derived wordforms of *amour* [a love]. Unfortunately, a consequence of this approach is that *amour* gets linked to *amphithéâtre* [amphitheater], *ample*, *amplifier* [to amplify], *amuser* [to amuse], etc. and not to any of the conjugated forms of *aimer* [to love].

30 Cf. (Reinert 1999).
31 Cf. (Brunet, 2011).

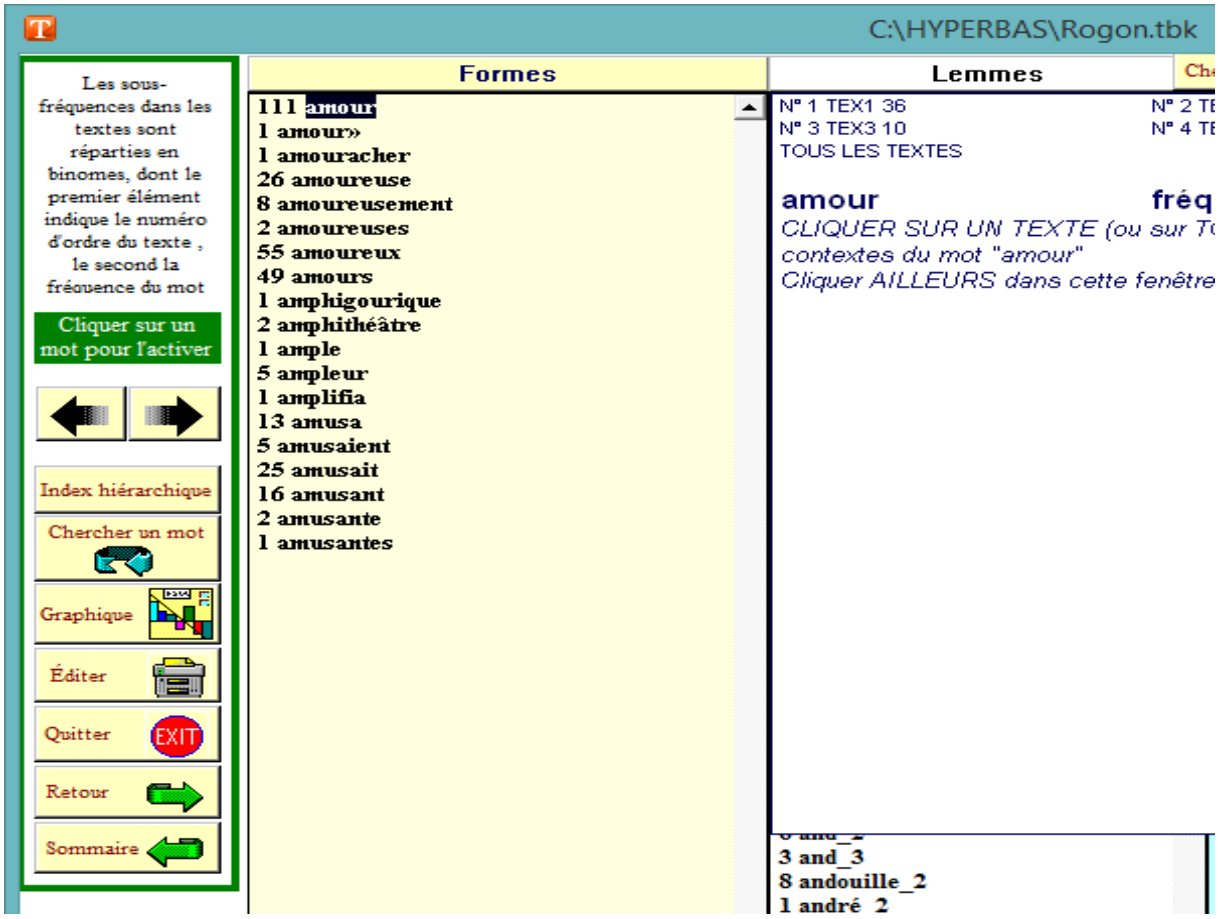

Figure 8: List of wordforms proposed by Hyperbase for the requested term “amour” [love]

Generally, orthographical parsers are not reliable. For instance, if users are looking for all France-related sentences in newspaper archives, selecting all the wordforms that match the regular expression "fran.*" would produce a large number of false positives, such as *franchir* [to cross (a river)], *frangipane* [almond cake], *franchise* [Honesty], while missing terms such as *refranciser* [re-frenchify].

6 Word clouds and semantic networks

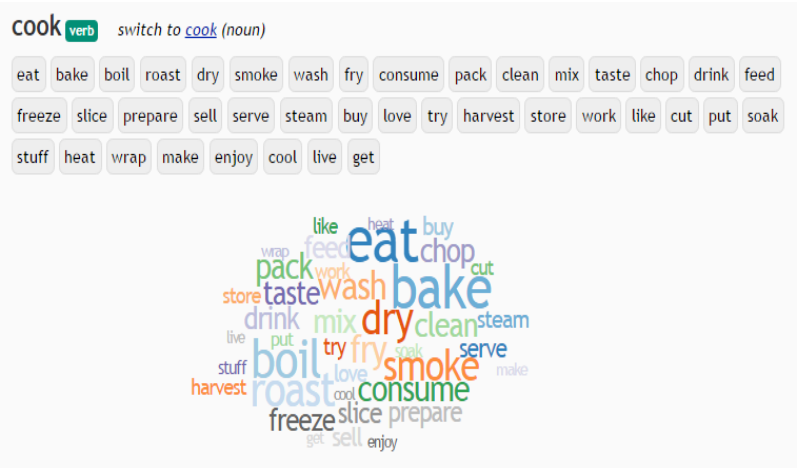

Figure 9: A word cloud for the wordform *cook* provided by Sketch Engine.

To help users study a concept in their text corpora, several corpus processors display a cloud of related words such as the one in figure 9. However, most of the words shown in these clouds are not particularly interesting; for example, there is no benefit in presenting grammatical words such as *get, like, make* or *try* to the users, and the words *clean* (11 meanings listed in Wiktionary), *dry* (13), *pack* (19), *consume* (7), *work* (6), have too many meanings to be useful. A better approach would be to let users control what terms should be looked for, using precisely defined queries such as the following:[32]

```
COOK = <bake> | <boil> | <brew> | <broil> | <cook> | <cuisine> | <fry> | <grill> |
<roast> | <simmer> | <steam> | <stew> | <toast>
```

Each user should be able to define lexical fields according to their specific needs. For instance, a psychologist might want to aggregate occurrences of the verbs *to love* and *to hate* when analyzing the emotivity of a patient, whereas a political scientist who analyzes speeches of various political figures might want to oppose these verbs. For an analysis of tourism-related brochures, the *COOK* lexical field might include terms such as *coffee shop, cooking class, diner, fast-food, gastronomy, gourmet store, restaurant*; for an analysis of dietary-related medical articles, it could include the terms *anorexia, appetite, bulimia, obesity, snacking*; for yet other studies, *breakfast, dinner, kitchen, lunch, oven, pot,* etc.

7 Homography and lexical ambiguity

In the previous sections, we have seen that recognizing a certain ALU often requires the corpus processor to take into account multiple wordforms. The reverse is also true: most wordforms potentially correspond to multiple ALUs. For example, the wordform *card* has a dozen unrelated meanings (*business card, credit card, green card,* etc.); the wordform *plant* too (*anchor plant, hay plant, nuclear plant, packing plant, pot plant*, etc.), the wordform *house* too (*acid house, auction house, country house, shotgun house, white house,* etc.).

Homography is an important characteristic of natural languages: we estimate that each simple French simple noun has over five meanings on average.[33] Moreover, the more frequent a wordform is, the more meanings it corresponds to. For instance, the French noun *carte* [board, chip, card, pass, map, etc.] has over 30 different meanings (=ALUs): *carte de crédit* [credit card], *carte mère* [mother board], *carte d'identité* [ID card], *carte de visite* [business card], *carte mémoire* [memory chip], *carte routière* [road map], etc.

If a wordform corresponds to multiple unrelated meanings, processing it as a single unit is non sensical: what can a political scientist deduce from a statistical analysis that reveals that the wordform "plant" had a peak of

32 In NooJ, symbols in angle brackets represent sets of morphological-related wordforms. For instance, <bake> = bake | bakes | baking | baked | baker | bakers | bakery | bakeries. Users can add constraints to symbols; for instance, <bake,V> represents only the verbal forms bake, bakes, baking and baked.

33 Based on the size of the relative size of the French DELAS (35,000 simple nouns) and the DELAC (200,000 compound nouns) dictionaries, cf. (Courtois, Silberztein 1990). Some French nouns have over thirty meanings, for instance *carte*: *carte de crédit* [credit card], *carte routière* [road map], *carte postale* [postcard], *carte-mère* [mother-board], etc. The DELA systems of dictionaries describe natural languages' standard vocabularies: if one adds to them specialized and technical terms, this ratio will grow significantly.

frequency in a newspaper published in January 2022: was this term related to an FBI agent in a covert operation (as in *a plant*)? or to the legalization of cannabis (as in *cannabis plant*), to a power outage due to an electrical plant failure?

A corpus processor that processes texts at the wordform level might detect that the wordform *matter* has an abnormally high frequency in a certain article of a newspaper and classify it in the scientific section (*gray matter*, *dark matter*, etc.), whereas this wordform would just appear in adverbs such as "as a matter of fact," "for that matter," "in these matters," etc.

In reality, when they analyze their corpus, most researchers in the humanities and the social sciences are looking for specific units of meanings rather than highly ambiguous graphical wordforms. Corpus processors must therefore be able to infer the units of meanings from the wordforms that occur in texts.

Lexical disambiguation is a general and complex problem that we cannot develop here; (Silberztein 2010) showed how verbs meaning disambiguation can be achieved using detailed dictionaries associated with refined syntactico-semantic grammars.[34] However, to retrieve terms — that constitute most of the needs of corpus users in the social sciences — a much simpler approach can be implemented. Its principle is that, instead of processing *card* as an highly ambiguous wordform, it is better to process its multiword units as independent units: *business card*, *credit card*, *green card* etc. These explicit forms are listed in dictionaries such as DELACs.[35]

In practice, multiword units are often abbreviated, as can be seen in:

> *... A third of doctors have asked for mental health support during the Covid pandemic, according to one of the UK's most senior medical organisations. The Royal College of Physicians (RCP) in Wales said Covid had thrown the NHS workforce into "sharp relief". One junior doctor Dr Emma Rengasamy said working through the pandemic was a "baptism of fire" (BBC News, 16 March 2021).*

Most multiword units can be abbreviated; for instance, a *charitable organization* is an organization; the *ministry of education* is a ministry, the *COVID pandemic* is a pandemic, etc. To automatically recognize multiword units even when they are abbreviated, a corpus processor could access grammar rules such as the following:

```
COVID = (COVID | COVID-19 | coronavirus) (crisis | epidemic | pandemic | surge)
```

When this grammar triggers a match in a text, the corpus processor would "activate" this ALU. After being activated, all its subsequent abbreviated occurrences (*e.g. crisis*, or *pandemic*) would then be aggregated and linked to the "COVID" ALU, until of course another ALU (e.g. *migrant crisis*) gets activated.

34 Dubois & Dubois-Charlier's dictionary *Les Verbes Français* (LVF) contains over 20,000 lexical entries, cf. (François et alii 2007). Each entry corresponds to a specific meaning of a verb, associated with a syntactico-semantic characterization of its context. (Silberztein 2010) describes how this dictionary was linked to syntactic and semantic grammars to automatically retrieve all occurrences of one specific meaning of the verb *abriter* (out of five meanings) in *Le Monde diplomatique*.
35 There are DELAC dictionaries available for several languages. The first DELAC dictionary was designed and constructed for French by (Silberztein, 1990).

8 Conclusion and perspectives

Today, more and more researchers in the humanities and the social sciences use corpus processors to explore and analyze their corpora. They are looking for units of meanings, rather than graphical wordforms. We have shown that exhaustive precise linguistic resources and methods can be used to link the wordforms that occur in the text files to the units of meaning that are of interest to the users.

— taking into account a large number of spelling, inflectional, derivational and synonymous variants of terms will lead to a much better recall rate;

— distinguishing between a wordform's different meanings will lead to a better level of precision.

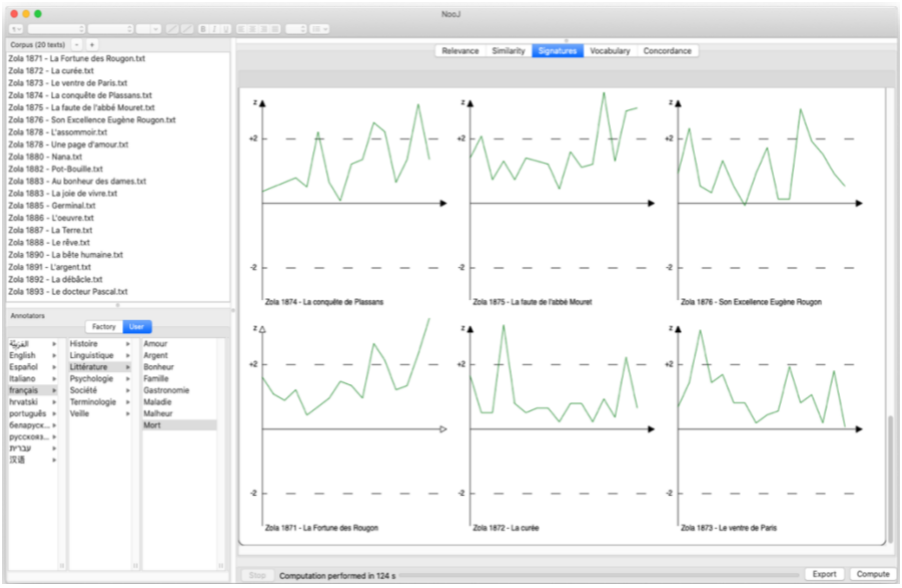

Figure 10: ATISHS' Standard-Score Analysis of the Lexical Field "mort" [death] in Zola's novels.

The ATISHS corpus processor[36] has been developed to provide users with all the statistical tools usually provided by existing corpus processors, applying these tools to units of meaning (ALUs) rather than graphical wordforms. For instance, figure 10 displays a standard score analysis for the occurrences of the lexical field *mort* [death] in the novels of Zola's Rougon-Macquart series. Here is the definition of the lexical field:

```
mort = (<déceder> | <mourir> | <périr> | <suicider>) |
       <cadavre> | <cimetière> | <crémation> | <crématorium> | <deuil> |
<enterrement> | funérailles | <DET> (<A> | <E>) <tombe> | <veuf> |
       <fatal> | <mortel> ;
```

ATISHS processes lexical fields defined by linguistic resources developed with the NooJ platform, that include lexicons of terms associated with their morphological and spelling variants, abbreviations and acronyms and synonyms, morphological grammars that represent their inflected and derived forms, as well as syntactic grammars that describe syntactico-semantic contexts used to eliminate false positives.

36 cf. http://www.nooj4nlp.org/atishs.html. All linguistic resources used by ATISHS have been developed using the NooJ linguistic development environmnent, cf. http://www.nooj4nlp.org/.

Although the linguistic approach has an initial high cost since it requires the development and maintenance of linguistic resources, we believe that it is the only valid approach if one wants to offer researchers in the humanities and the social sciences reliable tools to analyze their corpus.

9 References

Aouini, M. (2018). *Approche multi-niveaux pour l'analyse des données textuelles non-standardisées: corpus de textes en moyen français* (Doctoral dissertation, Bourgogne Franche-Comté University).

Brunet, E. (2011). Hyperbase : Logiciel hypertexte pour le traitement documentaire et statistiques des corpus textuels : http://hyperbase.unice.fr/hyperbase/doc/manuel.pdf

Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155-162.

Courtois, B., & Silberztein, M. (1990). Dictionnaires électroniques du français. *Langue française*, *87*(1), 3-4.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities*, *36*(2), 223-254.

Dickinson, M., & Ledbetter, S. (2012, May). Annotating Errors in a Hungarian Learner Corpus. In *LREC* (pp. 1659-1664).

François, J., Le Pesant, D., & Leeman, D. (2007). Présentation de la classification des Verbes français de Jean Dubois et Françoise Dubois-Charlier. *Langue française*, (1), 3-19.

Green, S., & Manning, C. D. (2010, August). Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 394-402).

Heiden, S. (2010). The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme. In *24th Pacific Asia conference on language, information and computation* (Vol. 2, No. 3, pp. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7-36.

Lamalle, C., & Salem, A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles*, 403-412.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Loubère, L., & Ratinaud, P. (2014). Documentation IRaMuTeQ. Download from: *http://www. iramuteq. org/documentation/fichiers/documentation_19_02_2014. pdf*.

Mathieu-Colas, M. (1990). Orthographe et Informatique: établissement d'un dictionnaire électronique des variantes graphiques. *Langue française*, (87), 104-111.

Reinert, M., (1999). Quelques interrogations à propos de l'" objet" d'une analyse de discours de type statistique et de la réponse" Alceste". *Langage & société*, *90*(1), 57-70.

Schmid, H. (1994). TreeTagger-a language independent part-of-speech tagger. Cf. : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

Silberztein, M. (1990). Le dictionnaire électronique des mots composés. *Langue française*, (87), 71-83.

Silberztein, M. (2010). La formalisation du dictionnaire LVF avec NooJ et ses applications pour l'analyse automatique de corpus. *Langages*, (3), 221-241.

Silberztein, M. (2016). Formalizing Natural Languages: the NooJ approach. Wiley Eds, Hoboken, NJ.

Silberztein, M. (2018). Using linguistic resources to evaluate the quality of annotated corpora. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing* (pp. 2-11).

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn treebank: an overview. *Treebanks*, 5-22.

Volokh, A., & Neumann, G. (2011, June). Automatic detection and correction of errors in dependency treebanks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 346-350).