Using Linguistic Resources to Evaluate the Quality of Annotated Corpora

Max Silberztein
Université de Franche-Comté
max.silberztein@univ-fcomte.fr

Abstract

Statistical and neural network based methods that compute their results by comparing a given text to be analyzed with a reference corpus assume that the reference corpus is complete and reliable enough. In this article, I conduct several experiments to verify this assumption and I suggest ways to improve these reference corpora by using carefully handcrafted linguistic resources.

1 Introduction

Nowadays, most Natural Language Processing (NLP) applications use stochastic methods that are, for example, statistical- or neural network-based, in order to analyze new texts. Analyzing a text involves thus comparing it with a "training" or reference corpus, which is a set of texts that have been either pre-analyzed manually or parsed automatically, and then checked by a linguist. Granted that the reference corpus and the text to analyze are similar enough, these methods produce satisfactory results.

Because natural languages contain infinite sets of sentences, these methods cannot just compare the text to be analyzed with the reference corpus directly at the sentence level. They rather process both the text and the reference corpus at the *wordform* level (i.e. contiguous sequences of letters). To analyze a sentence in a new text, they first look up how each wordform of the text was tagged in the reference corpus, and then they compare the context of the wordform in the text to be analyzed with similar ones in the reference corpus.

The basic assumption of these stochastic methods is that if the reference corpus is sufficiently large, the wordforms that constitute the text to be analyzed will contain enough occurrences to find identical, or at least similar, contexts. Reciprocally, if the reference corpus is too small or too different from the text to be analyzed, then the application will produce unreliable results. Therefore, evaluating the quality of an annotated corpus means answering the following questions:

- what is the minimum size of the annotated corpus needed to produce reliable analyses?
- how reliable is the information stored in an annotated corpus?
- how much information is missing in an annotated corpus, and how does the missing information affect the reliability of the analysis of new texts?

For this experiment, I have used the NooJ linguistic development environment[1] to study the *Slate* corpus included in the Open American National Corpus[2]. This sub-corpus, constituted by 4,531 articles/files, contains 4,302,120 wordforms. Each wordform is tagged according to the Penn tag set[3].

[1] NooJ is a free open-source linguistic development environment, distributed by the European Metashare platform, see (Silberztein, 2003) and (Silberztein, 2016).

[2] The Open American National Corpus (OANC) is a free corpus and can be downloaded at www.anc.org. We have looked at the Corpus of Contemporary American English (COCA), which has a subset that is free of charge: we will see in section 2 that it has problems related to vocabulary similar to those of the OANC. (Silberztein, 2016) has evaluated the reliability of the Penn treebank and found results similar to those discussed in section 4.

[3] In the OANC as well as in other annotated corpora such as the Penn treebank or the COCA, sequences of digits, punctuation characters and sequences that contain one or more dashes are also processed as linguistic units.

Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing, pages 2–11
Santa Fe, New Mexico, USA, August 20, 2018.

2 Vocabulary Coverage

2.1 Stability of the vocabulary

As a first experiment, I split the *Slate* corpus into two files: Even.txt contains all the articles whose original filename ends with an even number (e.g. "ArticleIP_1554.txt"), whereas Odd.txt contains all the articles whose original filename ends with an odd number (e.g. "ArticleIP_1555.txt"). These two corpora are composed of intertwined articles, so that vocabulary differences cannot be blamed on chronological or structural considerations.

As Figure 1 shows,[4] almost half of the wordforms in the vocabulary of the total corpus either occur in Even.txt but not in Odd.txt or occur in Odd.txt but not in Even.txt. In other words, for half of the wordforms in the corpus' vocabulary, the fact that they occur or not appears to be a random accident. The fact that the vocabularies of two random subsets of this 4-million-wordform corpus are so different shows that this corpus is still too small to have a stabilized vocabulary.

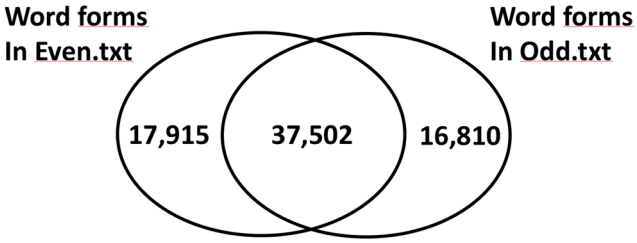

Figure 1. Vocabulary is unstable

Wordforms that occur in one sub-corpus but not in the other one, are not necessarily hapaxes or even rare wordforms: there are 4,824 wordforms that occur more than once in Even.txt, but never occur in Odd.txt, and there are 4,574 wordforms that occur more than once in Odd.txt, but never occur in Even.txt. The following are examples of wordforms that occur 10 times or more in one sub-corpus, but never occur in the other one:

- In Odd.txt: *cryptography* (12 occurrences), *mammary* (12), *predation* (12), *selector* (15), etc.
- In Even.txt: *irradiation* (13), *jelly* (17), *obsolescence* (11), *quintet* (16), *sturgeon* (10), etc.

The vocabulary covered by this 4-million-wordform corpus is still massively unstable, which indicates that the corpus size is much too small to cover a significant portion of the vocabulary.

2.2 Evolution of the vocabulary

As a second experiment, I studied the evolution of the size of the vocabulary in the full corpus. As we can see in Figure 2, the number of different wordforms (dashed line) first increases sharply and then settles down to a quasi-linear progression. That is not surprising because in a magazine, we expect to find a constant flow of new proper names (e.g. *Abbott*) and new typos (e.g. *achives*), as new articles are added to a corpus.

In NooJ, every element of the vocabulary is described by one, and only one, lexical entry. If a given wordform is polysemous, i.e. corresponds to more than one vocabulary element (e.g. *to milk* vs. *some milk*), then it is described by more than one lexical entry.

NooJ's Atomic Linguistic Units (ALUs) are either lexical entries (e.g. *eat*), inflected forms of a lexical entry (e.g. *eaten*) or derived forms of a lexical entry (e.g. *eatable*). NooJ handles affix ALUs (e.g. *re-*, *-ly*), simple ALUs (e.g. *table*), compound ALUs (e.g. *blue collars, in spite of*) as well as discontinuous ALUs (e.g. *turns ... off*, *took ... into account*). Note that contracted forms (e.g. *cannot*) and agglutinated forms (e.g. *autodialed*), very frequent in languages such as Arabic or German, are not ALUs: they are processed as sequences of ALUs.

[4] The total number of wordforms shown in Figure 1 is lower than 88,945, because we unified the forms that have different cases in the two sub-corpora. For instance, the wordform *Deductions* (only in Odd.txt) and the wordform *deductions* (only in Even.txt) are counted as one wordform.

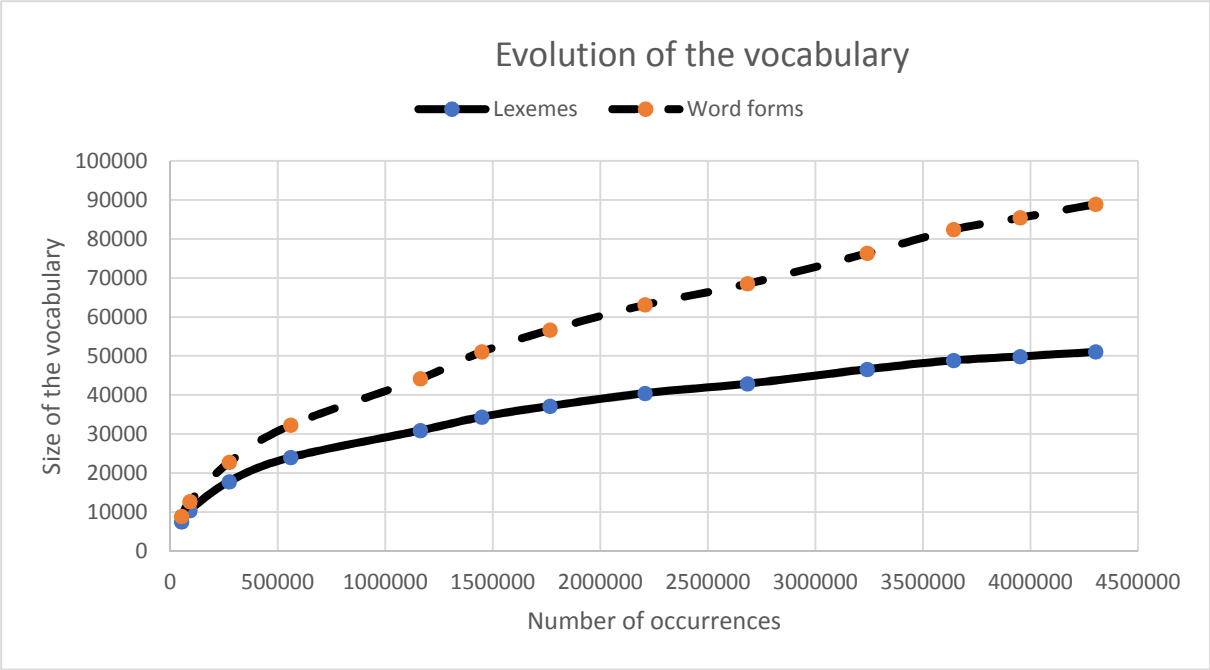

Figure 2. Evolution of the vocabulary

When evaluating the coverage of a reference corpus, it is important to distinguish ALUs from wordforms that have no or poor intrinsic linguistic value, such as numbers, uncategorized proper names and typos. To estimate the evolution of the vocabulary, I applied the English DELAS dictionary to the *Slate* corpus. The DELAS dictionary[5] represents standard English vocabulary, i.e. the vocabulary used in general newspapers and shared by most English speakers. It also includes non-basic terms such as *aardvark* (an African nocturnal animal) or *zugzwang* (a situation in chess), but it does not contain the millions of terms that can be found in scientific and technical vocabularies (e.g. the medical term *erythematosus*), nor terms used in regional variants all over the world (e.g. the Indian-English term *freeship*). The DELAS dictionary contains approximately 160,000 entries, which correspond to 300,000 inflected forms. Applying it to this corpus recognizes 51,072 wordforms, approximately a sixth of the standard vocabulary. Here are a few terms that never occur in the *Slate* corpus:

abominate, acarian, adjudicator, aeolian, aftereffect, agronomist, airlock, alternator, amphibian, anecdotic, apiculture, aquaculture, arachnid, astigmatism, atlas, autobiographic, aviator, awoken, axon, azimuth, etc.

The number of ALUs present in the corpus (solid line in Figure 2) grows slower and slower. By extrapolating it, I estimate that one would need to add at least 16 million wordforms to this 4-million-wordform corpus to get a decent 1/2 coverage of the standard vocabulary. Manually tagging such a large corpus — or even checking it after some automatic processing — would represent a considerable workload,[6] obviously implying a much larger project than simply constructing an English dictionary such as the DELAS.

Processing verbs correctly is crucial for any automatic parser because verbs impose strong constraints upon their contexts: for instance, as the verb *to sleep* is intransitive, one can deduce that any phrase that occurs right after it (e.g. *Joe slept last night*) has to be an adjunct rather than an argument (as in *Joe enjoyed last night*). Knowing that the verb *to declare* expects a subject that is a person, or an organization, allows automatic software to retrieve the pronoun's reference in the sentence: *They*

[5] The DELA family of dictionaries were created in the LADL laboratory, see (Courtois and Silberztein 1990; Klarsfeld 1991; Chrobot et al., 1999).

[6] 20 million wordforms to check, one second per check, 8-hours a day, 5-day a week, would take 3 years. By contrast, it takes between 6 and 9 months for a linguist to construct a NooJ module for a new language (that includes a DELAS-type dictionary).

declared an income, etc. However, if the reference corpus does not contain any occurrence of a verb, a statistical or neural-network based parser would have no means to deduce anything about its syntactic context nor its distributional restrictions, and therefore they would not be able to reliably process sentences that contain the verb.

The 4-million-wordform *Slate* corpus contains 12,534 wordforms tagged as verbal forms (tags VB, VBD, VBG, VBN, VBP or VBZ), which represents only a fifth of the 62,188 verbal forms processed by NooJ. Following are examples of verbs that never occur in the *Slate* corpus:

acerbate, acidify, actualize, adjure, administrate, adulate, adulterate, agglutinate, aggress, aliment, amputate, aphorize, appertain, arraign, approbate, asphyxiate, etc.[7]

2.3 Compound words

In the *Slate* corpus, a few compound words that contain a hyphen have been tagged as units, e.g. "a-capella_JJ". But these same exact compounds, when spelled with a space character, were processed as sequences of two independent units, e.g. "a_DT capella_NN". At the same time, a large number of sequences that contain a hyphen, but are not compounds, have also been tagged as linguistic units, e.g.:

abide-and, abuse-suggesting, activity-regarded, adoption-related, Afghan-based, etc.

Similarly, in the COCA, we can find *left-wing*, *wing-feathers* and *ultra-left-wing* tagged properly as ALUs, whereas *left wing*, *wing commander*, *wing nuts* are processed as sequences of two independent units. It seems that there is a systematic confusion between *compound ALUs* and sequences that contain a dash[8] in annotated corpora. In reality, most compounds do not contain a hyphen. For example, all occurrences of the adverb *as a matter of fact* have been tagged in the *Slate* corpus as:

As_IN a_DT matter_NN of_IN fact_NN

and in the COCA as:

as	as	ii
a	a	at1
matter	matter	nn1
of	of	jj32_nn132
fact	fact	nn1

This type of analysis makes it impossible for any NLP applications to process this adverb correctly. One would not want a MT system to translate this adverb word by word, nor a search engine to return these occurrences when a user is looking for the noun *matter*. In fact, a Web Semantic application should not even try to link these occurrences to the entities *dark matter* (2 occurrences), *gray matter* (2 occ.), *organic matter* (1 occ.) nor *reading matter* (1 occ.), etc.

NooJ's DELAC dictionary[9] contains over 70,000 compound nouns (e.g. *bulletin board*), adjectives (e.g. *alive and well*), adverbs (e.g. *in the line of fire*) and prepositions (e.g. *for the benefit of*). These entries correspond to approximately 250,000 inflected compound words. By applying the DELAC dictionary to the corpus, NooJ found 166,060 occurrences of compound forms, as seen in Figure 3.

These 166,060 compounds correspond to approximately 400,000 wordforms (i.e. 10% of the corpus) whose tags are either incorrect, or at least not relevant for any precise NLP application.

---

[7] Some of these forms do occur in the *Slate* corpus, but not as verbs. They have rather been tagged as adjectives, e.g. *actualized, amputated, arraigned,* etc.

[8] Silberztein (2016) presents a set of three criteria to distinguish between analyzable sequences of words and lexicalized multiword units: (1) the meaning of the whole cannot be completely computed from its components (e.g. a *green card* is much more than just a *card* that has a *green* color), (2) everyone uses the same term to name an entity (e.g. compare *a washing-machine* with *a clothes-cleaning device*), and (3) the transformational rules used to compute the relation between its components has some idiosyncratic constraints (compare the function of the adjective *presidential* in the two expressions: *presidential election* (*we elect the president, *presidents elect someone*) and *presidential race* (**we race the president, the presidents race against each other*)).

[9] Silberztein (1990) presents the first electronic dictionary for compounds (French DELAC), designed to be used by automatic NLP software. The English DELAC is presented in Chrobot et al. (1999).

These 166,060 occurrences represent 25,277 different compound forms, which amounts to only 10% of the English vocabulary. Standard terms such as the following never occur in the *Slate* corpus:

abandoned ship, *access path, administrative district, aerosol spray, after hours, agglutinative language, air bed, album jacket, ammunition belt, anchor box, appeal court, aqueous humor, arc welder, assault charge, attitude problem, auction sale, aviator's ear, awareness campaign, axe to grind, azo dye*, etc.

Moreover, most compound words actually found in the *Slate* corpus do not occur in all their forms: some nouns only occur in their singular form, whereas others only occur in their plural form. For instance, there are no occurrence of the singular forms of the following nouns:

absentee voters, *access codes*, *additional charges, affinity groups, AID patients, Alsatian wines, amusement arcades, ancient civilizations, appetite suppressants, armed extremists, assembly operations, attack helicopters, audio guides, average wages, ax-grinders*, etc.

Even if encountering an occurrence of any inflected or derived form for a lexical entry would allow an automatic system to correctly parse all its other inflected and derived forms, the *Slate* corpus only covers 10% of the compounds of the vocabulary. I estimate that one would need to add over 32 million wordforms to the corpus to get a decent 1/2 coverage of the English compounds.[10]

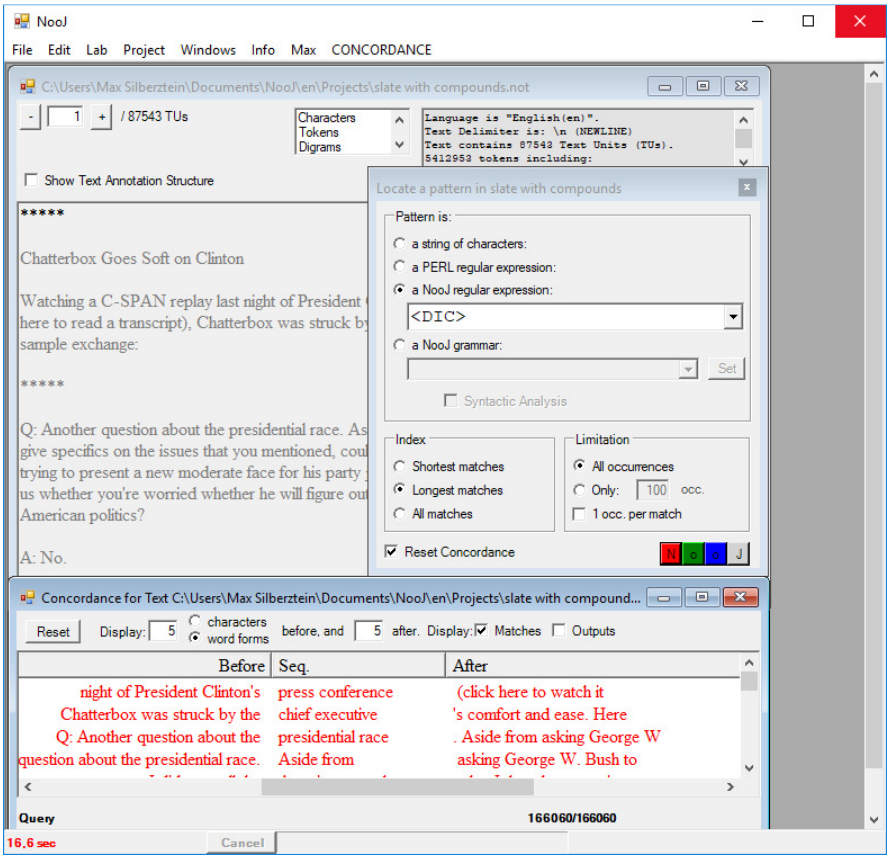

Figure 3. Compounds in the *Slate* corpus

2.4 Phrasal Verbs

Any precise NLP application must take into account all multiword units, even those that are discontinuous. Examples of discontinuous expressions include idiomatic expressions (e.g. *to read … the*

[10] 32 million wordforms to check, one second per check, 8-hours a day, 5-day a week, would take over 4 years. By contrast, it typically takes a year for a linguist to construct a DELAC-type dictionary.

riot act), verbs that have a frozen complement (e.g. *to take … into account*), phrasal verbs (e.g. *to turn … off*) and associations of predicative nouns and their corresponding support verb (e.g. *to take a* (*<E> | good | long | refreshing*) *shower*).

For this experiment, I applied NooJ's dictionary of phrasal verbs[11] to the *Slate* corpus. This dictionary contains 1,260 phrasal verbs, from *act out* (e.g. *Joe acted out the story*) to *zip up* (*e.g. Joe zipped up his jacket*). NooJ recognized over 12,000 occurrences[12] of verbal phrases, such as in:

… acting out *their predictable roles in the…*
… I would love to ask *her* out,…
… would have backed *North Vietnam* up…
… Warner still wanted to boss *him* around…
… We booted up *and victory!...*

However, less than 1/3 of the phrasal verbs described in the dictionary had one or more occurrences in the *Slate* corpus. For instance, phrasal verbs such as the following have no occurrence in the corpus: *argue down, bring about, cloud up, drag along, eye up, fasten up, goof up, hammer down*, etc.

3 Hapaxes

3.1 Wordforms and ALUs

In most applications that use statistical approaches (e.g. Economics, Medicine, Physics, etc.), hapaxes — i.e. statistical events that only occur once — are rightfully ignored as "accidents," as they behave like "noise," by polluting analysis results.

In linguistics, a hapax is a wordform that occurs only once in a reference corpus. There are reasons to ignore hapaxes during a text analysis since the unique syntactic context available cannot be used to make any reliable generalization. Following are examples of hapaxes that occur right after a verb in the OANC:

- an adjective, e.g. … *one cage is left* **unrattled**…
- an adverb, e.g. … you touched **unerringly** on all the elements…
- a noun, e.g. that score still seemed **misogynous**…
- an organization name, e.g. the center caused **Medicare** to pay for hundreds…
- a person name, e.g. *Lewinsky told* **Jordan** *that…*
- a verbal form, e.g. the deal might **defang** last year's welfare reform...
- a foreign word, e.g. *they graduated* **magna** *cum laude…*
- or even a typo, e.g. *whipped mashed* **potatos** *and…*

It is only by taking into account multiple syntactic contexts for a wordform that one can hope to describe its behavior reliably. If a wordform in the text to be analyzed corresponds to a hapax in the reference corpus that occurs right after a verb, it would be very lucky if a statistical or neural-network based parser were tagging it correctly.

There are 31,275 different hapaxes in the *Slate* corpus, out of 88,945 different wordforms, i.e. a third of the vocabulary covered by the *Slate* corpus, which covers itself a sixth of the standard vocabulary. Consequently, statistical parsers that do not exclude hapaxes will produce unreliable results for up to one third of the wordforms present in the reference corpus' vocabulary.

[11] Peter Machonis is the author of the Lexicon-Grammar table for English Phrasal Verbs, which has been integrated into NooJ via a linked couple dictionary / grammar (Machonis, 2010).

[12] There are a few false-positives, i.e. phrasal verbs that were recognized but do not actually occur in the text, such as: *The Constitution* grew out *of a convention*; they represent less than 2% of the matches. Half of these errors could be avoided by using simple local grammars, such as giving priority to compounds (so that the recognition of the compound preposition *out of* would block the recognition of the phrasal verb *grew out* in the latter example).

3.2 Compound words

As we have seen previously, the OANC and the COCA (as most reference corpora) contain no special tags for compound words, which are nevertheless crucial for any precise NLP application. To automatically identify them, some researchers use statistical methods to try to locate colocations.[13] Their idea is that if, for instance, the two wordforms *nuclear* and *plant* occur together in a statistically meaningful way, one may deduce that the sequence *nuclear plant* corresponds to an English term. Even if one subscribes to this principle, statistical methods cannot deduce that a sequence of wordforms probably corresponds to a term if it occurs only once. In the *Slate* corpus, there are 9,007 compound words that only occur once, e.g.:

absent without leave, accident report, adhesive tape, after a fashion, age of reason, air support, alarm call, American plan, animal cracker, application process, artesian well, assault pistol, atomic power, automatic pilot, aversion therapy, away team, axe to grind, etc.

If one removes these hapaxes from the list of compound words that occur in the corpus, the number of compound words that could theoretically be detected as co-locations is reduced to 25,277 – 9,007 = 16,270 occurrences, i.e. only 6% of the vocabulary.

Note that the colocation criterion does not really make any sense from a linguistic point of view. It is not because a sequence occurs often that it is necessarily an element of the English language vocabulary (e.g. the sequence *was in the* occurs 69 times), and reciprocally it is not because a sequence only occurs once in a corpus (e.g. *after a fashion*) that it is a lesser element of the English vocabulary. In the same manner that it would not make any sense to state that *alright* is not an element of English vocabulary because it only occurs once in the *Slate* corpus, it does not make sense to state that *artesian well* is not a term because it only occurs once in the same corpus.

3.3 Polysemy

To be reliable, statistical-based disambiguation techniques need to process units that are frequent enough. For example, in the *Slate* corpus, the wordform *that* is tagged 42,781 times as a subordinating conjunction (IN), out of 62,286 occurrences. It is then fair to predict that any of its occurrences has a 70% probability of being a subordinating conjunction.

However, if a corpus contains only one occurrence of a polysemous wordform, predicting its function in a new text can only produce unreliable results. For instance, the wordform *shrivelled* occurs only once in the *Slate* corpus:

… an orange that was, in Zercher's words, shrivelled and deformed...

It has been correctly tagged as an adjective (JJ), but this is not a reason to deduce that this wordform will always function as an adjective, as one can see in sentence: *The lack of rain has shrivelled the crops*. In the *Slate* corpus, there are 2,285 wordforms that have two or more potential tags, but occur only once, e.g.:

aboveboard (adjective or adverb), accusative (adjective or noun), advert (noun or verb), aflame (adjective or adverb), agglomerate (adjective, noun or verb), airdrop (noun or verb), alright (adjective or adverb), amnesiac (adjective or noun), angora (adjective or noun), apologetics (singular or plural noun), aqueous, armour (noun or verb), astringent (adjective or noun), attic (adjective or noun), auburn (adjective or noun), Azerbaijani (adjective or noun), etc.

Any parser that processes these wordforms as monosemous (because they only occur once in the reference corpus) produces unreliable results.

13 See, for instance, the European PARSEME initiative `http://typo.uni-konstanz.de/parseme` and the program of the SIGLEX-MWE (Special Interest Group on Multiword Expressions) workshops `http://multiword.sourceforge.net/PHITE.php?sitesig=MWE`.

4 Reliability

All statistical or neural-network based NLP applications that compare a reference corpus to the texts to analyze assume that the reference corpus can be relied upon: if the tags used to compute an analysis are incorrect, then one cannot expect these applications to produce perfect analyses. The fact that reference corpora contain errors is well known to NLP researchers.[14] Even though, I believe that the actual number of errors has been largely ignored or minimized.

The Open American National Corpus has been tagged by an enhanced version of GATE's ANNIE system,[15] using the Penn tag set. However, during this series of experiments, every superficial look at the *Slate* corpus[16] has uncovered mistakes. For instance, when looking at the form *that* in section 3.3., I found the following analyses for this wordform:

- 41,622 times as a subordinating conjunction (IN),
- 10,151 times as a Wh-determiner (WDT),
- 10,491 as a determiner (DT),
- 124 times as an adverb (RB).

All WDT (Wh-determiner) and RB (adverb) tags for the wordform *that* are incorrect: in fact, they correspond to pronoun uses of *that*. Here are a few examples of mistakes:

- Incorrect WDT: … Publications that refuse to… phrase that describes this… stigma that came with…
- Incorrect RB: … likes to boast that, … to do just that, and delightfully … know that, …

Thus, at least 25% of the occurrences of the wordform *that* have been tagged incorrectly. There is a systematic confusion between Wh-determiners and pronouns (WDT instead of WP); occurrences of the wordform *that* that are followed by a period or a comma have been systematically tagged as adverbs. The "systematic" aspect of these mistakes in the corpus is unsettling, because it means that enlarging the corpus to 10 or even 100 million words will not enhance the usefulness of the statistical methods: there is no *really useful* information added to the corpus if it was added automatically. As a matter of fact, a superficial look at the full OANC corpus shows the same systematic mistakes as in the *Slate* sub-corpus, which suggests the use of automatic disambiguation rules.

If — as systematic mistakes imply — the tagger used automatic rules to disambiguate wordforms, it is essential that we are able to look at them, so that we can correct them. One may even argue that these automatic disambiguation rules could even be inserted directly in the final NLP application: in that case, the reference corpus becomes less and less useful.

To estimate the accuracy of the tags in the corpus, I have compiled a list of its 84,386 useful wordforms[17] associated with their tags, and parsed it with NooJ:

- Out of the 49,323 wordforms that were not tagged as proper names, 11,019 tags — i.e. over 20% — are considered as incorrect by NooJ. Examples of incorrect tags are:

abbreviate, abduct, abhor, abhors, etc. (not nouns),
about, agonized, bible, cactus, California, etc. (not adjectives)
expenditures, Japanese, many, initiatives, wimp, etc. (not verbs)
anomaly, back, because, by, of, out, particular, upon, etc. (not adverbs)

A number of derived or agglutinated forms have been tagged as ALUs, e.g.:

[14] See for instance Green and Manning (2010) about tagging errors in Arabic corpora, Kulick et al. (2011) and Volokh and Neumann (2011) about errors in tree-banks, and Dickinson (2015) about methods to detect the annotation errors.
[15] See `http://gate.ac.uk`.
[16] The Open American Corpus has been tagged
[17] Tags such as CD (Cardinal Number), LS (List item marker), POS (Possessive ending), SYM (symbol) and TO (to) are not useful in the sense that they do not add any information to the word they describe; they can be automatically added with simple SED-type replacements such as s/\([0-9]*\)/\1_CD/.

audienceless, autodialed, balancingly, bioremediation, barklike, etc.

However, there are good linguistic reasons to consider these wordforms to be analyzable sequences of two ALUs (i.e. two tags), in order to process prefixes such as *auto-, bio-* and suffixes such as *-less, -ly, -like* as ALUs themselves.

- Out of the 35,063 uppercase wordforms that have been associated with an Proper name (tagged NNP or NNPS), we also get a large number[18] of incorrect tags, e.g.:

Abacuses, Abandoned, ABATEMENT, Abattoir, Abbreviated, Ablaze,
Abnormal, Abolished, Abuse, Abstract, Accidental, ALMOST, etc.

The Penn tag set does not contain tags for typos: all uppercase typos were tagged as proper names (NNP or NNPS), e.g.: *Aconfession, Afew, AffairsThe, Allpolitics*, etc. Most other typos have been associated with a noun tag, e.g.:

absentionists (NNS), achives (NNS), accrossthe (NN), afteryou (NN), alwaysattend (NN), etc.

18,949 sequences that include the em dash[19] have been incorrectly processed as compound words, e.g. *aggression–that* (JJ), *believe–that* (JJ), *chance–a* (JJ), etc. Finally, one needs to mention that there are over 200 typos in the tags themselves, for instance:

a,DTn, believ_VBe, classi_JJc, di_VBDd, JFK-styl_NNPe, etc.

Adding the 10% irrelevant tags (e.g. "as_IN" in "as a matter of fact") to the 20% incorrectly tagged sequences that include a hyphen or an em dash (e.g. "minister—an_JJ") , to the 20% impossible tags (e.g. "many_VB"), as well as the typos, makes the *Slate* corpus unreliable at best.

5 Granularity of the linguistic information

Dictionaries for NLP applications actually need to associate their lexical entries with more information than just their part of speech. Nouns need to be classified in the very least as Human, Concrete or Abstract, because there are syntactic and semantic rules that do not apply in the same way to human, concrete or abstract noun phrases. Verbs need to be associated with a description of their subjects, prepositions and complements. For instance, in order to analyze correctly the following sentence:

The classroom burst into laughter

a computer program needs to know that the (compound) verb *to burst into laughter* expects a human subject, that *classroom* is not a human subject, and therefore it needs to activate a special analysis (e.g. metonymy) to process (e.g. index/link/translate/etc.) it.

In the previous evaluations, I counted wordforms as if each were representing all its corresponding ALUs: for instance, when the wordform *agent* occurs in the corpus, I counted it as if it represented two lexical entries: the person (in *a secret agent*), and the product (in *a bleaching agent*). But when parsing a text, NooJ produces over 2 potential analyses for each wordform in average. In consequence, if we were to take this linguistic information into account by enriching the tag set (e.g. add a +Human or +NonHuman feature for nouns, add a +Transitive or +Intransitive feature for verbs, distinguish between multiple meanings of each lexical entry, etc.), the coverage of the corpus would be half of our previous estimate.

[18] It is not possible to know if an uppercase word is or is not a proper name without a syntactic analysis of its context. For instance, "Black", "Carpenter", "Hope", etc. are common words as well as proper names, depending on their context. That said, by scanning over the list of the words tagged as proper names, I estimate that e at least 15% of them are mistakes.
[19] The em dash is represented by two consecutive dash characters in the OANC *Slate* corpus.

6 Conclusion and proposals

On the one hand, a large number of NLP applications rely on reference corpora to perform sophisticated analyses, such as intelligent search engines, automatic abstracts, information extraction, machine translation, etc. On the other hand, linguists have spent years carefully handcrafting a large quantity of linguistic resources. Using these resources could enhance reference corpora significantly:

- a first operation would be to compare the set of tagged wordforms in the reference corpus with the lexical entries in an English dictionary such as the DELAS; this operation would allow one to correct typos as well as "impossible" tags (e.g. *many* should never be tagged as a verb);
- a second operation would be to tag compound words (e.g. *as a matter of fact*) and discontinuous expressions (e.g. *ask … out*) already listed and described in dictionaries, as processing these linguistic units as a whole is crucial for any precise NLP application;
- a third operation would be to develop a set of simple local grammars to correct tags in the text that are not compatible with the English grammar (e.g. *that* should not be tagged as a Wh-determiner in *phrase that describes this*). These local grammars could even replace the meaningless algorithms (e.g. *that* before a comma tagged as an adverb) used to tag the reference corpus;
- a fourth, more ambitious project would be to enhance the tag set in order to distinguish human, concrete and abstract nouns, as well as to classify verbs according to their complements. That would allow sophisticated NLP applications such as Information Extraction and MT software to be trained on a semantic-rich tagged corpus.

References

Agata Chrobot, Blandine Courtois, Marie Hamani, Maurice Gross, Katia Zellagui. 1999. *Dictionnaire Electronique DELAC anglais : noms composés*. Technical Report #59, LADL, Université Paris 7: Paris, France.

Blandine Courtois and Max Silberztein (editors). 1990. *Les dictionnaires électroniques du français*. Larousse: Paris, France.

Markus Dickinson. 2015. Detection of Annotation Errors in Corpora. In *Language & Linguistics Compass*, vol 9, Issue 3. Wiley Online Library, `https://doi.org/10.1111/lnc3.12129`.

Spence Green and Christopher Manning. 2010. Better Arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Confeence on Computational Linguistics* (COLING), pages 394-402.

Gaby Klarsfeld. 1991. *Dictionnaire morphologique de l'anglais*. Technical Report, LADL, Université Paris 7: Paris, France.

Seth Kulick, Ann Bies, Justin Mott. 2011. Further Developments in Treebank Error Detection Using Derivation Trees. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, pages 693–698, Portland, Oregon, USA.

Peter Machonis. 2010. English Phrasal Verbs: from Lexicon-Grammar to Natural Language Processing. *Southern Journal of Linguistics*, vol. 34, n01: 21-48.

Max Silberztein. 2003. *The NooJ Manual*. Available for download at `www.nooj-association.org`.

Max Silberztein. 1990. Le dictionnaire électronique DELAC. In Les dictionnaires électroniques du français. Larousse: Paris, France.

Max Silberztein. 2016. *Formalizing Natural Languages: the NooJ Approach*. Cognitive Science Series. Wiley-ISTE: London, UK.

Alexander Volokh, Günter Neumann. 2011. Automatic detection and correction of errors in dependency tree-banks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technology*: short papers, vol. 2, pages 346-350.