Fine-tuning a Subtle Parsing Distinction Using a Probabilistic Decision Tree: the Case of Postnominal *"that"* in Noun Complement Clauses vs. Relative Clauses

Zineddine Tighidet
Université Paris Cité
F-75013 Paris, France
tighidet.zineddine@gmail.com

Nicolas Ballier
CLILLAC-ARP and LLF, Université Paris Cité
& CNRS F-75013 Paris France
nicolas.ballier@u-paris.fr

Abstract

In this paper we investigated two different methods to parse relative and noun complement clauses in English and resorted to distinct tags for their corresponding *that* as a relative pronoun and as a complementizer. We used an algorithm to relabel a corpus parsed with the GUM Treebank using Universal Dependency. Our second experiment consisted in using Tree-Tagger, a Probabilistic Decision Tree, to learn the distinction between the two complement and relative uses of postnominal *"that"*. We investigated the effect of the training set size on TreeTagger accuracy and how representative the GUM Treebank files are for the two structures under scrutiny. We discussed some of the linguistic and structural tenets of the learnability of this distinction.

1 Introduction

English has relative clauses (*the man that I saw*) and noun complement clauses (*the fact that I saw a man*) that may have similar surface representations (often the definite article, a noun, often immediately followed by *that*) but different structural properties (Ballier, 2004). For POS-tagging systems based on trigrams, the distinction between these constructions can be challenging, not to mention the case of ambiguous sentences such as "the suggestion that he was advancing was ridiculous" (Huddleston, 1984).This is an issue for information retrieval, as conceptual argumentation makes heavy uses of noun complement clauses (Ballier, 2007), the governors of these noun complement clauses being "shell nouns" (Schmid, 2000). Complement taking nouns (Bowen, 2005) are crucial for the expression of stance (Charles, 2007) in documents, which is why this distinction may matter more than is usually assumed.

The Penn Treebank (Marcus et al., 1993) tagset (Santorini, 1990) does not make strict distinctions between the part-of-speech (POS) tag of

"that" when used as a relative pronoun (WDT) or when used as a conjunction when complementizing nouns: it uses IN when complementizing verbs or nouns. Even though the CLAWS8 [1]. (University Centre for Computer Corpus Research on Language, 1995-2004) tagset encodes this distinction with the *CST*[2] and *WPR*[3] tags, this tagger is not free and remains the property of the University Centre for Computer Corpus Research on Language (UCREL). To the best of our knowledge, the precision and recall of these two tags (and their corresponding syntactic structures) have not been reported.

Admitting POS-tagging systems have reached an overall satisfactory precision rate for standard English tagsets, we claim that this is not necessarily the case for tags that reflect such a subtle distinction which may have very similar surface representations. Discussing such POS-tags involves parsing issues of the *that*-clause that follows the noun. Our research question is mostly based on the ability of a system to identify noun complement clauses as apposed to (restrictive) relative clauses, but this can be addressed by analysing dependency relation labels (parsing) or distinct tags that encode this syntactic distinction (POS-tagging). We present the two strategies in two experiments, exploring whether such specific Universal Dependency labels can be learnt. In this paper, we only investigate overt complementizers as we are also investigating how that is tagged and do take into account noun complement clauses with zero complementizer, like in the example *Plus the fact I'm a coward* from the British National Corpus (Consortium et al., 2007).

The rest of the paper is structured as follows. Section 2 details the data we used for our exper-

---

[1]CLAWS, the Constituent Likelihood Automatic Word-tagging System, is the name of the tagset and of the POS-tagging software for English text, CLAWS (Garside, 1987)

[2]*"that"* as a conjunction

[3]*"that"* as a relative pronoun

iments. Section 3 analyses the Universal Dependency (UD) GUM Treebank for English in terms of precision for the dependency labels of these two structures as well as their distribution across the training, testing and development sets. We describe an experiment replicating one of the specific features of the GUM Treebank. Section 4 details an experiment based on algorithm adapting the UD annotation generated with GUM. Section 5 explains how Treetagger can be used to learn distinct tags for that used as a relative pronoun (WPR) or as a complementizer (CST). Section 6 discusses our results and section 7 outlines our future research.

2 Material and Methods

2.1 Test Sets

For our validation procedure, we used two test sets NCCtest and RCtest, one including 194 noun complement clauses (NCC), the other one included 189 relative clauses (RC). As language is complex, some sentences included other syntactic realisations, and a couple of "distractors" representative of the alternate structure were therefore included in our two test sets. We specify in Table 3 the expected (gold) label counts for each test set. Two annotators agreed on these gold labels of these two test sets ($\kappa = 1$).

2.2 Brown Corpus

We used the Brown corpus (Kucera et al., 1967), which is rather small with its 1 M tokens by contemporary standards, but well-balanced and freely available. Its current distribution in the NLTK python library (Bird, 2006) has been POS-tagged with the Penn Treebank, this is the substrate we used for our re-annotation experiment with TreeTagger. Treetagger is a probabilistic tagger which uses decision trees for probability transitions, which is robust for its retraining and claims accuracy above 96 % (Schmid, 1994).

2.3 Universal Dependency Annotation with UDPipe

UDPipe (Straka, 2018) is a pipeline that takes as input a text file and renders a CoNLL-U[4] file which contains the language-specific part-of-speech tag (*XPOS*), lemma or stem, the *DEPREL* (universal dependency relation) etc.
A file annotated in Universal Dependency contains among other columns the *XPOS* (part of speech) for

---

[4]https://universaldependencies.org/format.html

each token and the dependency relation, *acl:relcl* for relative clauses and (just) *acl* for noun complement clauses, though this more general category (*acl* corresponds to clausal modifier of noun, adnominal clause) also includes non-finite clause.

Clausal modifier of noun (*acl*)
acl stands for finite and non-finite clauses that modify a noun. The governor (head) of the *acl* dependency relation is the noun that is modified, and the dependent is the predicate of the clause that modifies the noun. In Figure 1 the finite clause "as he sees them" modifies the noun "the issues".
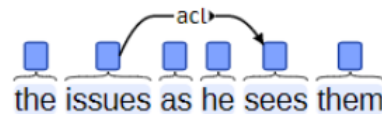

Figure 1: Example of clause modifier of noun (*acl*).

As evidenced by this example taken from the UD documentation, acl is a label that encompasses more than *that* noun complement clauses.

Relative clause modifier (*acl:relcl*)
A relative clause modifier of a noun is a clause that modifies the antecedent. The *acl:relcl* relation points from the governor (the antecedent) head of the modified nominal to the dependent (verb) of the relative clause. In Figure 2 the relative clause "which you bought" modifies the nominal "the book".
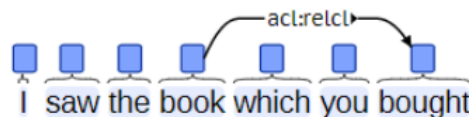

Figure 2: Example of relative clause modifier (*acl:relcl*).

Several treebanks for English are available [5] for the Universal dependency annotation (McDonald et al., 2013). We focused on the GUM Treebank (Zeldes, 2017), based on the Georgetown University Multilayer (GUM) Corpus [6] as its CoNLL-U format [7] contains a specific column that reports the dependency relation and the governor. Our next section analyses the accuracy of these two tags when labelling noun complement clauses and rel-

---

[5]https://universaldependencies.org/treebanks/en-comparison.html
[6]https://gucorpling.org/gum/
[7]an adaptation of the CoNLL-X format, (Buchholz and Marsi, 2006), https://universaldependencies.org/format.html

deprel	Train	Dev	Test
acl:that	65 (0.513)	13 (0.65)	14 (0.69)
acl:relcl	1419 (11.21)	258 (12.92)	216 (10.70)

Table 1: Frequency of "acl:relcl" and "acl:that" in the GUM Treebank files raw (normalized per 1000 tokens).

ative clauses in the development (DEV), training (TRAIN) and testing (TEST) sets of the GUM Treebank based on the GUM corpus (Levine, Lauren and Zeldes, Amir, 2017).

3 Revisiting the GUM Treebank

We noticed some debatable annotations for some cases where ellipsed *such* and *so* led some *that*-clauses expressing consequence to be labelled as acl as in *"As a result, wikiHow is still at the size that every editor eventually gets to know other editors"*. We computed the proportion of Relative Clauses (RC) in relation to noun complement clauses (NCC).

3.1 Frequency of RC and NCC in the GUM Treebank

As can be seen in Table 1, there are at least 15 times more relative clauses (RC) than noun complement clauses (NCC) in the GUM Treebank.

One of the benefits of the GUM Treebank is that it contains extra information, the ninth column conflates the dependency relation (acl) and *that* for noun complement clauses, we have tried to exploit this `acl:that` tag by building a UDPipe model based on this treebank and by trying to recapture this information by an algorithm.

3.2 Replicating the GUM Ninth Column

In the ninth column of the GUM corpus, we were specifically interested in the *"acl:relcl"* and *"acl:that"* annotations to improve the detection of noun complement clauses, since the standard deprel (dependency relation) column only provides the *"acl"* label and does not distinguish between finite and non finite uses of adnominal clauses. We trained a UDPipe model using the training, development and test sets of the GUM Treebank on Github[8]. However, once we applied the model on the same unannotated corpus, the ninth column

was empty. It seems that UDPipe only captures the standard columns of the treebanks.

3.3 Emulating the Ninth column

We were therefore interested in reconstructing this column by implementing a heuristic. Once the acl:relcl have been copied from the deprel column, the algorithm consists in exploiting the seventh (Head of the current word) and eighth (Universal dependency relation to the HEAD) columns such that:

Algorithm 1 : **Heuristic to emulate acl:that labels in the ninth column**

---

  for each $sentence \in corpus$ **do**
    for each $token \in sentence$ **do**
  1. Combine the seventh and eighth columns of the $token$ that were generated by the previously trained UDPipe model.
  2. If *"that"* is right after the word to which the seventh column of the $token$ points to, then add *"that"* to the ninth column.

---

4 Learning to tag with TreeTagger

This retagging experiment (Gaillat et al., 2014) relies on the ability of TreeTagger (Schmid, 1994) to be used not only as a POS-tagger but as a tool which can be trained to learn how to tag, provided a specific tagset and sample data are provided. We used samples from the Brown corpus in its NLTK distribution and modified the Penn Treebank tagset to distinguish *that* as WPR (relative pronoun) and *that* as CST (complementizer). In the learning phase, TreeTagger sees a vocabulary file and tokens associated to their tags and generates a .par model file to be used for POS-tagging. This section describes how we modified the tags to train the system [9]. After the annotation of the Brown corpus by UDPipe, a heuristic was applied on the results in order to introduce the *WPR* and *CST* tags which are not previously used in the tagset. To do that, the *DEPREL* label was used, so our method assumes that the UDPipe trained with the English GUM corpus provides a sufficiently correct *DEPREL* label for noun complement clauses:

The aim of this experiment is to see how the TreeTagger accuracy increases as a function of the

---

Algorithm 2 : **Heuristic for Brown re-annotation**

for each *sentence* ∈ *corpus* **do**
  for each *token* ∈ *sentence* **do**
- If the *token* is a verb (i.e. XPOS = VB) and is a clausal modifier of noun (i.e. DEPREL = acl) then go steps before that *token* to see if there is any *"that"*, if so, label it as *CST*.
- If the *token* is a verb (i.e. XPOS = VB) and is part of a relative clause (i.e. DEPREL = acl:relcl) then go steps before that *token* to see if there is any *"that"*, if so, label it as *WPR*.

deps column	Train	Dev	Test
acl:that	0.78	0.76	0.71
acl:relcl	0.92	0.92	0.94

Table 2: Accuracy of acl:relcl and acl:that annotations in the "deps" column recreated by combining the "head" and "deprel" columns for each of the GUM Treebank files.

training set size. To do this, the TreeTagger received different proportions of a training set as input. To be more specific, there are 500 training files representing the annotated Brown corpus, for the first training the first 10 files were used, and then the 30, 100, 200, 300, 400 and finally the 500 training files. For each training a *.par* file that corresponds to the model was returned.

5 Results

5.1 Emulating the Ninth column

To assess this algorithm that selects only *that-* (finite) clauses among the acl clauses, we tested it with the GUM treebank, comparing our results in our reconstructed column with the original data. The heuristic gave good results for the annotations of relative clauses "acl:relcl" with an accuracy that exceeds 90% (see table 2).

Nevertheless, the algorithm works less well for "acl:that", this is partly due to some coordinated NCC clauses and to multi-word-units (like *quid pro quo*).

5.2 Re-annotating with TreeTagger

We used our specifically designed testing files that contain respectively 189 *"that"* as *WPR* and 194 *"that"* as *CST*. The first one named

RCtest (Relative Clause) was used to compute the accuracy for *WPR* and the second one named NCCtest (Noun Complement Clause) for *CST* (see Figure 3 and 4). We used these specific files because they are manually annotated and each one of them contains a majority of the two tags we are interested in, which makes it convenient for our experiments.
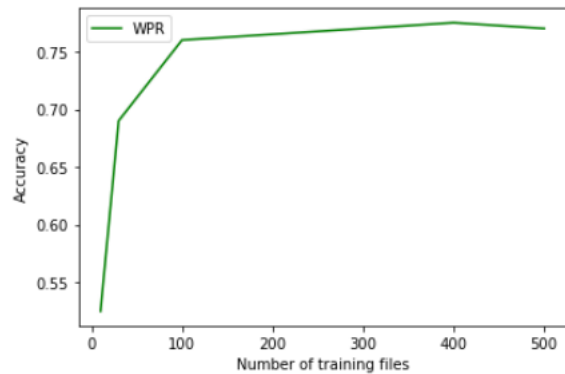

Figure 3: TreeTagger accuracy curve for *WPR* tag (computed using the RCtest data).
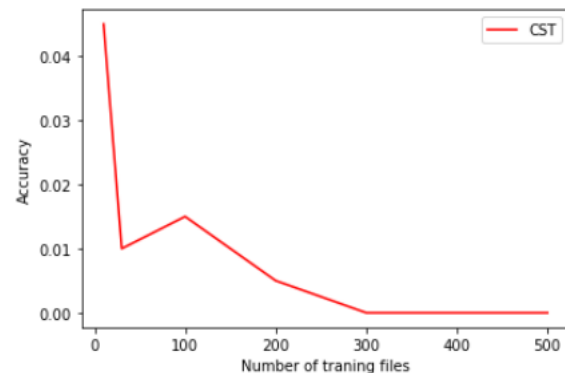

Figure 4: TreeTagger accuracy curve for *CST* tag (computed on the NCCtest data).

As shown in Figures 3 and 4 the TreeTagger accuracy increases with the number of training files for the RCtest data. This is a natural behaviour from a probabilistic model such as the TreeTagger, the probability increases as the weight of Relative clauses increases in the data it has. However, there is a drastic decrease in the *CST* curve around 100 training files, and the TreeTagger did not perform very well annotating the *"that"* with *CST* tag, as shown with Figure 4 the accuracy is very low, and as the number of training files increases, the accuracy goes down. Whatever technique is used,

the detection of noun complement clauses is more challenging than for relative clauses.

6 Discussion

Two approaches for explaining the results obtained for the CST tag can be either statistically or linguistically motivated. Starting with the first approach, as shown in Table 3, as the number of training files goes up, the number of other tags increases, especially for the IN tag, which in this case represents a confusion by the TreeTagger to annotate with the right tag, in fact IN is not a specific tag but rather a generic tag as it also corresponds to verbal *that*-clauses, therefore, this shows that the TreeTagger generated noise due to a confusion on the annotation of "that" (this is better illustrated in the figure 9 especially for the graph that represents the IN tag in blue.) The second approach consists in analysing the competing labels for *that*.

6.1 Accuracy in relation to other categories

The Penn Treebank tagset (Santorini, 1990), even though it does not acknowledge the whole complex range of functional realisations of *that*, e.g. adverbial, proform vs deictic uses, see (Ballier et al., 2022) can help visualise the complex interaction of the learning process of the identification of the different functional uses of *that*. As the training data increases, the variable proportions of the different functional realisations of *that* probably changes, so that a probabilistic tagger generates models variable in their results for this tagging task. The tagger has to learn the different competing tags for *that*. Our two test datasets allow us to monitor the evolution of the training phase as the size of the training data increases. Whereas we tried to train Treetagger to learn CST for NCC *that* and WPR for relative pronouns, we also computed the distribution of other tags that "*that*" may take, such as "WDT" (*that* when used as a relative pronoun, but also "WH"-determiners such as *which*), "DT" (Determiners), and "IN" (Subordinating conjunction, whether for nouns or for verbs) for each of the RC and NCC corpus. Table 3 recaps the changes observed when we evaluated the labels with our two testing sets (RCtest and NCCtest). For each testing set, we indicate the expected count of each label in the columns RCtest GOLD and NCCtest GOLD.

Here is an example of these potential mishaps in the POS-tagging: "that meeting that|IN [vs DT]

	RCtest	RCtest GOLD	NCCtest	NCCtest GOLD
10 training files				
WPR	107	189	20	17
CST	22	26	10	194
IN	95	0	183	0
DT	7	15	16	14
30 training files				
WPR	146	189	28	17
CST	5	26	3	194
IN	72	0	189	0
DT	8	15	9	14
100 training files				
WPR	158	189	25	17
CST	2	26	3	194
IN	65	0	194	0
DT	6	15	7	14
200 training files				
WPR	156	189	27	17
CST	1	26	1	194
IN	66	0	196	0
DT	8	15	0	14
300 training files				
WPR	157	189	22	17
CST	2	26	0	194
IN	67	0	202	0
DT	5	15	5	14
400 training files				
WPR	159	189	21	17
CST	2	26	0	194
IN	64	0	199	0
DT	6	15	7	14
500 training files				
WPR	158	189	23	17
CST	1	26	4	194
IN	65	0	188	0
DT	7	15	7	14

Table 3: Statistics about *WPR*, *CST*, *IN* and *DT* tags obtained for each of the 7 models (i.e. trained with 10, 30, 100, 200, 300, 400 and 500 files).
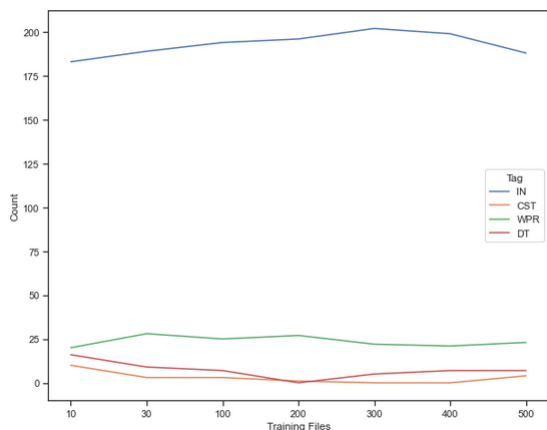
Figure 5: Evolution of IN, CST, WPR and DT tags with training files in the NCCtest corpus.
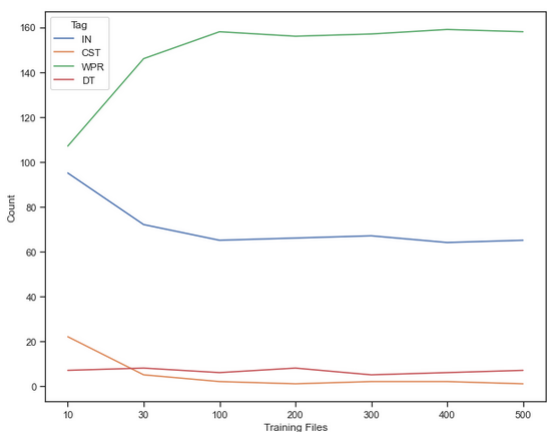

Figure 6: Evolution of IN, CST, WPR and DT tags with training files in the RCtest corpus.

morning was about a public case that|IN [vs WPR] we might make". The first deictic *that* was properly labelled, the second one was erroneously labelled as a subordinating conjunction and for the third occurrence, the relative pronoun was tagged as a subordinating conjunction (see additional examples in the Appendix).

6.2 Weakness of the TreeTagger-based heuristic

We re-annotated a corpus initially tagged with the Penn Treebank, which means that we modified some IN tags to CST and some IN tags to WPR for relative pronouns but the Brown corpus data retained some WDT labels. As shown in Table 3, there are many *WDT* tags, this is simply because the *WDT* tag is both an older and more general version of the *WPR* tag, and seemingly the TreeTagger kept the older version. So the *WDT* and *WPR* tags

are likely labels for relative pronouns considered as equivalent in the computing of the metrics, even though strictly speaking some WDT tokens in the Brown corpus may correspond to WH-determiners such as *which*. The main objection to our method is that we only relabelled a portion of the IN tags, so that the system has to learn a WPR versus CST distinction while still being fed with some examples of IN. In this sense, we can only partially monitor the behaviour of Treetagger when subjected to more examples. Figure 5 and Figure 6 plot the evolution of the tagging of the NCCtest and RCtest sets (respectively) as the corpus size increases. We expect the system to learn to relabel IN as either WPR or CST but this is hardly the case for CST. It should be noted that we did not control the input of the respective number of examples with CST and with WPR when increasing the data size of the training data. We only report the total counts of the tags assigned to *that*, we did monitor the individual behaviour of the tagging system for each occurrence of *that*.

6.3 Long-Distance Dependencies

As already pointed out, noun complement clauses can follow a relative clause for the same noun (but not the other way round). *That*-relative clauses tend to be adjacent to their antecedents (and are often restrictive relative clauses) whereas (*that*-) noun complement clauses can be separated from their governor. So we explored a simple metric which is the distance (i.e. number of tokens) separating a *"that"* (annotated either with *CST* or *WPR*) and the last noun before it. As shown in the boxplots in Figure 7 there is a tendency showing that the *"that"* tagged with *CST* using a verb with a *DEPREL = acl* have a higher distance separating them from the last noun before them. This can probably cause some ambiguity due to the higher distance. However, as we can see for the *"that"* tagged with *WPR* using a verb with a *DEPREL = acl:relcl* the distance with the last noun is smaller, and there are less misclassifications (i.e. less noise) for the *"that"* used as *WPR*. This is just a statistical approach to see if there is any bias that can explain why the heuristic produces a lot of noise.

Our metric is rather crude but head nouns of NCCs need not be adjacent to the *that*-clauses, so that an inventory of structures in-between could be taken into account. The distance between the governor and the *that*-clause of these long-distance
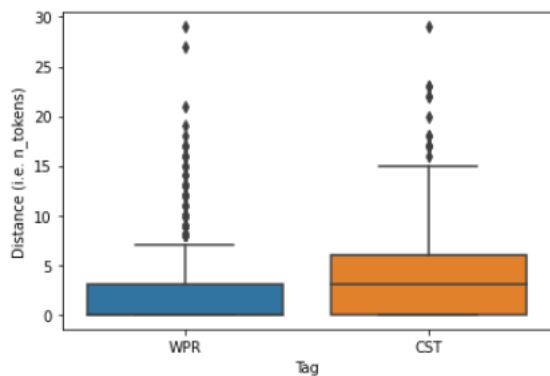
Figure 7: Distribution of the distance (number of tokens) separating a *"that"* and the last noun before it for each of the *WPR* and *CST "that"*.

dependencies (Osborne, 2019) could be more systematically investigated.

6.4 Relevance of UD Deprel Labels for NCC?

It should be noted that UD changed the dependency label for noun complement clauses, as explained on the UD website: "In earlier versions of SD/USD, complement clauses with nouns like fact or report were also analyzed as ccomp [clausal complement]. However, we now analyze them as acl. Hence, ccomp does not appear in nominals. This makes sense, since nominals normally do not take core arguments." We may challenge this view since ccomp implies a "clausal complement" and nouns may require a "core argument", even more so than for adjectives.[10] One of the unfortunate consequences is that adverbs like *now* in the sentence "Now that the world is in the age where lighting seems to be a daily necessity" are labelled as a governor of the "adnominal" clause. It maybe the case that acl is a debatable label, also used after verbs as for that verbal complement clauses ("if this seems incredibly far-fetched, comfort yourself that double chute failure in modern times is also extremely unlikely, and that you have already beaten worse odds"). Consequently, the (SUD) Surface-Syntactic Universal Dependencies (Gerdes et al., 2018) has suggested alternative labels for acl. Another approach might be to restrict noun complement clauses to a subcategory of acl specific to noun complement clauses (possibly labelled as acl:ncl).

---

[10]For a similar argumentation see (Osborne and Gerdes, 2019).

7 Further Research

7.1 Quality Monitoring of the Training Phase

We have only estimated the accuracy of the annotation on our testing sets but we have not monitored the qualitative aspect of the annotation. Are some sentences systematically mislabelled or can we observe some changes during the training phase? For example, this NCC gets to be interpreted as a relative clause: *"O'Neill had an emotional reaction that [tagged as WPR] the level of corruption was too high to do serious projects in Russia," Deripaska recalls.* Some configurations seem to remain challenging for parsing, and qualitative monitoring of the accuracy should take into account these sentences for which labelling improves or not. Controlling for frequency of exposure in the training data should prove to be very fruitful to maybe detect thresholds in frequency (or proportions) in the training data for accurate tagging. For example, an example in our appendix seems to suggest that a trigram sequence no/N/that (and corresponding identification of noun complement clauses) seems to be learned after exposure to the 100 training files (36 occurrences). As some of the examples of mislabellings in the Appendix also suggest, it is likely that our relabelling algorithm for WPR is too greedy, and a more elaborate version should filter out alternative relative pronouns that should inhibit the relabelling process. We should also apply stricter conditions on the type of *that* which can be re-tagged. Assuming the DT label is correct, only IN labels should be re-tagged.

7.2 More data?

More training files from the Brown corpus have been manually annotated and given to the TreeTagger, and an improvement in the *CST* accuracy was observed (see Figure 8). Though a plateau seems to be observed for the tag CST (*that* for NCC complementizer), one may wonder if more examples of NCCs in the training data would alter this curve. We have only analysed the GUM Treebank for the UD analysis, but no less than six treebanks are available on github for the Universal Dependency analysis of English.

7.3 Learnability and Dispersion

Our monitoring of the learning curve of the tag distinction in our TreeTagger experiment could be finer-grained: we did not control for genre types within the Brown corpus and the relative distribu-
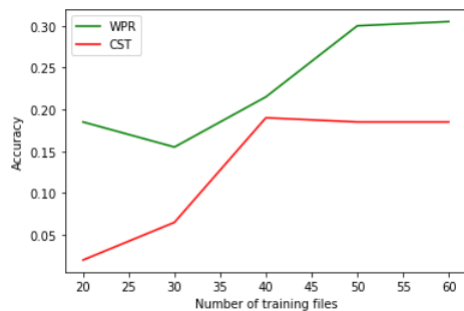
Figure 8: TreeTagger accuracy for *"that"* annotated with *CST* in red and *WPR* in green with more training files.

tions of the two structures. If relative clauses seem to be more frequent than NCCs in the GUM Treebank, NCCs are more likely to be more frequent in argumentative texts (Ballier, 2007). Our experiment only reported the effect of the number of the Brown files in the training data, not the specific distribution of the two structures across the different registers of the Brown corpus. The dispersion of these linguistic structures in the training data could be monitored across the corpus subparts using adequate dispersion measures (Gries, 2020) or by comparing the vocabulary growth curves (Evert and Baroni, 2007) of the two constructions across the Brown corpus files. Our Figure 9 crudely plots the distribution of the different tags in the training data as the size of the corpus increases (measured in number of files, but not with the corresponding text genres). Increasing the size of the corpus may require more attention to a frequency/textual diversity trade off.

8 Conclusion

In this study, we have experimented two methods to detect noun complement clauses, either by using the universal dependency GUM treebank or by retagging the Brown corpus with specific *WPR* and *CST* tags. We also explored an automated way to do this annotation using a specific heuristic. We have evidenced the longer distance between the noun and the *that*-clause for noun complement clauses. The detection of relative clauses does seem to be much more robust than for noun complement clauses, which remains a problem for information retrieval as text genres could be interestingly classified with this criterion. The difference in frequency and in adjacency may account for such a discrepancy in the accuracy of the identification of the clause type. We have only begun to explore the

parameters of the learnability of these tags corresponding to such a subtle linguistic distinction.

Appendix

8.1 Example of a noun complement clause where that gets properly tagged after 100 files in the training data (containing 36 occurrences of the *no N that* sequence)

*"However, there is no guarantee **that**[tagged as CST] only the genuine repentant will produce works of value to the society."*

8.2 Examples of remaining errors in our test sets

We include examples of persistent mislabelling in our test data. After 500 training files, 6 sentences with that in noun complement clauses are still tagged as if they were relative pronouns (with WPR).

- *The statement that|WPR the tribunal has made an "error of law" means no more or less than that|CST the construction placed upon the term by the court is preferred to that|DT of the tribunal.*

- *There was no dispute that|WPR Bunn throughout acted with the authority of the bank.*

- *This included a commitment that|WPR "if one of the two states should become the target of aggression, then the other side will give the aggressor no military aid or other support".*

- *We have received information that|WPR today, between 1400 and 1500, there was an explosion at the residence of Seyed Ali Khamenei.*

- *Recently there was the illusion that|WPR Hamas, while not a perfect partner, was at*
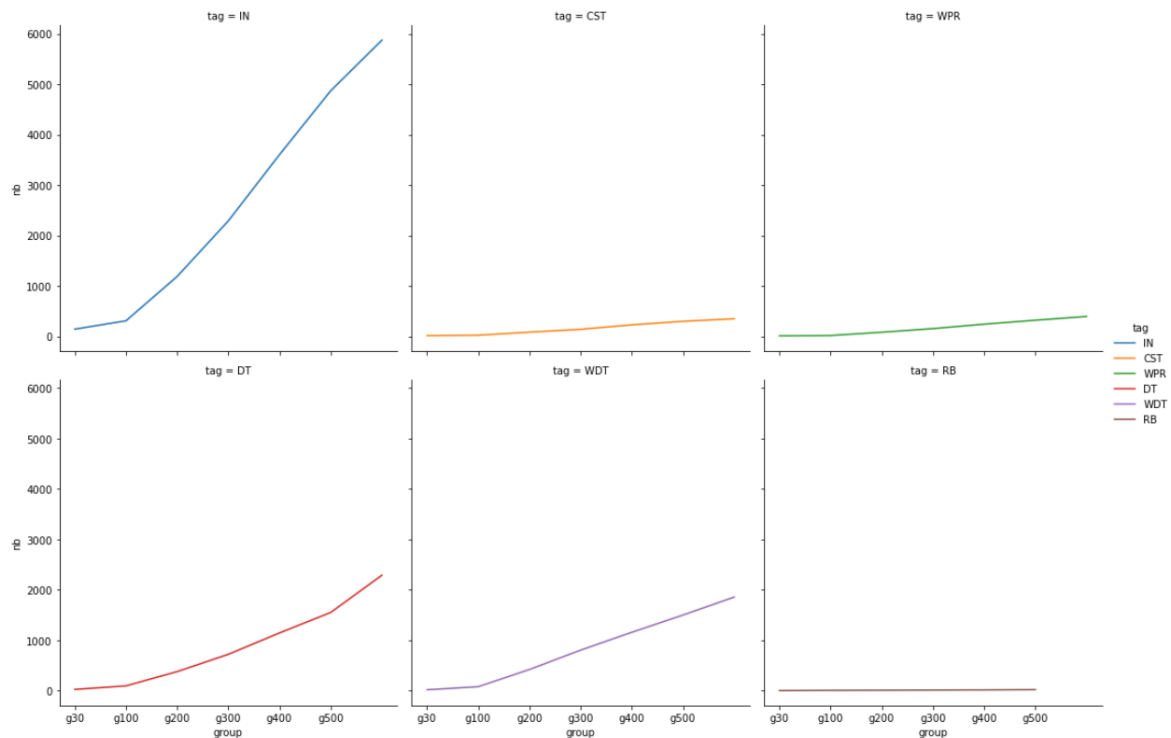
Figure 9: Evolution of the number of different tags for the re-annotated Brown corpus file groups (10, 30, 100, 200, 300, 400, 500 files.)

least a group that could implement decisions,"
he said.

- *Where there is a contract for the sale of goods by description, there is an implied condition **that**|WPR the goods shall correspond with the description.*

8.2.1 Example from our test sets that has been annotated with DT rather than with CST

We illustrate the complexity of the polyfunctionality of *"that"* by showing an example of overfitting for the deictic/pronominal uses of *"that"*.

*"A high-ranking official in the Clinton administration expressed shock **that**[tagged as DT rather than CST] "the kids" in the White House "did not stand up when the president entered the room."*

8.2.2 Examples from the RC test set that have been annotated with IN rather than with WPR

- *High death rates among children reduce the value **that** |IN parents place on education; and so on.*

- *The distinction **that** |IN matters is from that of 'patronage', which itself, as we shall see, is*

highly varied.

8.2.3 Examples from the NCCtest set that have been annotated with IN rather than with CST

- *They're living proof **that** asthma can be passed from generation to generation.*

- *Where there is a contract for the sale of goods by description, there is an implied condition **that** the goods shall correspond with the description.*

8.3 An example of false positives for the Brown relabelling heuristic

- *"... But one does not have to affirm the existence of an evil order irredeemable in **that**[tagged as WPR] sense, or a static order in which no changes will take place in time, to be able truthfully to affirm the following fact: there has never been justitia imprinted in social institutions and social relationships except in the context of some pax-ordo preserved by clothed or naked force ..."* (it should be *DT* rather than *WPR*). The relative clause is with WHICH, not with THAT.

References

Nicolas Ballier. 2004. *Praxis métalinguistiques et ontologie des catégories*. Habilitation thesis in linguistics, Université de Paris X-Nanterre.

Nicolas Ballier. 2007. La complétive du nom dans le discours des linguistes. *D. Banks, La coordination et la subordination dans le texte de spécialité, Paris, L'Harmattan*, pages 55–76.

Nicolas Ballier, Antonio Balvet, Taylor Arnold, and Thomas Gaillat. 2022. *Some metalinguistic assumptions behind tagsets for English: evidence from that in different versions of the Brown corpus*, pages 27–81. Peter Lang, Bern.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Rhonwen Bowen. 2005. Noun complementation in english : a corpus-based study of structural types and patterns.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164.

Maggie Charles. 2007. Argument or evidence? disciplinary variation in the use of the noun that pattern in stance construction. *English for Specific Purposes*, 26(2):203–218.

BNC Consortium et al. 2007. British National Corpus XML edition. *Oxford Text Archive. http://hdl.handle.net/20.500*, 12024:2554.

Stefan Evert and Marco Baroni. 2007. zipfR: Word frequency distributions in R. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 29–32.

Thomas Gaillat, Pascale Sébillot, and Nicolas Ballier. 2014. Automated classification of unexpected uses of *this* and *that* in a learner corpus of english. *Language and Computers*, 78:309–324.

Roger Garside. 1987. The CLAWS word-tagging system. *The Computational analysis of English: A corpus-based approach. London: Longman*, pages 30–41.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal dependencies workshop 2018*.

Stefan Th. Gries. 2020. Analyzing Dispersion. In *A Practical Handbook of Corpus Linguistics*, pages 99–118, Cham. Springer International Publishing.

Rodney Huddleston. 1984. *Introduction to the Grammar of English*. Cambridge University Press.

Issa M Kanté. 2017. Étude sémantico-syntaxique de la complétive nominale en anglais et en français. *Étude sémantico-syntaxique de la complétive nominale en anglais et en français*.

Henry Kucera, Henry Kučera, and Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press.

Levine, Lauren and Zeldes, Amir. 2017. GUM: The Georgetown University multilayer corpus.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Timothy Osborne. 2019. *A dependency grammar of English: An introduction and beyond*. John Benjamins Publishing Company.

Timothy Osborne and Kim Gerdes. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics*, 4(1):1–28.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the penn treebank project (3rd revision). *Technical Reports (CIS)*, page 570.

Hans-Jörg Schmid. 2000. *English abstract nouns as conceptual shells*. De Gruyter Mouton, Berlin.

Helmut Schmid. 1994. TreeTagger-a language independent part-of-speech tagger. *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/*.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD Shared Task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

University Centre for Computer Corpus Research on Language. 1995-2004. Constituent likelihood automatic word-tagging system 8.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.