Improving the parsing of French coordination through annotation standards and targeted features

Assaf Urieli
CLLE-ERSS
Université de Toulouse
assaf.urieli@univ-tlse2.fr
Joliciel Informatique
Foix, France
assaf@joli-ciel.com

Abstract

In the present study we explore various methods for improving the transition-based parsing of coordinated structures in French. Features targeting syntactic parallelism in coordinated structures are used as additional features when training the statistical model, but also as an efficient means to find and correct annotation errors in training corpora. In terms of annotation, we compare four different annotations for coordinated structures, demonstrate the importance of globally unambiguous annotation for punctuation, and discuss the decision process of a transition-based parser for coordination, explaining why certain annotations consistently out-perform others. We compare the gains provided by different annotation standards, by targeted features, and by using a wider beam. Our best configuration gives a 37.28% reduction in the coordination error rate, when compared to the baseline SPMRL test corpus for French after manual corrections.

1 Introduction

Coordinated structures (CS) are recognised as one of the main difficulties for automatic syntax parsers. They are particularly challenging for transition-based parsers, which operate sequentially from sentence start to end: indeed, even for a simple coordinated structure, it is virtually impossible to determine the first conjunct of the structure without examining the rest of the sentence. Consider the following three sentences, identical in French up to the coordinating conjunction:

Example 1.1 - J'ai mangé une pomme rouge et mûre. *(I ate a red and ripe apple)*
- J'ai mangé une pomme rouge et une orange. *(I ate a red apple and an orange)*
- J'ai mangé une pomme rouge et Georges a bu du thé. *(I ate a red apple and George drank some tea)*

In the above cases, selecting the correct conjuncts is simply a matter of examining the parts-of-speech immediately following the coordinating conjunction, except in the last case, where we have to decide whether or not George gets eaten. Nevertheless, nothing preceding the conjunction can help us make the decision. Often the situation is more complex, with adjuncts intervening between the conjunction and second conjunct, not to mention cases such as various forms of ellipsis, CSs with 3 or more conjuncts, and modifiers shared by two or more conjuncts.

In this article, after reviewing related work (section 2) and introducing CS annotation and transition-based parsing (section 3) and our data set and software (section 4), we follow a chronological outline in terms of our own research. In a previous study (Urieli, 2014) we successfully applied knowledge-rich targeted features to the pos-tagging of ambiguous functional words. In the present study we turn to parsing (section 5.1), and attempt to apply knowledge-rich targeted features for coordination to the SPMRL 2013 dependency corpus for French (Seddah et al., 2013). Although the results are not fully satisfactory, we discover while tuning the features that they can be very useful for pinpointing and correcting many of the coordination errors in the training and evaluation corpora (section 5.2). Also, while exploring the reason behind failure to coordinate correctly, we note that the way in which coordination is annotated in the corpus is responsible for a sizable proportion of errors. We then attempt automatic transformations

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Page numbers and proceedings footer are added by the organisers. Licence details: http://creativecommons.org/licenses/by/4.0/.

of this annotation and compare results for six different annotations (section 5.3). Finally, we combine annotation schemes with targeted features and a wider beam to find the maximal gain that can be attained (section 5.4).

2 Related work

Several studies have explored the annotation standards for coordination in English. However, the original Penn Treebank annotates only a subset of simple coordinated structures implicitly by grouping the items together in a single phrase. Maier et al. (2012) present an annotation scheme for Penn Treebank coordination which includes punctuation, critical in the case of constituency treebanks. They then (Maier and Kübler, 2013) train a classifier to attempt to recognise coordinating vs. non-coordinating commas, and achieve an f-score of 89.22 for the coordinating (difficult) class. Many of the phenomena they are trying to disambiguate in the constituency treebank by annotating punctuation are disambiguated in dependency treebanks more simply by using an appropriate set of dependency labels, e.g. in the case of apposition vs. coordination. In the present study, we thus take a somewhat opposite approach by removing annotation from punctuation in the dependency treebank context, in order to concentrate the decision-process on the conjuncts themselves.

Ivanova et al. (2013) measure performance for English using three different annotations for coordination, all of which are covered by the present study. With respect to annotation, they come to similar conclusions for English to ours for French, but concentrate on the lowest-accuracy conjunction-headed approach, as it is proned by the grammar-based parser in which they specialize.

Popel et al. (2013) perform a survey of many different dependency annotations for coordination, and develop a tool for lossless transformation between these annotations. They also describe in detail the various difficulties involved in annotating coordination, including the role of punctuation.

Schwartz et al. (2012) compare the "learnability" of various possible annotations for 6 structures in English, including coordination, where learnability is defined both by the annotation giving the highest attachment accuracy, and by the annotation which attains a target accuracy with the fewest training examples. They compare 2 possible annotations for coordination, and find, as we do, that using one of the conjuncts as head is far more learnable than using the conjunction as the head, across a variety of parsers. However, since the Penn Treebank does not annotate coordinated structures with more than 2 conjuncts, they explore fewer annotation possibilites than in the present study.

Tsarfaty et al. (2011) raise a similar question of evaluating parsers trained on different annotation standards, including for coordination, but take a radically different approach. They convert all annotations to directly comparable generalised functional trees, and find that apparent major differences in performance are considerably attenuated or disappear when considered in such a light. It would be interesting to apply their method to our different annotations for French data, and see to what extent it affects results.

In terms of annotation standards, the present study extends previous work by (a) applying similar experiments to French and consolidating certain conclusions, while concentrating on the case of 3 or more conjuncts, (b) highlighting the importance of a systematic annotation for punctuation, which is only possible when punctuation is not explicitly used to indicate coordination, and (c) comparing gains from annotation changes to those made by the addition of targeted coordination features or using a wider beam.

In terms of specific targeted features for coordination, Hogan (2007) achieves statistically significant improvements in noun phrase (NP) coordination in English, in the context of a history-based constituency parser, by introducing features for NP head semantic similarity. Shimbo and Hara (2007) leave out semantics, and instead use features incorporating the syntactic "edit-distance" between competing structures. Both studies apply to consitituency parsers with a higher complexity than our linear transition-based parser.

Other studies have attempted introducing generic "rich" features without specifically aiming at parallelism in coordination. Kübler et al. (2009) propose a method whereby the n-best PCFG parses are reranked, in order to improve the parsing of coordination in German. Their features are generic, but can cover the full parse trees since they are applied in reranking rather than during parsing. Our study differs

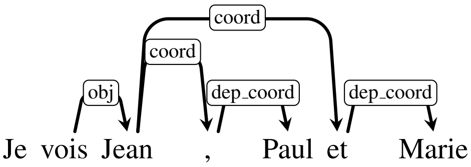


Figure 1: French SPMRL annotation for coordination

from theirs by applying the features during the first parsing pass of a linear-complexity transition-based parser, rather than requiring a large beam of n-best solutions at the outset.

Zhang and Nivre (2011; 2012) have shown the usefulness of generic "rich" features such as the valency (number of dependents) of a given token, the distance between two tokens, a list of current unique modifier labels for a token, and features looking at various characteristics a token's second order governor (its governor's governor). They find that these features are particularly useful for global-learning based parsers with a very high beam width (64)—this however comes with certain practical disadvantages, since parsing speed is linearly correlated to beam width, and analysing with a beam of 64 takes 64 times as long. In our study, we do not explore beam widths beyond 5 and do not apply global learning, and show nevertheless that highly specific targeted features give considerable gain even in such a context.

In terms of French, De la Clergerie (2014) introduces rich symbolic features into statistical transition-based parsing indirectly, by parsing each sentence first using FRMG, a TAG parser, and injecting features based on the FRMG parse into the transition-based parser. He attains excellent results for French (LAS=90.25 for the SPMRL `test` corpus with guessed pos-tags). However, given the need to parse with a TAG parser, this system is not directly comparable to linear-time transition-based parsing.

3 Annotations and analysis mechanisms

Let us consider the following sentence containing a 3-conjunct coordinated structure:

Example 3.1 Je vois Jean, Paul et Marie. *(I see John, Paul and Mary)*

Figure 1 shows the French SPMRL dependency annotation for this sentence: all conjuncts are governed by the first conjunct via the preceding comma or conjunction.

The next question is: how is such an annotation parsed? In this study we concentrate purely on transition-based parsing (Kübler et al., 2009). Parsing is thus defined as a series of transitions leading from one parse configuration to the next, where a parse configuration is defined as follows:

- σ: a stack, or ordered sequence of tokens which have been partially processed
- β: a buffer, or ordered sequence of tokens which have not yet been processed
- Δ: a set of dependency arcs of the form *label(governor, dependent)* that have already been added
- τ: a sequence of transitions allowing us to reach the current configuration from an initial one

We will use σ_0 to indicate the token currently on top of the stack, and $\sigma_{1..n}$ for tokens deeper in the stack. Similarly, β_0 indicates the next token to be processed on the buffer, and $\beta_{1..n}$ for tokens farther down the buffer. Parsing begins with *root* artefact on the stack and all other tokens on the buffer. Parsing ends when the buffer is empty. Our study uses the arc-eager transition system (Nivre, 2008), which defines the four transitions shown in table 1 for moving from one configuration to the next.

It is well known that transition-based parsers tend to favour short-distance dependencies over longer distance ones (McDonald and Nivre, 2007; Candito et al., 2010), since they will always compare two closer tokens before comparing two tokens which are farther away, and the decision regarding the two closer tokens is taken independently given the information available at this point. Thus, a closer token is never directly compared to a token farther away when making an attachment decision. This tendency can be somewhat curtailed by applying a beam search (Urieli and Tanguy, 2013).

transition	**effect**	**precondition**
left-arc$_{\text{label}}$	Create the dependency arc *label*(β_0,σ_0) and pop the stack	The reverse dependency *any*(σ_0,β_0) does not exist, and σ_0 is not the *root* node
right-arc$_{\text{label}}$	Create the dependency arc *label*(σ_0,β_0), and push the head of the buffer to the top of the stack	
reduce	Pop the top of the stack	The top-of-stack has a governor
shift	Push the head of the buffer to the top of the stack	

Table 1: The arc-eager transition system for shift-reduce dependency parsing

Now, as already seen in example 1.1, forward-looking features are required to correctly identify the first conjunct by guessing the second conjunct. Table 2 shows the exact sequence of transitions required for parsing the 3rd example sentence from example 1.1, from the moment when we first encounter the coordinating conjunction on the buffer to the moment when the CS itself has been fully parsed. Difficult decisions are shown in bold. Among these, the `reduce` transitions on lines n+1 and n+2 both require us to look farther down the buffer to guess the most likely second conjunct, since we can only reduce when there are no more dependents to be attached. The `shift` transitions on lines n+4 and n+5 are simpler, since we already know that the first conjunct is a verb. Still, we have to recognise that the second verb is composite, which is governed by convention by the past participle rather than the helper verb.

	transition	**stack**	**buffer**	**dependencies added**
n		*root*, mangé, pomme, rouge	et, Georges, a, bu, du, thé	
n+1	**reduce**	*root*, mangé, pomme	et, Georges, a, bu, du, thé	
n+2	**reduce**	*root*, mangé	et, Georges, a, bu, du, thé	
n+3	right-arc$_{\text{coord}}$	*root*, mangé, et	Georges, a, bu, du, thé	coord(mangé,et)
n+4	**shift**	*root*, mangé, et, Georges	a, bu, du, thé	
n+5	**shift**	*root*, mangé, et, Georges, a	bu, du, thé	
n+6	left-arc$_{\text{aux\_tps}}$	*root*, mangé, et, Georges	bu, du, thé	aux_tps(bu, a)
n+7	left-arc$_{\text{suj}}$	*root*, mangé, et	bu, du, thé	suj(bu, Georges)
n+8	right-arc$_{\text{dep\_coord}}$	*root*, mangé, et, bu	du, thé	dep_coord(et, bu)

Table 2: Arc-eager transition sequence for coordination, with difficult decisions in bold

The case of a CS with 3 or more conjuncts is even more complicated, since it requires lookahead features for the first two conjuncts, looking farther ahead than in the case of the 2-conjuct CS. In all cases, correctly guessing the final conjunct ahead of time is critical information to correctly annotating the coordination.

4 Data and software

4.1 Talismane

All of the experiments in this study use the Talismane parser[1]. Talismane (Urieli, 2013) is an NLP toolkit including a sentence detector, tokeniser, pos-tagger and transition-based parser. All four modules use a statistical supervised machine learning approach, and it is possible to apply a beam search to the last three modules, as well as defining sophisticated features and rules using an expressive feature definition syntax. For all experiments in the present study, we used a linear SVM model with $C = 0.25$ and $\epsilon = 0.01$. We applied a cutoff of 5, so that a feature has to appear at least 5 times in the training corpus to be considered.

4.2 French Treebank

The original input for this study is the dependency annotation for the French section of SPMRL (Seddah et al., 2013), itself derived from the French Treebank (Abeillé et al., 2003), via an automatic conversion of constituency structures to dependencies. We use the `train` (14,759 sentences, 412,879 tokens), `dev` (1,235 sentences, 36,272 tokens) and `test` (2,541 sentences, 69,922 tokens) divisions of this corpus as defined for SPMRL. All of our studies use the gold pos-tags from the treebank, in order to make an abstraction of pos-tagger errors and concentrate on parsing. The baseline LAS excluding punctuation is 89.57% (`dev`) and 89.45% (`test`). The baseline f-score for coordinated structures, calculated as the f-score for all individual coordination arcs, is 84.35% (`dev`) and 85.16% (`test`).

4.3 Initial error classification

We began this study by analysing coordination errors performed by Talismane in the `dev` corpus. Out of 240 errors analysed, 24% were annotation errors (of which over 60% were correctly annotated by Talismane), 14% were artefacts of the annotation scheme (the 2nd and 3rd conjunct were directly coordinated by Talismane unlike the original annotation), and 30% were errors where Talismane coordinated two different pos-tags, whereas the correct coordination involved the same pos-tag. If we group this together with other cases of simple parallelism (e.g. cases where Talismane coordinated different prepositions instead of the same prepostion), this climbs up to 38%. The remaining 24% covered various difficult cases, including elliptical coordinations. Only 12% involved cases where semantics were required to make the correct choice.

The cases where the mildly rich French morphology might help us are very rare: only three cases among the `dev` corpus errors. In the examples below and elsewhere in this article, the guessed conjuncts are shown in *italics* (non-italics for the English translation), the correct conjuncts are underlined, and the conjunction is shown in **bold**. In the first example, the feminine demonstrative pronoun *celle* indicates that we are coordinating with the feminine noun *présidence* rather than with *M. Michel Albert*:

Example 4.1 [. . .] on avait parlé de la présidence des AGF à la place *de M. Michel Albert* **ou** *de celle* du GAN occupée par M. François Heilbronner. *(. . . they spoke of the presidency of the AGFs instead* of Mr Michel Albert ***or*** of that *of the GAN occupied by Mr François Heilbronner.)*

In the second case, the masculine past participle *rejeté* should coordinate with the masculine past participle *opté* rather than the feminine *faite*:

Example 4.2 Le conseil d'administration [. . .] a opté pour la proposition de reprise *faite* par Bongrain **et** *rejeté* celle de Besnier. *(The board of directors chose the takeover proposal* made *by Bongrain **and*** rejected *the one made by Besnier.)*

In the final example, a plural adjective *répétitifs* is coordinated with a plural adjectival past participle *construits*, rather than a previous morphologically unadorned past participle *découvert* in a conjugated construction:

Example 4.3 [. . .] les Européens ont *découvert/VPP* l'immensité du stock japonais : [. . .] scénarios répétitifs/ADJ **mais** habilement *construits/VPP* [. . .] *(the Europeans* discovered *the immensity of the Japanese stock: repetitive **and** skillfully* constructed *scenarios. . .)*

[1] `http://redac.univ-tlse2.fr/applications/talismane.html`

Because of the rarity of such cases, we decided not to include morpholigical features in our experiments.

5 Experiments

5.1 Initial experiment with targeted features

We first decided to target the 38% of errors relating to simple parallelism (e.g. parallelism errors related to mismatched pos-tags or prepositions, rather than semantics).

Because of the importance of identifying a second conjunct before identifying the first one, we first constructed the following targeted feature:

- **Second conjunct identification:** attempts to correctly identify the second conjunct. Since all subsequent features depend on this second conjunct feature, it was critical to attain high accuracy. Also, since the feature is a component of features used to select the first conjunct, it can only make use of information available when a first conjunct candidate is at σ_0 and the conjunction at β_0 (steps 1, 2 and 3 in table 2): critically, it tries to guess the second conjunct with no knowledge of the correct first conjunct.

The most difficult cases for this feature are verbs, since both coordinated verbs need to be outside of subordinate, relative or comment phrases. Comment phrases, particularly numerous in journalistic text, and marked only by punctuation, word order, and lexical choices, are the most difficult to recognise. The following list shows examples of sentences with two conjugated verbs (in italics), and with the conjuncts underlined.

1. **Verb coordination**: Il s'*agit* ici d'un jour normal de la semaine **et** un inventaire scrupuleux *exigerait* que l'on prenne également en compte l'offre accrue du mercredi. *(We are* dealing *here with a normal weekday,* ***and*** *a scupulous inventory would* require *us to take into account the increased offer on Wednesdays.)*

2. **Verb coordination**: Les chiffres parlent d'eux-mêmes : les Japonais *occupent* 30 % du marché américain **et** leurs exportations *représentent* près de 75 % du déficit commercial global annuel. *(The numbers speak for themselves: the Japanese* occupy *30% of the American market* ***and*** *their exports* represent *almost 75% of the annual global commercial deficit.)*

3. **Comment phrase**: A Lourdes, nous *signale* notre correspondant Jean-Jacques Rollat, la venue **et** la circulation des pèlerins ont été très *perturbées*. *(At Lourdes,* signals *our correspondent Jean-Jacques Rollat, the arrival* ***and*** *circulation of pilgrims was considerably* disrupted*.)*

4. **Relative clause**: Les émissions d'éveil qui ont *fait* la richesse des chaînes de service public entre 1975 **et** 1985 ont toutes *disparu*. *(The discovery programmes which* constituted *the richness of public channels between 1975 and 1985 have all* disappeared*.)*

We tested this feature on the training corpus, by applying it whenever a conjunction was found in σ_0, and seeing how often it correctly guessed "true" when the token in β_0 was the second conjunct, and "false" when the token in β_0 was not the second conjunct, while ignoring knowledge of the first conjunct. The accuracy for the "true" result is 99.07%, and for the "false" result is 94.54%.

We then used this feature to construct various features attempting to recognise parallelism in CS within the framework of transition-based parsing. Most of these features compare the item currently at the top-of-stack to the second conjunct guess, and check to see if there is a better candidate deeper in the stack. The following features were used:

- **Pos-tag mismatch:** if the first conjunct candidate at the top-of-stack has a different pos-tag from the second conjunct guess, does a candidate with the same pos-tag exist deeper on the stack?

- **Mismatched prepositions:** if the first candidate at the top-of-stack and the second conjunct guess are two different prepositions, does the same preposition exist deeper on the stack?

- **Pos-tag match:** if the first conjunct candidate at the top-of-stack is the same pos-tag as the second conjunct guess, are there any other candidates with this pos-tag deeper on the stack?
- **3 conjunct parallelism:** when two tokens of the same pos-tag, separated by a comma, are being compared, is the second token followed by a coordinating conjunction and then a third token with the same pos-tag as the first two? We allow for various intervening modifiers depending on the pos-tag being considered.
- **Parentheses:** is the first conjunct candidate at the top-of-stack inside parentheses and the second conjunct guess outside of them?

When we first attempted to apply these features to our dev (and test) corpora, our f-score for coordination (coord and dep_coord combined) improved from 84.34% to 85.52% (85.16% to 86.97% for test), giving a fairly modest error reduction of 7.54% (12.20% for test). In terms of significance, McNemar's test gives a p-value < 0.001 for coordination label changes in both dev and test.

Now, there are of course cases in the training corpus with valid non-parallel structures, such as the following coordination between an adjective and prepositional phrase:

Example 5.1 Au mieux, la reprise sera lente/ADJ **et** de/P faible ampleur. *(At best, the recovery will be slow **and** of limited extent.)*

These, however, are few and far in between when compared to the very large number of errors concerning clear pos-tag parallelism. We will examine some errors introduced by applying targeted parallelism features to non-parallel CSs in our final error analysis found in section 5.4.

5.2 Improvements through manual correction

The targeted feature definition involved several iterations in which features were projected onto the training corpus, and any unexpected results were analysed. Among the unexpected results were a very large number of annotation errors. Given that 24% of the original errors in the dev corpus were annotation errors, and our efficient method for pinpointing and correcting such errors by projecting targeted features, we decided to apply these targeted manual corrections to the entire SPMRL French corpus (train, dev and test).

Specifically, these manual corrections involved:

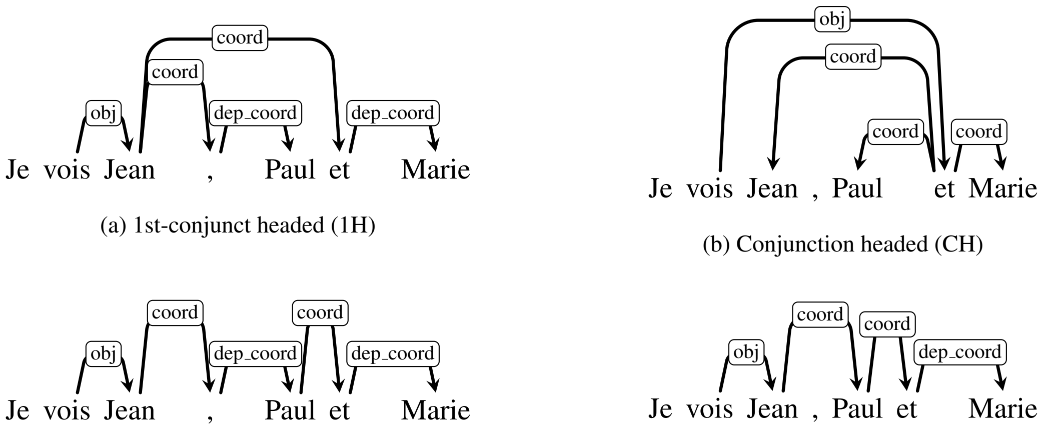
- Fixing any coordination where the dependent preceded the governor (impossible in the original annotation standard)
- Reviewing and standardizing all cases of *ni…ni…* (neither…nor…) and *soit…soit…* (either…or…).
- Projecting the above targeted features onto the corpus via Talismane, and correcting any items where the feature yielded unexpected results.

The total corrections are 1,488 for train (out of 21,061 coordination relations = 7.07%), 106 for dev (out of 1,743 coordination relations = 6.08%) and 274 for test (out of 3,420 coordination relations = 8.01%). Multi-word expressions (MWEs) were left as is, except on rare cases where a modifier inside the MWE was coordinated to a modifier outside of it.

	dev base	**dev fix**	**test base**	**test fix**
train base	*84.34*	85.08	*85.16*	85.54
train fix	83.99	**85.75**	84.99	**86.75**

Table 3: Coordination f-score after targeted manual error correction

Table 3 shows the coordination f-score with and without targeted error correction in both training and evaluation. Fixing errors in the training corpus is only useful when equivalent errors are fixed in the


coord
coord
obj
dep_coord
dep_coord
Je vois Jean , Paul et Marie
obj
coord
coord
coord
Je vois Jean , Paul et Marie
coord
coord
obj
dep_coord
dep_coord
Je vois Jean , Paul et Marie
coord
coord
obj
dep_coord
Je vois Jean , Paul et Marie

(a) 1st-conjunct headed (1H)

(b) Conjunction headed (CH)

(c) Previous conjunct headed (PH)

(d) Previous conjunct headed 2 (PH2)

Figure 2: Different annotations for coordination

evaluation corpora. If we consider the corrected evaluation corpora only, fixing errors in the training corpus gives an f-score error reduction of 4.49% for `dev` (8.37% for `test`).

The remainder of this study uses the manually corrected corpora as a baseline. Although this is not satisfying in terms of comparisons with other studies, we found ourselves constrained to do so because our automatic conversions from one annotation scheme to another required a clean and consistent annotation to begin with. In order to simplify comparisons, we have generated a difference file to apply to the original SPMRL corpus, available upon request.

5.3 Comparing annotation schemes

As seen in section 4.3, over 14% of the initial errors were artefacts of the annotation scheme for a CS with more than 2 conjuncts, where Talismane systematically attached the conjunct to the previous conjunct, whereas the original annotation scheme systematically attaches it to the first conjunct. Indeed, the previous conjunct attachment is more natural for transition-based parsers: since the comma is a highly ambiguous indicator for coordination, the coordination is often missed between the first and second conjuncts, and the first conjunct is reduced. By the time the parser reaches the coordinating conjunction, only the second conjunct is left on the stack. This suggested that changing the CS annotation scheme could lead to considerable improvements.

We therefore decided to experiment with four different equivalent CS annotation schemes, as shown in figure 2. Subfigure 2a gives the original **1H** (1st conjunct headed) annotation used in the SPMRL 2013 dependency corpus for French. The first conjunct always heads the CS, and governs the coordinating commas and conjunction with a `coord` label, which in turn govern the remaining conjuncts with a `dep_coord` label. Subfigure 2b shows the **CH** (conjunction headed) annotation, used by a wide variety of grammars: the conjunction governs all of the conjuncts with a `coord` label. Subfigure 2c shows the **PH** (previous conjunct headed) annotation, in which each conjunct governs the following coordinator (whether a comma or a conjunction) with the `coord` label, and the coordinator governs the following conjunct with the `dep_coord` label. Finally, subfigure 2d shows the **PH2** annotation, in which we skip the comma, so that conjuncts separated by a comma are directly governed by the previous conjunct using the `coord` label. In the case of a simple CS with 2 conjuncts, the PH and PH2 annotations are identical to the 1H annotation.

Notice that there is no loss of information between these four annotations, so that round-trip conversions can restore the original annotation. Post-positioned shared modifiers (e.g. *"Jean, Paul et Marie Dupont"*, where all three are members of the Dupont family) can be indicated by having the conjunction govern the shared modifier in the CH annotation, and having the 1st conjunct govern it in the other annotations. This annotation becomes non-projective (i.e. involves crossed dependency arcs) in 1H, PH and PH2 when the modifier applies to the objects of a prepositional phrase coordination, e.g. *"Je parle* *de Jean, de Paul, et de Marie Dupont"* ("I'm talking about John, Paul and Marie Dupont"). Since we use a projective parser in the present study, we change the governor to the final conjunct when required to avoid non-projectivity, thus losing some information. The 1H, PH and PH2 have no simple way of distinguishing ante-positioned shared modifiers from modifiers of the first conjunct, e.g. *"Chers Jean, Paul et Marie"* ("Dear John, Paul and Mary"). Moreover, none of these annotation schemes provide a clear solution for elliptical coordinations, e.g. *"J'ai vu Jean et Paul hier, et Marie aujourd'hui"* ("I saw John and Paul yesterday, and Mary today").

Another possibility for annotation was suggested by detailed analysis of the actual transition sequences for the first 20 coordination errors, revealing two cases in which, if a comma followed the first conjunct, the first conjunct was erroneously reduced. This suggested that having to take a decision when the comma was found at β_0 led to errors which could be eliminated if the comma were immediately attached and only used as a feature for further decisions. Now, if we look at the French Treebank annotation for punctuation outside of coordinated structures, the label is always `ponct`, but the choice of the punctuation's governor seems fairly arbitrary. Parser confidence is thus very low for punctuation attachment decisions, and as a result, when applying a beam search, the beam is often filled with alternative arbitrary punctuation attachment decisions instead of true syntactic ambiguities. We therefore decided to experiment as well with attaching punctuation systematically to the previous non-punctation token (or to the root artefact when punctuation opens the sentence), except in the case of coordinating commas for the 1H and PH annotations. Indeed, for the CH and PH2 schemes, we were forced to apply this punctuation "fix" in order to avoid generating a large number of non-projective punctuation arcs when transforming the corpus. In these latter two annotations, where coordinated commas are not used to annotate the CS, applying a punctuation fix results in systematic annotation for all punctuation in the corpus, thus resulting in a systematic application of the `right-arc`$_{\text{ponct}}$ and `reduce` transitions.

We thus make the hypothesis that transition-based parsers will favour those annotations which rely on shorter-distance dependencies, specifically PH and PH2. Our second hypothesis is that systematic annotation for commas (PH2) helps improve annotation by removing a needless source of ambiguity.

Scheme:	**1H**	**1H+P**	**CH+P**	**PH**	**PH+P**	**PH2+P**
Dev						
Coord f-score	85.75	85.60	73.20	86.68	86.96	89.21
Coord prec.	99.55	99.55	98.88	99.49	99.49	99.41
Coord recall	75.31	75.09	58.11	76.79	77.24	80.91
LAS no punct.	89.69	89.69	87.44	89.74	89.82	90.11
UAS no punct.	91.71	91.64	89.39	91.74	91.78	92.02
LAS	87.34	91.00	89.13	87.38	91.11	91.45
UAS	89.10	92.69	90.81	89.12	92.82	93.10
Test						
Coord f-score	86.75	86.94	73.09	88.20	88.44	90.29
Coord prec.	99.70	99.52	99.38	99.75	99.50	99.71
Coord recall	76.78	77.18	57.80	79.04	79.59	82.50
LAS no punct.	89.63	89.81	87.19	89.76	89.94	90.16
UAS no punct.	91.63	91.79	89.17	91.75	91.94	92.13
LAS	87.19	91.12	88.93	87.29	91.24	91.49
UAS	88.93	92.85	90.64	89.01	92.98	93.20

Table 4: Comparing CS annotation

Table 4 shows results for the six annotation schemes (where +P indicates the punctuation fix was applied): 1H, 1H+P, CH+P, PH, PH+P, PH2+P. All results are after targeted manual correction. For ease of comparison with previous studies, we show LAS and UAS both with and without punctuation. Unsurprisingly, in the schemes without the punctuation fix, hence with arbitrary attachment for punctuation, we systematically lose 2% when we include punctuation in the LAS/UAS, whereas in the schemes with the punctuation fix we systematically gain over 1%.

In the coordination results, we include both the `coord` and `dep_coord` labels, since different schemes have different proportions for these. Precision is very high because of the strong markers for coordination. Recall is much lower, because of the difficulty of finding the first conjunct. The conjunction-headed scheme CH+P is a clear loser in transition-based parsing—hardly a surprising result, since it requires far more lookahead features. All of the previous-conjunct headed schemes (PH, PH+P, PH2+P) outperform the first-conjunct headed schemes (1H, 1H+P) by over 1.5% when it comes to the coordination f-score, which validates our hypothesis based on the analysis of errors in section 4.3. Finally, the clear winner is the PH2+P scheme, where all attachment ambiguity is transposed from punctuation to the conjuncts, with 2.0% gain in coordination f-score with respect to the PH+P scheme. The coordination f-score error reduction between the original 1H scheme and PH2+P is 24.28% for `dev` (26.72% for `test`). In terms of statistical significance for both the `dev` and `test` corpora (McNemar's test applied to identifying individual conjuncts), the differences between 1H, 1H+P, PH and PH+P are not significant (p-value > 0.05). The differences between any other schema and CH+P or PH2+P are highly significant (p-value < 0.001).

5.4 Combining with targeted features

In our final experiment, we combine the PH2+P annotation scheme with the targeted features presented in section 5.1, to see to what extent the gains are cumulative. We also test at different beam widths to see how much additional gain can be had at higher beams.

Beam:	**Beam 1**				**Beam 2**				**Beam 5**			
Scheme:	**1H**		**PH2+P**		**1H**		**PH2+P**		**1H**		**PH2+P**	
Features:	∅	+	∅	+	∅	+	-∅	+	∅	+	∅	+
Dev												
Coord f-score	85.8	86.4	89.2	90.0	87.0	87.2	90.3	90.5	87.2	87.4	90.8	90.7
Coord prec.	99.6	99.4	99.4	99.4	99.6	99.5	99.5	99.5	99.4	99.4	99.6	99.5
Coord recall	75.3	76.3	80.9	82.2	77.2	77.5	82.7	83.0	77.6	78.0	83.3	83.4
LAS no pnct	89.7	89.7	90.1	90.3	90.2	90.3	90.5	90.6	90.4	90.4	90.7	90.7
UAS no pnct	91.7	91.8	92.0	92.2	92.2	92.3	92.5	92.6	92.4	92.5	92.6	92.7
LAS	87.3	87.4	91.5	91.6	88.0	88.1	91.8	91.9	88.2	88.3	91.9	92.0
UAS	89.1	89.2	93.1	93.2	89.8	89.9	93.5	93.6	90.0	90.1	93.6	93.7
Test												
Coord f-score	86.8	88.5	90.3	91.3	87.8	89.3	90.5	91.6	88.6	89.6	90.6	91.7
Coord prec.	99.7	99.6	99.7	99.7	99.8	99.7	99.6	99.6	99.8	99.6	99.6	99.6
Coord recall	76.8	79.5	82.5	84.3	78.4	80.9	83.0	84.8	79.6	81.5	83.1	85.0
LAS no pnct	89.6	89.9	90.2	90.3	90.3	90.4	90.6	90.7	90.5	90.6	90.6	90.8
UAS no pnct	91.6	91.9	92.1	92.2	92.2	92.4	92.5	92.6	92.5	92.6	92.6	92.7
LAS	87.2	87.4	91.5	91.6	88.0	88.2	91.8	92.0	88.4	88.4	91.9	92.0
UAS	88.9	89.2	93.2	93.3	89.7	89.9	93.5	93.6	90.0	90.1	93.6	93.7

Table 5: Combining annotation schemes and targeted features at different beam widths

Table 5 shows the results at beams 1, 2 and 5, for the original scheme 1H and the best scheme PH2+P, and with (+) or without (∅) targeted features. Gains are clearly centered on coordination recall. Table 6 shows the same information in terms of f-score error reduction with respect to the baseline configuration (1H annotation, baseline features, beam 1), with a maximal reduction of 35.09% for the `dev` corpus, and 37.28% for `test`. The three parameters tested are to a large extend cumulative. Individually, changing the annotation standard gives the most gain, followed by targeted features and then increasing the beam size to 2. In terms of statistical significance for the test corpus (McNemar's test applied to identifying individual conjuncts), all combinations are significant (p-value < 0.05) except for: PH2+P/∅/1-2 to PH2+P/∅/5; PH2+P/+/2 to PH2+P/+/5; and a few other combinations going from 1H/+ to PH2+P/∅.

	None	**Features**	**Scheme**	**Both**
Dev: base f-score = 85.75				
Beam 1	0.00	4.28	24.28	29.89
Beam 2	8.49	9.82	32.14	33.40
Beam 5	9.82	11.44	35.09	34.95
Test: base f-score = 86.75				
Beam 1	0.00	12.91	26.72	34.64
Beam 2	8.15	19.02	28.53	36.83
Beam 5	13.58	21.81	28.98	37.28

Table 6: Coordination f-score error reduction with respect to 1H, baseline features, beam 1

In terms of time performance, these changes have a vastly different cost. All tests were run on an Intel Xeon E3-1245 V2 machine, with a 3.4GHz clock speed, 4 cores, 8 threads, and 8 Mb cache, running the Ubuntu 12.04.2 LTS 64-bit operating system. The baseline setup takes 171 seconds to parse the `test` corpus (+133 seconds to load the model and lexicon), giving about 400 tokens/second. Changing the schema from 1H to PH2+P speeds up analysis slightly (×0.93). Changing the beam width results in a linear increase in time, ×2 for a beam of 2, and ×5 for a beam of 5. Finally, targeted features result in a ×22 increase in time.

We also performed a detailed error analysis for `dev` corpus, on the remaining errors in the PH2+P corpus with targeted features at beam 1. Although the number of erroneous coordinations analysed has reduced from 241 to 151, the percentage of errors relating to simple parallelism (pos-tag mismatch, preposition mismatch, etc.) remains stable, down from 38% to 36%. Annotation errors are reduced from 24% to 11%. Artefacts of the annotation scheme in which conjuncts are attached to the first or second conjunct are reduced from 15% to 5%. Finally, the complicated cases have climbed significantly, with ellipses climbing from 5% to 13% and cases where only semantics can help us decide climbing from 12% to 23%. The latter results indicates that introducing semantic resources might be worthwhile for the remaining errors.

There are a few cases of CSs coordinating unlike categories, where the new features introduced errors. We have a two cases of true non-parallelism, as in the following case, where an adjectival past participle is coordinated with a prepositional phrase:

Example 5.2 [...] celle *d'/P* une part significative des programmes et des productions réalisées/VPP **ou** *en cours de/P* réalisation. *(... that* of *a significant part of programs and productions that are already finished* ***or*** *currently* being *prepared.)*

We have a similar valid case of a non-parallel copula coordinating an adjective with a pronoun :

Example 5.3 Ce n'*est/V* pas forcément la plus économiquement souhaitable/ADJ, **mais** celle/PRO qui fera le moins de vagues, compte tenu de l'agitation dans les campagnes, *entendait/V*-on [...] *(It*'s *not necessarily the most economically desirable,* ***but*** *the one which will make the least waves, given the restlessness in the countryside, we were* told... *)*

The remaining cases are related to spelling errors in the original text, or to tokenisation and pos-tag errors in the gold pos-tags. For example, in the following case, the journalist misspelt the second *baisser* (to lower) as an infinitive verb whereas it should have been the homophone past participle *baissé*:

Example 5.4 Quant au dollar lui-même, il a monté/V quand on croyait qu'il allait *baisser/VINF* [...] **et** *baisser/VINF* derechef quand le marché commençait à se convaincre... *(As for the dollar itself, it rose when we thought it would* lower, ***and*** lower[ed] *once again when the market started to convince itself...)*

A second case involves the MWE *conformément aux* (in conformance with), which should probably be marked as a single preposition rather than ADV+P:

Example 5.5 *Dans le cas des/P* céréales, **et** conformément/ADV *aux/P* orientations souhaitées par les organisations professionnelles [...] *(*In the case of *cereals,* ***and*** *in conformance* with *the desires of professional organisations, ...)*

Similar cases involve the pos-tagging of *généraux* as a noun (generals in an army) rather than an adjective (general):

Example 5.6 [. . .] à l'ensemble des *présidents/NC* des conseils régionaux/ADJ et *généraux/NC*. *(. . . to all of the* presidents *of* regional ***and*** general *councils.)*

6 Conclusions and perspectives

In the present study, we attempted to improve the parsing of coordinated structures in French through changes to the annotation scheme and the application of targeted features. Both methods were successful, with annotation scheme changes reducing the `test` corpus coordination f-score error rate by 26.72%, targeted features reducing it by 12.91%, and the two combined reducing it by 34.64% (36.83% at beam 2, 37.28% at beam 5).

However, the application of targeted features comes at a considerable practical cost in terms of time performance ($\times$22 increase in time). This is partly due to the fact that features are described in configuration files using a declarative syntax, so that certain operations (e.g. looking forward in the buffer) are repeated thousands of times. Indeed, forward-looking features do not rely on partial parsing information, and could even be cached for any given token for the entire sentence parse, across parse configurations. If features were programmed and compiled, this could be made far more efficient, but we would lose the advantage of external configuration files.

In addition, we introduced a method for efficiently correcting training corpus errors through the projection of targeted features, a method which could be extremely useful for corpus constructors. Finally, we highlighted the usefulness of removing all ambiguity from the annotation of punctuation.

In a future study, we would need to test these methods with guessed pos-tags rather than gold pos-tags in order to check their sensitivity to pos-tag errors. It would also be interesting to apply our methods to other languages, and to include targeted semantic features based on semantic resources automatically constructed using semi-supervised methods. For languages with a richer morphology than French, it might well be worthwhile to introduce features based on morphological parallelism as well. Finally, various methods would have to be explored for improving the time performance of targeted features, if possible without losing the configurability and flexibility of declarative feature files.

Acknowledgements

I would like to thank the anonymous reviewers for their in-depth reading and many helpful suggestions.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks*. Kluwer.

Marie Candito, Joakim Nivre, Pascal Denis, and Enrique Henestroza Anguiano. 2010. Benchmarking of statistical dependency parsers for french. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 108–116. Association for Computational Linguistics.

Deirdre Hogan. 2007. Coordinate noun phrase disambiguation in a generative parsing model. In *Annual Meeting - Association for Computational Linguistics*, volume 45, page 680.

Angelina Ivanova, Stephan Oepen, and Lilja Øvrelid. 2013. Survey on parsing three dependency representations for english. *ACL 2013*, page 31.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency parsing*. Morgan & Claypool Publishers.

Sandra Kübler, Wolfgang Maier, Erhard Hinrichs, and Eva Klett. 2009. Parsing coordinations. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 406–414. Association for Computational Linguistics.

Wolfgang Maier and Sandra Kübler. 2013. Are all commas equal? detecting coordination in the penn treebank. In *The Twelfth Workshop on Treebanks and Linguistic Theories (TLT12)*, page 121.

Wolfgang Maier, Erhard Hinrichs, Sandra Kübler, and Julia Krivanek. 2012. Annotating coordination in the penn treebank. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 166–174. Association for Computational Linguistics.

Ryan T McDonald and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *EMNLP-CoNLL*, pages 122–131.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

Martin Popel, David Marecek, Jan Štepánek, Daniel Zeman, and Zdeněk Žabokrtskỳ. 2013. Coordination structures in dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Éric Villemonte de la Clergerie. 2014. Jouer avec des analyseurs syntaxiques. In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2014)*, pages 67–78, Marseille, France.

Roy Schwartz, Omri Abend, and Ari Rappoport. 2012. Learnability-based syntactic annotation design. In *COLING*, pages 2405–2422.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiorkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clérgerie. 2013. Overview of the spmrl 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages: Shared Task*, Seattle, WA.

Masashi Shimbo and Kazuo Hara. 2007. A discriminative learning model for coordinate conjunctions. In *EMNLP-CoNLL*, pages 610–619.

Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2011. Evaluating dependency parsing: robust and heuristics-free cross-nnotation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 385–396. Association for Computational Linguistics.

Assaf Urieli and Ludovic Tanguy. 2013. L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 188–201, Les Sables d'Olonne, France.

Assaf Urieli. 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail.

Assaf Urieli. 2014. Améliorer l'étiquetage de "que" par les descripteurs ciblés et les règles. In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2014)*, pages 56–66, Marseille, France.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *ACL (Short Papers)*, pages 188–193.

Yue Zhang and Joakim Nivre. 2012. Analyzing the effect of global learning and beam-search on transition-based dependency parsing. In *COLING (Posters)*, pages 1391–1400.