How to Evaluate and Compare Tagsets? A Proposal

Hervé Déjean

Seminar für Sprachwissenschaft
Universität Tübingen
dejean@sfs.nphil.uni-tuebingen.de

Abstract

We propose a methodology which allows an evaluation of distributional qualities of a tagset and a comparison between tagsets. Evaluation of tagset is crucial since the task of tagging is often considered as one of the first tasks in language processing. The aim of tagging is to summarise as well as possible linguistic information for further processing such as syntactic parsing. The idea is to consider these further steps in order to evaluate a given tagset, and thus to measure the pertinence of the information provided by the tagset for these steps. For this purpose, a Machine Learning system, ALLiS, is used, whose goal is to learn phrase structures from bracketed corpora and to generate formal grammar which describes these structures. ALLiS learning is based on the detection of structural regularities. By this means, it can be pointed out some non-distributional behaviours of the tagset, and thus some of its weaknesses or its inadequacies.

1. Introduction

> A typical and familiar case of corpus annotation is grammatical tagging (also called word-class tagging, part-of-speech tagging or POS-tagging). In this case a label or tag is associated with a word [...] to indicate its grammatical class. (Garside et al., 1997)

As G. Leech has written, tagging is one of the major tasks in corpus annotation. More and more this annotation is done automatically or semi-automatically by means of taggers. The emergence of these annotated corpora allows the development of learning techniques using these corpora in order to learn taggers or parsers (Brill, 1993; Charniak et al., 1993; Brant, 1999). One of the questions which is asked when you want to annotate a corpus or to train a tagger is: which tagset do you use and why? For this purpose several tagsets have to be compared. Whereas methodologies for evaluating tools (taggers or parsers) exist (Paroubek, 1998; LREC, 1998; Caroll et al., 1999), the problem of evaluation of tagsets seems to be less investigated.

The existing methods proposed until now try to modify an existing tagset in order to improve the tagging accuracy (Section 7.). (EAGLES, 1996a; Baker et al., 1998) propose guidelines which offer an abstract model of features sets, but these recommendations rather concern the standardisation of tagset and provide no criterion for evaluating tagsets[1].

In this article we propose a methodology which distinguishes the problem of tagset evaluation from tagging evaluation, and tries to evaluate the pertinence of tagsets relatively to parsing. The general idea relies on this observation: *Part-of-Speech tagging is often seen as the first stage of a more comprehensive syntactic annotation, which assigns a phrase marker, or labelled bracketing to each sentence of a corpus, in the manner of a phrase structure grammar* (Garside et al., 1997). If we can substitute tags for words, and nevertheless correctly parse sentences (in our case only phrase structures are parsed), we can consider that no essential information has been lost during tagging. In the contrary case (some structures can not be correctly parsed), some tags do not contain enough information (contained at the word level) to achieve a partial parsing of the sentence. We will show that the number of tags is not an appropriate criterion for evaluating a tagset, and that the quality of a given tagset rather depends on its distributional properties.

How to do practically this evaluation? We use a Machine Learning system, ALLiS, which generates a grammar of a given structure from a bracketed corpus. ALLiS main task is to determine whether or when a given tag belongs to a given phrase structure (PS), and we can indirectly use it to evaluate tagsets. ALLiS first learns the canonical behaviour of a tag, and then identifies its deviant behaviours. They correspond to non distributional behaviours.

The plan of this article is as follows: Section 2. calls to mind some goals and properties of tagset. Section 3. explains the methodology proposed here. The next section provides an illustration of an evaluation of tagsets. In Section 5. three tagsets are compared. Section 6. puts the question of the influence of the structures definition used for evaluating tagset. Finally, Section 7. discusses some related works.

The tagsets used in this article are the Penn Treebank tagset (hereafter Penn) (Marcus et al., 1993), the CLAWS2 tagset (Garside et al., 1987) and the SUSANNE tagset (Sampson, 1995).

2. About Tagsets

2.1. What kind of Tagsets and what Kind of Evaluation

Tagsets can be various and specific to one or another task (among morphological annotation, syntactic annotation, Information Extraction). In this paper, we focus the study on tagsets used for syntactic annotation. We propose a methodology for judging the linguistic (or "external") quality of a tagset (*the extent to which it allows retrieval of all important grammatical distinction in the language* (Sampson, 1995, page 29)), and not the "internal" quality (*the extent to which a particular tag is useful in aiding the disambiguation process, and increasing the accuracy of tagging*). (Elworthy, 1994), comparing both cri-

[1] Even if standardisation helps indirectly tagset evaluation.

teria, concludes that the external (linguistic) criteria should be followed. The "structural" part is not taking into account (use or not of features, organisation in hierarchical tree).

2.2. Notion of Domain

We can consider tagging as "the annotation of the words in a text with tags indicating their syntactic properties" (Halteren, 1999). The notion of *distribution* is often used when building up a tagset. If this notion is often employed or cited, the notion of *domain* (Harris, 1954, page 159) is less known or not enough explicitly used. Harris wrote: "All the statements about dependence and substitutability apply within some specified domain", and Harris cited word, phrase and clause as common types of domain. We consider that all the tags generated by a distributional method are in relation with a specific domain, and that they have to report their membership of these domains. One tag has to belong to one and only one domain. Although, the use of general tags simplifies the work of parsers, this tagging is not *optimal* since the parser has to come down to the lexical level to determine the syntactic behaviour of a word, namely its domain. The Penn tag *VBG* belongs to several domains and rather reflects a morphological property. Only using this tags, the syntactic behavior (which domain it belongs to) of a word tagged *VBG* is not possible.

3. A Proposal for Evaluating Tagsets

3.1. Using parsing for evaluating tagsets

The idea of using parsing for evaluating tagset is quite obvious, since the purpose of tagging is to help syntactic parsing and since these two levels are strongly correlated:

> In fact there is a strong argument that these [tagging and parsing] are not really distinct levels at all: grammatical tagging is merely a specification of the leaves [...] of the phrase structure tree which is a favoured model for syntactic annotation. (Garside et al., 1997)

In (Giguet and Vergne, 1997) or (Karlsson et al., 1995) tagging and parsing are done simultaneously, providing very good result at the tagging level, even though a great many work uses a sequential approach (tagging then parsing). The quality of tagsets is, in the last case, very important since parsing uses information provided by tagging. If a tag has no clear syntactic properties, its use does not allow an economy in the description and in the processing. In this case, the parser has to retagged the word in order to solve the problem of its membership to a given structure, and for this reason the use of this tag is non-optimum.

For evaluating tagsets, we consider the quality of information transmitted by tagging for parsing. One way to judge this quality is to answer this question: *Is it possible to parse a sentence using only tags?* If the answer is positive, the tagset transmits the necessary syntactic information containing in the words, and has thus performs its purpose. If the answer is negative, the tagset contains some classes which are not well enough defined. Table 1 illustrates this problem: a same sequence of tags (here **DT NN JJ NN**) can have more than one analyse, and the information contained in the tags themselves is not enough to find the right analyse.

[the	world	automotive	market]
DT	NN	JJ	NN
[the	shuttle]	[last	year]

Table 1: Two possible parsings of the sequence DT NN JJ NN (Wall Street Journal corpus).

The domains we proposed for this evaluation are the non-recursive phrase structures (PS) and simple clauses. Evaluation using a full parsing does not seem realistic presently since parsing prepositional phrases requires generally lexical resources (Collins and Brooks, 1995). The use of PS and clauses allows a fully coverage of many tagsets.

3.2. The Theoretically Minimal Tagset

We would like to point out that the quality of a tagset does not depend on the quantity of tags. For this purpose, we build up the minimum tagset necessary to parse sentence whatever the domains are. The first idea is to use a tagset with one tag per structure. Suppose we want to parse only NP and VP. We simply need three tags: NP and VP and O (other, for elements which do not belong to NP or VP). Words belonging to the NP structure are tagged NP and similarly for VP. But this tagset is not enough. The following example illustrates the problem: a sequence of adjacent NP tags can correspond to a sequence of structures.

[NP the	way	the	economy NP]	[VP will	perform VP]
NP	NP	NP	NP	VP	VP
the	way	the	economy	will	perform

In order to solve this problem, we have to add one feature to each tag: the break property. A tag takes the feature $B+$ if it introduces a break into a sequence of elements belonging to the same kind of structure (in general these elements correspond to specific classes such as determiner in English). The word *the* must be tagged NP_{B+}. The new parsing of the preceding sentence is :

[the	way]	[the	economy]	will	perform ...
NP_{B+}	NP_{B-}	NP_{B+}	NP_{B-}	VP_{B+}	VP_{B-}
the	way	the	economy	will	perform

Thus a tagset composed of one tag pro Phrase Structure with the feature B+ or B- is enough in order to segment sentences into those structures (but this tagset does not allow the inner parsing of a structure). We estimate that a tagset of about 20 tags is enough to parse a sentence into PS and clause structures.

3.3. The Learning System: ALLiS

The apparition of annotated and bracketed corpora has developed the utilisation of Machine Learning techniques in Natural Language Processing (Wermter et al., 1996), (Nerbonne and Osborne, 1999). Generally annotated corpora are used as "training data" and then the learning system is evaluated with test data. In this work, the opposite

is done: we use a Machine Learning technique in order to evaluate linguistic data.

ALLiS[2] (Architecture for Learning Linguistic Structures) (Déjean, 2000) is a learning system which uses the theory refinement in order to learn non-recursive PS. Using bracketed corpora as input ALLiS learns a regular expression grammar which describes PS. This grammar is then used by the CASS parser (Abney, 1996) or by the Xerox Finite State Tools (Karttunen et al., 1997). The learning task is composed of two steps. The first step is the generation of an *initial grammar*. In this grammar, each tag is assigned the value: "belong to the structure" or "does not belong to the structure". This initial grammar provides an incomplete analysis of the data. The second step is the refinement of this grammar. During this step, the validity of the rules of the initial grammar is checked and the rules are improved (refined) if necessary. This refinement relies on the use of two operations: the *contextualisation* (in which contexts such a tag belongs or not to the PS) and *lexicalisation* (such a word belongs to the phrase). ALLiS generates a list of problematic points of the tagset encountered during the learning phase. This identification of problemtic points can only be provided by symbolic learning systems, since statistical methods can just provide directly a segmentation into structures (Tjong Kim Sang and Veenstra, 1999).

3.3.1. Notion of refinement

The notion of "refinement" is the central notion for ALLiS. When a rule is learned (for example: the tag *VBN* does not belong to NP), ALLiS tries to find out exceptions to this rules. In order to refine this rule, it disposes of two operators: the *contextualisation* and the *lexicalisation* (Table 2). The contextualisation provides a list of contexts where an element categorised in one category can appear in another category. The tag *VBN* is categorised by default as occurring out of an NP[3], but ALLiS can detect contexts in which it always occurs inside an NP. The lexicalisation points out words whose behaviour is constant. In our training corpus the word *increased* occurring before a noun belongs to an NP 10 times and only one time outside whatever the preceding context is. Both operations can be redundant: *the/DT increased/VBN labor/NN costs/NNS*.

Since ALLiS can provide *contextual* rules in order to improve the parsing, these are not beyond the actual tagging technology which mainly relies on the notion of local contexts. These situations can be detected at the tagging level.

3.4. Notion of Recoverability

This methodology considers that tagsets must contain the most possible syntactical information, and all the tags which do not follow this principle are thus negatively judged.

An opposite point of view is develop in the Penn Treebank (Marcus et al., 1993) which uses the notion of *recoverability*:

> A key strategy in reducing the tagset was to eliminate redundancy by taking into account both lexical and syntactic information. (Marcus et al., 1993, 314)

The idea is to reduce the size of the tagset used as starting point (the Brown tagset), so that the redundancy contained in the corpus is also reduced. Then, even if the Penn tagset does not contain enough information in order to determine the syntactic role of each element, tags function can be found out using lexical information (the word) or syntactical one (bracketing):

> We would like to emphasize that the lexical and syntactic recoverability inherent in the POS tagset version of the Penn Treebank corpus allows end users to emply a much richer tagset than the small one described in Section 2.2 if the need arises (Marcus et al., 1993, 315).

It seems that this need systematically arises for processing applying a cascaded approach, since this bottom up approach can not use higher level in order to identifiy the syntactic role of tags. But this identification is easy by using the Penn Treebank since the syntactic information is present. We understand the philosophy of the recoverability as being: if you know the word and its structure, you can find its tag. But from the "automatic processing" point of view, the situation is generally opposite: if the word and its tag are known, can we determine its syntactic structure?

We can also notice that this notion of recoverability seems to be in contradiction with the following remark found in the same article: *By contrast [with the Brown corpus], since one of the main role of the tagged version of the Penn Treebank corpus is to serve on the basis for a bracketed version of the corpus, we encode a word's syntactic function in its POS tag whenever possible.* (Marcus et al., 1993, page 316)

This remark follows the methodology used in this article in order to evaluate tagsets.

```
VBN non rel
contextualisation:
  rel left in  VBN  POS   13    1
  rel left in  VBN  JJ     7    3
  rel left in  VBN  PRP$  18    0
  rel left in  VBN  CD     4    1
  rel left in  VBN  DT   221    3
  rel left out VBN  TO     3    0
  rel left out VBN  IN    74    1
  rel left out VBN  VBG   13    2
lexicalisation:
  VBN increased     10   1
  VBN discontinued   7   0
  VBN increased      5   0
  VBN Posted         4   0
```

Table 2: Refinement of the tag VBN for the NP structure. First line: the tag VBN after the tag POS occurs 13 times in an NP and one time outside.

[2] A demo of the chunker for NP and VP can be used at: `http://www.sfb441.uni-tuebingen.de/~dejean/chunker.html`.

[3] The occurrences of *VBN* in an NP represent 15% of the total occurrences

4. How to Evaluate of a Tagset

We present two kinds of evaluation: a global and a local one. If it could seem to be interesting to get a quantification of the quality of a tagset, we will see that a qualitative approach is preferable, and at any rate is mandatory to identify intrinsic weaknesses of a tagset. All the tagsets studied provide roughly similar result and they only differ about some specific points whose frequency is quite low.

But, first of all, we have to choose a domain (Section 2.2.). Three are studied: the based-NP, based-VP, and Based-PP. The definition of these structures is provided by the Penn Treebank (Section 6. discusses the importance of this choice regarding the evaluation). Sections 15-18 of the Wall Street Journal corpus serve as training data for ALLiS, and Section 20 serves as test data. The data are tagged using the Penn tagset and CLAWS2 tagset (ACQUILEX tagger (Garside et al., 1987)).

4.1. Global Evaluation

Global evaluation consists of evaluating the initial grammar generated by ALLiS. A "perfect" tagset, according to our criteria, would provide a score of 100% (or nearly), and would not be improved by contextualisation. The initial grammar generated by ALLiS for NP with the Penn tagset is (CASS formalism):

```
:np
NB = PRP| EX | WP | WDT;
AB = DT | POS | PRP$ | WP$;
A= JJS | JJ | JJR | $ | #;
aA = RBR|RBS;
N= CD | NN | NNP | NNS | NNPS;
NP -> NB |
      AB* ((aA* A)* N)+;
```

This grammar is only composed of tags which are categorised by ALLiS as belonging by default to NP (Déjean, 2000).

Table 3 shows a global evaluation of the Penn and CLAWS2 tagsets for NP, PP and VP

Structure	Tagset	Initial Grammar
NP	Penn	86.33
	CLAWS2	86.30
PP	Penn	84.30
	CLAWS2	90.60
VP	Penn	83.15
	CLAWS2	81.87

Table 3: Evaluation of two tagsets: the Penn Treebank and the CLAWS2 tagset. the rate is $F = \frac{2*precision*recall}{recall+precision}$.

From this table, it is, in fact, difficult to determine the best tagset. The result depends on the structure used. Concerning NP, the evaluation of the initial grammar provides two closed scores. This global evaluation offers no information at all about weaknesses and strengths of the different tagsets. For instance, the initial NP grammar using the Penn tagset gets the best score thanks to few specific points: the use of the tag *WDT* for the word *that*, used as *wh-determiner*, when the CLAWS used the *CST*, merging the classes of conjunction and of relative pronouns. Despite this weakness, which concerns a frequent structure and penalises the CLAWS evaluation of 0.5%, the CLAWS tagset counterbalances this using other tags which are more distributional than the Penn tags (Section 5.). The bad score of the Penn tagset concerning PP is due to the merging of prepositions and complementisers into one class (*IN*). Regarding VP, the Penn tagset gets a better score. This is only due to a higher error rate with the CLAWS2 tagsets, errors which hide the better distributional properties of some CLAWS2 tags.

This kind of evaluation is biased by the fact that, most of the time, the improvements due to distributional tags only concern some specific points whose frequency is small. For example, tagging the word *including* as preposition (word tagged *VBG* in the Penn Treebank) improves the score of the VP parsing by only 0.45%. The rate of the errors of tagging is largely higher (5-7%) and hides such positive points of a tagset. It is also more interesting to restrict the evaluation to specific constructions (Section 4.2.) in order to make the differences between tagsets emerged.

4.2. Local Evaluation

Another way to evaluate a tagset is to consider the behaviour of each tag regarding a given domain, using the notion of *reliability*. The reliability of an element corresponds to the ratio between its frequency in the structure over its total frequency in the corpus. If its reliability is 1 (resp. 0), the tag always belongs (resp. does not belong) to the structure and its syntactical behaviour is predictable. An ideal tagset would provide tags which only belong to one unique structure. The tagset used by (Giguet and Vergne, 1997) defines tags from three structures: non-recursive NP, non-recursive VP and simple clause, and a tag only belongs to one domain (except coordinations). Parsing with this tagset is thus straightforward. Unfortunately, tagsets often provide elements which can belong to several structures. We can be tempted to use this information to furnish a direct quantifiable criterion of comparison between tagsets, but some tries do not lead to conclusive results. One possible evaluation might be to estimate the number of reliable tags for a given tagset, but this feature offers no easy way to compare tagsets: a reliable tag in one tagset can correspond to several reliable tags in another tagset, and thus the last tagset is privileged (an operation of mapping would be necessary in order to compare tagset by this way). However, this criterion can be used in order to evaluate the intrinsic quality of a tagset (a very good tagset being only composed of reliable tags). Table 4 provides some examples of the reliability of some tags for the Penn, CLAWS2 and SUSANNE tagsets.

In some cases the subcategorisation of some general tags can be useful (RB), and sometimes useless (NN). We can make two general remarks: First it is always useful to subdivide non-reliable tags, so that the subdivision introduces reliable tags (the subdivision of the non-reliable tag *RA* (CLAWS2) leads to the creation of two reliable tags: *RAa* and *RAh* (SUSANNE)). Second, the subdivision of a reliable tag might not be useful for a syntactical purpose,

Tagset	Tag	NP Rel.	Tag	NP Rel.
Penn				
	RB	13%	NN	98%
CLAWS2				
	RL	7%	NN1	98%
	RR	18%	NNT1	99%
	RA	22%	NNU	98%
	RT	49%	NNL1	98%
	RG	70%	NNJ1	99%
SU				
	RAa	0%	NNT1c	100%
	RAh	100%	NNT1h	100%
	RTt	100%	NNT1m	100%
	RGi	100%		

Table 4: Syntactic reliability of some adverbial and nominal tags in the NP structure (Penn, CLAWS2 and SUSANNE tagsets)

but can be required by other purposes (Information Extraction for example).

The following table shows that these new tags correspond to classes including few words having a very specific behaviour.

RAa	*ago*
RAh	*am, pm, o'clock*
RTt	*today, tomorrow, yesterday, tonight*
RGi	*around, about, circa, getting_on_for, over some, under, up_to*

This example perfectly illustrates the remark by Creissels: *Among adverbs, it seems that we find a certain number of units whose distribution is so specific that is is not obvious that their categorisation to one or the other big category allows achieving an economy in the description [...].* (Creissels, 1995)

We can also find another positive point when using more precise tags, point which concerns manual annotation. In Sections 15-16 of the Wall Street Journal corpus, the word *about* is tagged *RB* when it occurs before a numeral or a measure. But in Sections 17-18, this word in the same context is systematically tagged with another tag: *IN*. We think that the use of a specific tag (as *RGi* used in the SUSANNE corpus) for this word in this particular context would have avoided this error of annotation.

This list of the non-reliable tags constitutes a precious resources in order to improve tagset. It points out weaknesses of the tagset, namely tags which possess no clear syntactic behaviour, and which require subcategorisation (introduction of new reliable tags). This list is used in order to compare tagsets as described in the next section.

5. Comparison between the Penn, CLAWS2 and SUSANNE Tagsets

We compare here three tagsets: the Penn tagset, the CLAWS2 tagset and the SUSANNE tagset. We do not dispose of a tagger using the SUSANNE tagset, and we thus have to use two different corpora to achieve this comparison. The Penn tagset is a simplification of the Brown tagset (Francis, 1980), when the CLAWS2 is an extention of it:

> On the other hand, when the UCREL group began to move on from the problem of automatic word-tagging to the larger problem of parsing, they found it necessary to introduce a few new wordtags for words having a special role with respect to higher levels of grammatical structure. (Garside et al., 1987, page 166)

The SUSANNE tagset is also a refinement of the CLAWS2 tagset.

We present here several specific points where ALLiS detects difficulties. These points were noted during the processing of the Penn tagset and we will see how the two other tagsets process them.

5.1. Adverbs

The study of the traditional "class" of adverbs is interesting since it is generally one of the most non-distributional in the grammar. The Penn tagset uses just one tag, *RB*, for its class. CLAWS2 and SUSANNE tagsets present a dozen of classes which correspond more or less to the tag *RB* (the mapping is not direct). Let look at how the three tagsets process the sequence *[NP about a year NP]*:

- about_ RB/IN a_DT year_NN (Penn)
- about_RG a_AT1 year_NNT1 (CLAWS2)
- about_RGi a_AT1 year_NNT1c(SU)

The Penn just specifies than *about* is tagged as adverb or as proposition/complementiser (depending on the sections of the corpus), and ALLiS can not learn that, in this context, the word *about* belongs to the NP (*IN* follows by *DT NN* traditionally corresponds to a PP). Only the presence of a cardinal (*CD*) allows the correct parsing of the word *about* (Table 5) The case is similar which the CLAWS2 tagset (RG: class of the general adverbs).

```
IN non rel
              IN CD     About   11   0
              IN CD     about  128   1
              IN        the     11   0
              IN        under    1   0
              IN TO CD  up       8   2
              IN $ CD   around   6   0
rel spec      IN JJS CD         25   0
rel left out  TO IN             11   3
```

Table 5: Processing of the tag IN (Penn tagset).The presence of the word *the* tagged *IN* is due to errors of tagging.

SUSANNE uses the tag *RGi* which exactly corresponds to the current structure:

> Words tagged RGi and other functioning similarly preceding a sequence of numeral and measure are analysed as forming a IC tagma [upper group] without internal grouping. (Sampson, 1995, page 226)

RGi is categorised by ALLiS as a reliable tag which always belongs to NP. The syntactic behaviour of this tag is fully

learnable without contextualisation or lexicalisation. We consider then that the SUSANNE tagset best covers this structural point. The CLAWS2 and Penn tagset have similar behaviour.

Another interesting structure is noun phrase functioning adverbially followed by the word *ago*. The sequence *a year ago* is encoded as an Adverbial Phrase (ADVP) by the Penn Treebank and by SUSANNE (RX:t) and is tagged:

- a_DT year_NN ago_RB (Penn)
- a_AT1 year_NNT1 ago_RA (CLAWS2)
- a_AT1 year_NNT1c ago_RAa (SUSANNE)

In order to build the temporal adverbial phrase from these sequences of tags, ALLiS has to use lexical information with the Penn tagset: the tag *RB* is not reliable even in the context *DT NN*, and the problem can be only solved thanks to the lexicalisation of RB (*ago*). With CLAWS2, the contextual information (RA after NNT1) is enough to built the adverbial structure. And finally, the SUSANNE tagset needs no contextual at all: the information is directly contained in the tag *RAa*, composed of just one word: *ago* and which occurs only in this kind of structure. We then establish the following classification for this structure: first SUSANNE, then CLAWS (need of contextualisation), and then Penn (need of lexicalisation).

5.2. A "Break in the Time"

The second main point we want to consider in order to compare these tagsets concerns nouns describing period of time functioning adverbially at the clause level.

In the Wall Street Journal corpus, words such as *yesterday* or *tomorrow* are tagged *NN* and compose NP. Since we use the NP definition of the Wall Street Journal corpus, we consider these words as member of NP. The problem of these words is illustrated Table 1: contrary to other words tagged *NN* and *NNP*, these words do not constitute a unique structure with the immediate preceding noun. This behaviour of *breaker* can only be detected through an operation of lexicalisation. The table 6 can be read as: If a tag *NN* is preceded by a noun, then both mainly belong to the same NP (line *NN non breaker 6861/123*). But there are 123 exceptions *of* this behaviour, and 99 (2+74+23) of these exceptions are covered by the three words *tomorrow*, *yesterday* and *today*. The case is similar with the tag *NNP* whose exceptions concern dates.

Concerning the CLAWS2 tagset, the wordclasses *NNT(1/2)* correspond to temporal nouns (day, week, month) and their break property is gotten directly. The problem concerns words such as *yesterday*. These words are tagged *RT* as well as words such as *then* and *now*. The class *RT* corresponds to quasi-nominal adverbs of time. ALLiS categorises the tag *RT* as non-member of NP by default (it mostly occurs out of NPs), and then uses contextualisation and lexicalisation to correctly parse words which are considered inside an NP. But contrary to the Penn tagset, the breaker property of the tag *RT* is automatically detected and no lexicalisation is needed (Table 7). This choice (adverb and not noun) penalises the CLAWS2 during the first step of the global evaluation (Table 3). But, for the second step

```
NN non breaker   5861/123
exceptions:
   NN  tomorrow    2  0 left
   NN  yesterday  74  2 left
   NN  today      23  0 left
[...]
NNP non breaker  5726/156
exceptions:
   NNP HUD         3  0 left
   NNP Wednesday  16  0 left
   NNP Dec.        2  0 left
   NNP Friday     25  4 left
   NNP Nov.        6  0 left
   NNP Monday     10  3 left
   NNP Sept.       2  0 left
   NNP Mr.         5  0 left NN
   NNP Oct.        8  1 left
   NNP Thursday    8  0 left
   NNP Tuesday    12  0 left
```

Table 6: Detection of the Break Property for the tags NN and NNP (Penn).

(contextualisation), the CLAWS2 offers better results (the detection of the break property is automatic whereas the Penn needs lexicalisation).

```
RT outer by default
contextualisation:
  rel left out RT VV0   9   3
  (EXP) now  RT VV0 3 3 2
  rel left out RT IF    4   1
  rel left out RT VVD  26   4
  (EXP) then RT VVD 3 3 2
  rel left out RT II   18   4
  (EXP) then RT II 4 4 2
  rel left out RT VVG   4   1
  rel left out RT VVN  10
  rel left out RT ICS  10   3
  rel right out RT IW   3
  rel right out RT .   44   9
  (EXP) now  RT . 8 9 2
  rel right out RT IF   3
  rel right out RT II  25   9
  (EXP) now  RT II 3 3 2
  (EXP) then RT II 4 4 2
  then RT ICS 8/9 -> 1 left tag
lexicalisation:
   RT tomorrow      8   0
   RT tonight       2   0
   RT Yesterday    15   0
   RT yesterday   150   0
   RT Today         8   0
   RT today        54   2
   RT then          8   1 ICS
RT breaker    9/93
```

Table 7: Processing of the tag RT tag (CLAWS2).

The SUSANNE tagset is perfectly distributionally regarding this problem, since it uses a wordclass *RTt* which only includes *yesterday*, *today*, *tomorrow* and *tonight*, and wordclasses for words indicating a time period (like

CLAWS2). The syntactic information is entirely contains in tags and the tagset is totally distributionally regarding this point.

This structure describing a period of time is distributionally well-marked[4], and the introduction of a specific phrase (and thus specific tags only belonging to this phrase) is strongly recommended (introduction present in SUSANNE). It is interesting to note that all the modifications done for the SUSANNE tagset are validated by ALLiS.

6. Interaction between Tagsets and Structures

We would like here to point out the importance of the definition of the domain used during the evaluation. Since the definition of a structure can vary from one corpus to another, it is interesting to see how well can a tagset adapt itself to a PS definition. We can think that the use of the NP definition provided by the Penn Treebank favours the Penn tagset to the detriment of the others. But, as we will see, using the Penn Treebank definition is not always an advantage for its tagset.

In the Wall Street Journal corpus, the words *object* and *subject* are tagged NN but do not belong to NP in specific contexts:

```
NN rel
   subject/NN      TO   14   0
   order/NN    IN  TO   11   1
   [...]
```

This behaviour is in contradiction with the general behaviour of the other words tagged NN, and the lexicalisation is needed in order to solve the problem. Regarding the CLAWS2 tagset, these words are respectively tagged *BTO* and *JJ* and are not considered as NP. SUSANNE uses the same kind of annotation but with ditto (BTO22 and II21). The definition used is thus an advantage of the two last tagsets and an inconvenient for the Penn tagset.

The opposite case can happen: a word tag as *NN* (noun), and then considered as belong to NP by the WSJ corpus, can be tagged differently by other tagsets. The following table gives the examples of the words *example* and *instance* which are tagged *NN* by the Penn tagset and occur in an NP, but which are tagged *REX* by the CLAWS2 tagset (adverb introducing appositional constructions).

```
REX outside by default
exceptions:
          REX instance   21   0
          REX example    38   0
left out  REX REX        59   2
```

This tag is considered as non-member of NP by default, but is identified as reliable when another tag REX occurs before it. Information provided by the lexicalisation is redundant (the context *REX REX* exactly corresponds to the sequences *for example* and *for instance*). These two words as tagged *REX22* in the SUSANNE corpus, and this tag is directly reliable (the word *for* is tagged *REX21*).

[4]Even at the phrase level

We see then that the two tagsets (CLAWS2 and SUSANNE) can be adapted to the NP definition provided by the Penn Treebank. As said in the preceding section, the SUSANNE tagset adapts itself very easily concerning the "yesterday" problem (use of a specific tag: *RTt*). The adaption, although difficult to quantify, seems to be a very good criterion in order to evaluate the distributionality of a tagset. Tagsets (such as SUSANNE) being adaptable are composed of tags whose syntactical properties are clearly defined into a given domain.

7. Other Approaches

The works already proposed about tagset evaluation concern rather the internal criteria (Section 2.). The purpose is thus to modify an existing tagset so that the tagging accuracy increases (EAGLES, 1996b). This improvement proceeds from the merge of some (generally) ambiguous wordclasses. The purpose of (Brants, 1995) is to reduce the size of the tagset in order to increase the frequency of some rare n-grams used by HMM, and thus to improve learning. (Elworthy, 1994) studies the relationship between tagging accuracy and tagset size with the conclusion that tagset size has weak influence on the tagging accuracy. (Schütze, 1995) proposes an algorithm for automatically merging distributional classes in order to improve the tagging accuracy. If some mergings have no or little effect on the PS segmentation (merging of NN(S), NNP(S)), then other are catastrophic (merging of RB, RP, RBR and RBS). These tags are certainly merged because of their lack of distributional properties.

(Hughes and Atwell, 1994) propose a way to evaluate automatically inferred wordclasses, but these classes can not by used directly for annotating corpora. (Wynne et al., 1998) combines the result of two different tagsets in order to improve tagging.

We can cite some tries in order to map tagsets (Hughes et al., 1995; Teufel, 1995). Similarly to ours, this technique might be indirectly used in order to compare tagsets (ease of one tagset to be map in one another, another manner to consider the degree of adaption of a tagset).

Few works propose manual comparison of tagsets. (Müller, 1997) compares the METS tagset and other English tagsets, more specially concerning *wh pronouns*. In (Sampson, 1995), a comparison is sometimes done between the SUSANNE tagset and the CLAWS2 tagset, in order to explain the added modifications.

8. Conclusion

We propose here a methodology in order to evaluate tagsets regarding syntactic parsing. This method uses a system, ALLiS, which systematically identifies deviant behavious of a tag. If the point of view used for evaluating tagset is not original (it was used during the elaboration of the SUSANNE tagset), it is useful to possess a software allowing a systematic evaluation of problematic tags. The method requires a corpus where domains are bracketed, resource which is mostly only available for English. If the quantitative evaluation offers a general estimation of the tagset, the detection of the weaknesses of the tagset and the comparison among several tagsets can only be done

through a detailled analysis for each tag. The same methodology can be use in order to evaluate tagsets used at the phrase and clause level.

9. Acknowledgements

This research is funded by the TMR network: Learning Computational Grammars[5]. I would like to thank Erik Tjong Kim Sang and Miles Osborne who provided data for this comparison.

10. References

Abney, Steven, 1996. Partial parsing via finite-state cascades. In *In Proceedings of the ESSLLI '96 Robust Parsing Workshop*.

Baker, J.P., Lou Burnard, A.M. McEnery, and A. Wilson, 1998. Corpus validation guidelines. Technical report, European Language Resources Association.

Brant, Thorsten, 1999. Cascaded markov models. In *EACL 99*.

Brants, Thorsten, 1995. Tagset reduction without information loss. In *proceedings of ACL-95, student session*.

Brill, Eric, 1993. *A Corpus-Based Approach to Language Learning*. Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.

Caroll, John, Guido Minnen, and Ted Brisco, 1999. Corpus annotation for parser evaluation. In *EACL workshop on Linguistically Interpreted Corpora (LINC)*.

Charniak, Eugene, Curtis Hendirckson, Neil Jacobson, and Mile Perkowitz, 1993. Equations for part-of-speech tagging. In *Proccedings of the Eleventh National Conference on Artifial Intelligence*.

Collins, Michael and James Brooks, 1995. Prepositional phrase attachment through a backed-off model. In *Third Workshop on Very Large Corpora*.

Creissels, Denis, 1995. *Éléments de Syntaxe Générale*. Presses Universitaires de France.

Déjean, Hervé, 2000. Theory refinement and natural language learning. In *COLING'2000*. Saarbrücken.

EAGLES, 1996a. Morphosyntactic annotation. Technical Report EAG-CSG/IR-T3.1, Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

EAGLES, 1996b. Study of the relation between tagsets and taggers. Technical Report EAG-CLWG-Tags/V, Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

Elworthy, David, 1994. Tagset design and inflected languages. In *EACL-94 SIGDAT*.

Francis, W. N., 1980. *Studies in English Linguistics*. Longman, pages 192–209.

Garside, Roger, Geoffrey Leech, and Anthony Mc Enery (eds.), 1997. *Corpus Annotation*. Longman, London and New York.

Garside, Roger, Geoffrey Leech, and Geoffrey Sampson (eds.), 1987. *The computational Analysis of English - A Corpus-based Approach*. London: Longman.

Giguet, Emmanuel and Jacques Vergne, 1997. From part-of-speech tagging to memory-based deep syntactic analysis. In *IWPT'97*. MIT, Boston,Massachussets, USA.

Halteren, Hans Van (ed.), 1999. *Syntatic Wordclass Tagging*. Book News, Inc., Portland.

Harris, Zellig, 1954. Distributional structure. *Word*, 10(2-3):146–162.

Hughes, John and E. Atwell, 1994. The automated evaluation of inferred word. In A. Cohn (ed.), *Proceedings of the 11 European Conference on Artificial Intelligence (ECAI-94)*.

Hughes, John, Clive Souter, and John Atwell, 1995. Automatic extraction of tagset mappings from parallel-annotated corpora. In *EACL'95*.

Karlsson, Fred, Atro Voutilainen, Huha. Heikkilä, and Artro Anttila (eds.), 1995. *Constraint Grammar, a language independent system for parsing unrestricted text*. Berlin and New York Mouton de Gruyter.

Karttunen, Lauri, Tamás Gal, and André Kempe, 1997. Xerox finite-state tool. Technical report, Xerox Research Centre Europe, Grenoble.

LREC (ed.), 1998. *First International Conference on Language Ressources and Evaluation*. Granada.

Marcus, Mitchell, Béatrice Santorini, and Marc Ann Marcinkiewicz, 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19(2):313–330.

Müller, Frank Henrik, 1997. *Tagsets für das Englische im Lichte der Wortartendiskussion*. Master's thesis, Westfälische Wilhelms-Universität Münster.

Nerbonne, John and Miles Osborne, 1999. Learning computational grammars. Technical report, TMR Project Nr. ERBFMRXCT980237, 1st Year Report.

Paroubek, Patrice, 1998. Experience in grace tagging evaluation. In *First International Conference on Language Ressources and Evaluation*. Granada.

Sampson, Geoffrey, 1995. *English for the Computer. The SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon Press.

Schütze, Hinrich, 1995. Distributional part-of-speech tagging. In *EACL*. Dublin.

Teufel, Simone, 1995. A support tool for tagset mapping. In *Workshop SIGDAT, EACL '95*.

Tjong Kim Sang, Erik and Jorn Veenstra, 1999. Representing text chunks. In *Proceedings of EACL'99, Association for Computational Linguistics*. Bergen.

Wermter, Stefan, Ellen Riloff, and Gabriele Scheler, 1996. Connectionist, statistical and symbolic approaches to learning natural languages processing.

Wynne, Martin, Roger Garside, Geoffrey Leech, and Andrew Wilson, 1998. Parallel wordclass tagging. In *First International Conference on Language Ressources and Evaluation.*

[5] `http://lcg-www.uia.ac.be/`