TAGGING FRENCH WITHOUT LEXICAL PROBABILITIES – COMBINING LINGUISTIC KNOWLEDGE AND STATISTICAL LEARNING

EVELYNE TZOUKERMANN
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974–0636
evelyne@research.att.com

DRAGOMIR R. RADEV*
Department of Computer Science
450 Computer Science Building, Columbia University
New York, NY 10027
radev@cs.columbia.edu

AND

WILLIAM A. GALE
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974–0636
gale@research.att.com

Abstract. This paper explores morpho-syntactic ambiguities for French to develop a strategy for part-of-speech disambiguation that a) reflects the complexity of French as an inflected language, b) optimizes the estimation of probabilities, c) allows the user flexibility in choosing a tagset. The problem in extracting lexical probabilities from a limited training corpus is that the statistical model may not necessarily represent the use of a particular word in a particular context. In a highly morphologically inflected language, this argument is particularly serious since a word can be tagged with a large number of parts of speech. Due to the lack of sufficient training data, we argue against estimating lexical probabilities to disambiguate parts

*The work was achieved while the author was at AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974–0636 of speech in unrestricted texts. Instead, we use the strength of contextual probabilities along with a feature we call "genotype", a set of tags associated with a word. Using this knowledge, we have built a part-of-speech tagger that combines linguistic and statistical approaches: contextual information is disambiguated by linguistic rules and n-gram probabilities on parts of speech only are estimated in order to disambiguate the remaining ambiguous tags.

1. Introduction

This paper explores morpho-syntactic ambiguities for French to develop a strategy to disambiguate part of speech labels that a) reflects the nature of French as an inflected language, b) optimizes the estimation of probabilities, c) allows the user flexibility in tagging.

Problems in tagging French:
French has a rich inflectional morphology and words can have up to eight different morphological analysis depending on the choice of tags. For example, let us take a common word of French, the word "moyenne" (meaning *average* as a noun, verb, or adjective) shown in Table 1. The word has seven distinct morphological analyses. Column 3 gives the full morphological analysis of the word, column 4 represents the tag corresponding to it from the large tagset and column 5, the tag from the small tagset (large and small tagsets are discussed in Sections 5.2 and 7.1.1). The seven tags in the large tagset get reduced to five in the small one.

word	**base form**	**morphological analysis**	**tagset1**	**tagset2**
"moyenne"	<moyen>	adjective, fem. sing.	JFS	jfs
"moyenne"	<moyenne>	noun, feminine sing.	NFS	nfs
"moyenne"	<moyenner>	verb, 1st pers., sing., pres., ind.	V1SPI	v1s
"moyenne"	<moyenner>	verb, 1st pers., sing., pres., subj.	V1SPS	v1s
"moyenne"	<moyenner>	verb, 2nd pers., sing., pres., imp.	V2SPM	v2s
"moyenne"	<moyenner>	verb, 3rd pers., sing., pres., ind.	V3SPI	v3s
"moyenne"	<moyenner>	verb, 3rd pers., sing., pres., subj.	V3SPS	v3s

TABLE 1. Morphological analyses of the word "moyenne".

In a given sentence where the word "moyenne" occurs, multiple other tags appear, as exemplified in Table 2. The second column of the table shows all the tags for the word in column 1. The correct tag is in bold, followed by the meaning of the correct tag in column 3.

Word	**tag from morphology**	**Meaning of the tag**
$<S>$	^	beginning of sentence
La	**rf** b nms u	article
teneur	**nfs** nms	noun feminine singular
moyenne	**jfs** nfs v1s v2s v3s	**adjective feminine singular**
en	**p** a b	preposition
uranium	**nms**	noun masculine singular
des	**p** r	preposition
rivières	**nfp**	noun feminine plural
,	**x**	punctuation
bien_que	**cs**	subordinating conjunction
délicate	**jfs**	adjective feminine singular
à	**p**	preposition
calculer	**v**	verb

TABLE 2. Sample output of a sentence chunk with the word "moyenne".

The goal of tagging is to find the most appropriate tag associated with a word. It has often been suggested that lexical probabilities should be used on word forms in order to find the most likely tag for a word. This approach is somewhat limited for tagging richly inflected languages, especially when in addition to the part of speech, the output of the system needs to contain morphological information (such as number, tense, and person). The problem with extracting lexical probabilities from a limited training corpus is related to the fact that statistics may not necessarily represent the use of a particular word in a particular context. In French, a word can have up to eight parts of speech, and it is very unlikely that all corresponding forms will be present in the training corpus in large enough numbers.

Our goal is to identify approaches that allow for a better estimation of the variability of tag distributions for all words that appear in the test corpus. Several paradigms have been used for disambiguating parts of speech in French. Whether one or another should be used depends on the availability of large training corpora as well as on the amount of information that the tags are used to convey. The next section explores different strategies to handle the morphological variability of French, and proposes a solution which captures variability on one hand, and frequency of patterns on the other. Section 3 gives some evidence on the power of contextual probabilities vs. lexical ones for French. Finally, the paper presents a part of speech tagger that takes into account both linguistic knowledge and statistical learning. Its novelty relies on several features: (a) the estimation of probabilities based on genotypes, (b) a fully modular architecture that allows the user flexible ordering of all independent modules, (c) an expanded tagset that gives the user the flexibility to use any derived subset, (d) the exportability of the system to other languages, and (e) the use of a mixed linguistic and statistical approach. Results are provided, as well as directions for future use of the model.

2. Strategies to capture morphological variants

Given that a word can have from two to eight different morphological types (based only on six morphological categories, such as syntactic category (noun, adjectives, verbs, etc.) and mood, tense, person, number, gender), an important step in designing a tagger is to decide which features the tagset should capture. Then, given the multitude of morphological variants (one single French verb can have up to 45 inflected forms), what is the best way to optimize the training corpus? It is clear that learning the distribution of a large variety of tags is very difficult with sparse training input. Morphological variants could be obtained via:

- **base forms:** in Table 1, the word "moyenne" has three different base forms, the masculine adjective "moyen", the feminine noun "moyenne", and the verb "moyenner". One way to capture these morphological variants could be to take the paradigm of base forms and to estimate probabilities on the different inflections. For example, in the word *moyenne*, one could estimate the probabilities of the verbal base form "moyenn-er" by the frequency of occurrences of the following endings 1ST-PERSON-SINGULAR-PRESENT-INDICATIVE, 1ST-PERSON-SINGULAR-PRESENT- SUBJUNCTIVE, 2ND-PERSON-SINGULAR-PRESENT-IMPERATIVE, 3RD- PERSON-SINGULAR-PRESENT-INDICATIVE, 3RD-PERSON-SINGULAR- PRESENT-SUBJUNCTIVE. This would almost rule out forms such as 2ND-PERSON-SINGULAR-PRESENT-IMPERATIVE, since imperative forms would be less likely to occur in narrative texts than indicative forms[1]. Also, 1st person forms would be given lower probabilities, since they are less likely to appear in news articles.
- **surface forms:** another way to capture the information could be to estimate the lexical probabilities of the words in a text. That is, for each word such as "moyenne", estimate the probability of the word given the eight morphological distinct forms. This would necessitate an extremely large body of texts in order to cover all the inflectional variations for a given word. Taking into account that there is no dis-

[1] Of course, this would also depend on the genre of the text; imperative forms would be more frequent in cookbooks for example.

ambiguated corpus of that size for French, this approach does not seem feasible.

Taking into account these previous points, we have used a new paradigm to capture the inflection of a word on the one hand, and the analyses associated to this word on the other. We call a **genotype** the set of tags that a given word inherits from the morphological analysis. For example, the French word "le" (meaning *the* or the direct object *it, him*) has two parts of speech: BD3S [PERSONAL-PRONOUN-DIRECT-3RD-PERSON-SINGULAR] and RDM [DEFINITE-MASCULINE-ARTICLE]. Thus, its genotype is the set [BD3S RDM]. Similarly, the genotype for the word "moyenne" is [JFS, NFS, V1SPI, V1SPS, V2SPM, V3SPI, V3SPS] or [jfs, nfs, v1s, v2s, v3s], depending on the tagset (see Sections 5.2 and 7.1.1 for a description of the tagsets).

Section 3.2 demonstrates that words falling in the same genotype have similar distributions of parts of speech. We will also show that using genotypes for disambiguation reduces the sparseness of training data. In some sense, this is comparable to the approach taken in Cutting et al. (1992). In this approach, they use the notion of word equivalence or ambiguity classes to describe words belonging to the same part-of-speech categories. In our work, the whole algorithm bases estimations on genotype only, filtering down the ambiguities and resolving them with statistics. Moreover, the estimation is achieved on a sequence of n-gram genotypes. Also, the refinement that is contained in our system reflects the real morphological ambiguities, due to the rich nature of the morphological output and the choice of tags. There are three main differences between their work and ours. First, in their work, the most common words are estimated individually and the less common ones are put together in their respective ambiguity classes; in our work, every word is equally treated by its respective genotype. Second, in their work, ambiguity classes can be marked with a preferred tag in order to help disambiguation whereas in our work, there is no special annotation since words get disambiguated through the sequential application of the modules. Third, and perhaps the most important, in our system, the linguistic and statistical estimations are entirely done on the genotypes only, regardless of the words. Words are not estimated given their individual of class categories; genotypes are estimated alone (unigram probabilities) or in the context of other genotypes (bi- and tri-gram probabilities).

3. Lexical Probabilities vs. Contextual Probabilities

There has been considerable discussion in the literature on part of speech tagging as to whether lexical probabilities are more important for probability estimation than contextual ones, and whether they are more difficult to obtain, given the nature of corpora and the associated problem of sparse data. On one hand, Church (1992) claims that it is worth focusing on lexical probabilities, since this is the actual weakness of present taggers. On the other hand, Voutilainen (Karlsson et al., 1995) argues that word ambiguities vary widely in function of the specific text and genre. He gives the example of the word "cover" that can be either a noun or a verb. He shows that even in the large collection of genres gathered under the Brown Corpus (Francis and Kučera, 1982) and the LOB Corpus (Johansson, 1980), the homograph "cover" is a noun in 40% of the cases, and a verb in the rest. The same homograph extracted from a car maintenance manual, *always* appears as a noun. Several experiments were run to figure out the types of ambiguities found in French and their distribution. In the tagger for French, we argue that contextual probabilities are in fact more important to estimate than lexical ones since a) there is no large training corpus for French, b) it would be nearly impossible to get a corpus covering all French morphological inflected forms. As Zipf's law predicts, even an arbitrary large training corpus would still be missing many word forms, since that corpus would have a large tail of words occurring very few times. Zipf's law holds even stronger for French.

3.1. HOW AMBIGUOUS IS FRENCH?

We selected two corpora[2], one with 94,882 tokens and the other with 200,182 tokens, in order to account for the morpho-syntactic ambiguity of French. Table 3 shows the distribution of these ambiguities for each French token. Columns 2 and 4 give the number of words corresponding to the tags in column 1. Column 3 and 5 show the percentage of words per tags in the corpus.

It is interesting to point out that despite the fact that one corpus is twice the size of the other, the distribution of the number of tags per word is nearly the same. Table 3 shows that a little more than half of the words in French texts are unambiguous, 25% of the words have two tags, 11% of the words have three tags, and about 5% of the words have from four to eight tags. Another way to quantify the word ambiguity is that, for the corpus of 94,882 tokens, there is a total of 163,824 tags, which gives an average ambiguity factor of 1.72 per word. Similarly, for the corpus of 200,182 tokens, there are 362,824 tags, which gives an ambiguity factor of 1.81 per word.

[2] Extract of the French newspapers Le Monde (Paris), September-October, 1989, January, 1990. Articles Nos. 1490 - 1875.

genotype size	**94,882 tokens**	**% of the corpus**	**200,182 tokens**	**% of the corpus**
1 tag	54570	57%	110843	58%
2 tags	24636	26%	50984	25%
3 tags	11058	11%	23239	11%
4 tags	634	.5%	3108	1%
5 tags	856	.9%	5963	2%
6 tags	2221	2%	4621	2%
7 tags	590	.5%	1069	.5%
8 tags	317	.5%	355	.1%

TABLE 3. Ambiguity of French words in two corpora of different sizes.

3.2. LEXICAL PROBABILITIES VS. GENOTYPES

In Table 4, a few words belonging to a very frequent genotype [nfs v1s v2s v3s] (noun-feminine-singular, verb-1st-person-singular, verb-2nd-person-singular, verb-3rd-person-singular) were extracted from the test corpus and probabilities were estimated with the information from the training corpora. The table shows the words in the leftmost column; the next three columns display the distribution in the three corpora (C_1, C_2, C_3), with the number of occurrences found in the training corpus ("occ" in the table), the number of times the word is tagged "nfs" and the number of times it is "v3s". Note that since these words were never "v1s" or "v2s" in the training corpus, there is no account for these parts-of-speech. Column 4 shows the total for the three corpora. Table 5 gives a total in percentage of the occurences of "nfs" and "v3s" in the training corpora. The sum of the 8 words is given followed, in the last line of the table, by the resolution of this genotype throughout the entire training corpus.

Tables 4 and 5 show that, if we were to estimate lexical probabilities, there would not be any information for the word "danse" (*dance*), since it does not appear in the training corpus. On the other hand, in capturing only the genotype [nfs, v1s, v2s, v3s] for the word "danse", the information from the training corpus of 89.15% "nfs", 10.85% "v3s" will be applied and "danse" will be correctly assigned the "nfs" tag. In this case, genotypes help with smoothing since the word itself (known or not from the training data) is ignored, and its membership in the genotype supercategory is used instead. Therefore, this strategy drastically reduces the problem of sparse data. In the Brown corpus, about 40,000 words appear five times or less (Church, 1992). When low frequency words occur with equal part-of-speech distribution, it is hardly possible to pick the right part-of-speech. Using

	Training C_1			Training C_2			Training C_3			Training C_{1-3}		
	10K words			30K words			36K words			76K words		
	occ	nfs	v3s	occ	nfs	v3s	occ	nfs	v3s	occ	nfs	v3s
laisse	1	0	1	0	0	0	0	0	0	1	0	1
masse	0	0	0	11	11	0	0	0	0	11	11	0
tâche	2	2	0	0	0	0	0	0	0	2	2	0
lutte	0	0	0	0	0	0	5	4	1	5	4	1
forme	3	3	0	61	57	4	1	1	0	65	61	4
zone	0	0	0	12	12	0	5	5	0	17	17	0
danse	0	0	0	0	0	0	0	0	0	0	0	0
place	4	4	0	10	10	0	12	12	0	26	26	0
Total:	10	9	1	94	90	4	23	22	1	127	121	6

TABLE 4. Comparing frequencies of words vs. genotypes.

	Total nfs	Total v3s
laisse	0.00 %	100.00 %
masse	100.00 %	0.00 %
tâche	100.00 %	0.00 %
lutte	80.00 %	20.00 %
forme	94.00 %	6.00 %
zone	100.00 %	0.00 %
danse	NO DATA	NO DATA
place	100.00 %	0.00 %
Total 8 words:	95.2 %	4.72 %
Total genotype:	89.15 %	10.85 %

TABLE 5. Total of “nfs” vs. “v3s” in Table 4.

genotypes is a practical way to solve that problem.

3.3. DISTRIBUTION OF GENOTYPES

To exemplify further the advantage of using genotypes, we measured their distribution through the same three corpora. Table 6 exhibits some convincing numbers: for a corpus of 10,006 tokens, there are 219 different genotypes, and for a corpus of 76,162 tokens, there are 304 unique genotypes. In other words, while the corpus is increased by 86%, the number of different genotypes increases by 27% only. Furthermore, the number of genotypes to estimate remain very low, since for a corpus of 76,162 tokens, genotypes represent only 3% of this number.

corpora	# of tokens	# of words	# of genotypes
Corpus$_1$	10006	2767	219
Corpus$_2$	34636	4714	241
Corpus$_3$	31520	5299	262
Corpus$_{1-3}$	76162	10090	304

TABLE 6. Genotype distribution.

4. Construction of the Tagger

The tagger is made of a series of stand-alone modular programs which can be combined in many different ways to allow flexibility of the system.

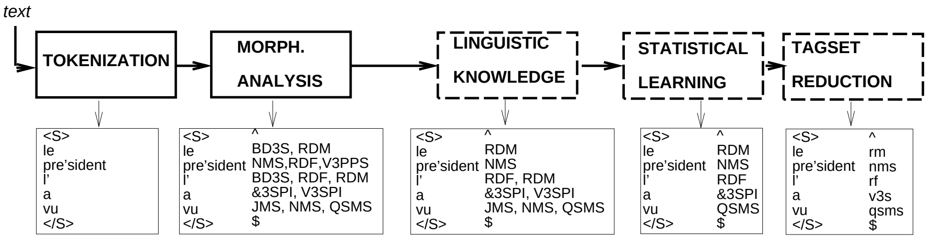

Figure 1. System Components.

Figure 1 presents a view of the algorithm. The text to be tagged is preprocessed and tokenized; then morphological analysis is performed followed by deterministic rules and statistical knowledge. Finally, the large tagset is reduced to a smaller one, as an example of tag reduction. Linguistic knowledge, statistical learning, and tagset reduction modules are surrounded by dashed boxes indicating that modules can be applied in arbitrary order. Figure 1 shows also an example of a few words in a text for each of the modules in order to demonstrate the disambiguation process.

The next sections describe the development of the tagger based on genotypes estimation; first, some issues related to tokenization are raised, then the linguistic knowledge and the statistical learning modules are explained followed by the results.

4.1. ISSUES WITH TOKENIZATION

The first step in tagging involves a series of text preprocessing modules that are used for the tokenization of the corpus.

- **Sentence boundaries:** places where sentences begin are identified and replaced by appropriate tags. As punctuation symbols play an important role in disambiguation, they are also assigned special tags.
- **Proper nouns:** the morphological dictionary contains common nouns and proper nouns, but there is a large number of proper nouns that appear in the corpus and need to be tagged; the number of proper nouns missing in the morphological dictionary is typically fairly high. Therefore, the tagger applies several heuristics. As an example, it treats each word starting a sentence as possibly having an additional *proper noun* tag; after morphological analysis, if the word inherits a new analysis, the latter one will prevail; if not, the word is identified as *proper noun* and is dynamically added to the PROPER_NOUNS dictionary. If a capitalized word is found in the middle of a sentence, it will inherit immediately the *proper noun* tag.
- **Accent restitution:** An additional difficulty due to the accents appears. In continental French, accented characters lose their accents if they become capitalized. This is valid in either sentence-initial position or in the middle of the sentence. Therefore, many accents in the text are missing. A phonology-based recovery technique is applied in order to attempt to recover these accents. Namely, an initial uppercase vowel will get an accent if it precedes a consonant in the following configuration:

    - if the word starts with the following pattern `ECV`, where `E` is the upper case character "`E`", `C` is one of the consonants or consonant pairs [b, bl, br, c, ch, cl, cr, d, dl, dr, f, fl, fr, g, gl, gr, h, j, j, l, m, n, p, ph, pl, pr, q, r, s, sl, sr, t, tl, tr, v, vl, vr, z], and V one the vowels [a, e, i, o, u, y], the acute accent is recovered.

    - if the observed word is "A" or "Etre", the accent will be respectively grave and circumflex.

- **Acronyms:** a treatment similar to the one of the proper nouns is applied here.
- **Compound words:** compound words or non-compositional words in French are to be tagged as a separate entity. They are recognized from our dictionary sources and are considered as a single lexical unit. For example, locutions such as "a priori" (*a priori*), "top secret" (*top secret*), or "raz de marée" (*tidal wave*) will be treated as single lexical entries.
- **Personal pronouns:** if two words are connected by a dash "-", and the second word is a personal pronoun, the two-word unit gets split. For example, the compound "dit-elle" (*said she*) becomes the two words "dit" and "elle".

- **Word splitting:** when all other stages are completed, the corpus is split into lexemes and translated from 8-bit characters to 7-bit ascii characters if necessary. Accents are marked with diacritic symbols following the accented letter. Example: "co^te's" is used for "côtés" (*sides*)[3].

5. Linguistic knowledge

Once the text is tokenized, morphological analysis is performed in order to disambiguate the words.

5.1. MORPHOLOGICAL ANALYSIS

Finite-state transducers (FST) are used to achieve morphological analysis. The FST is built on the model developed for Spanish morphology (Tzoukermann and Liberman, 1990) and handles mainly inflectional morphology with some derivational affixes, such as "anti-" in "anti-iranien" (*anti-iranien*), and "arrière-" (*great*) in "arrière-grand-père" (*great-grand-father*). The arclist dictionary – dictionary of finite-state transitions – was originally built using several sources, including the Robert Encyclopedic dictionary (Duval et al., 1992) and lexical information from unrestricted texts.

The FST used in the morphological stage of the tagger consisted of up to 4 distinct sub-FST's: the common names or main FST , the main proper-noun FST, which is dynamically generated from the learning corpus, and another proper-noun FST generated heuristically from the corpus to be tagged or test corpus. The last one is generated each time a new corpus is tagged. The main FST recognizes over 90,000 entries, i.e. all inflected forms, such as nouns, verbs, adjectives, as well as uninflected forms, such as adverbs, conjunctions, and other categories. Morphological analysis is performed with a high level of refinement. For example, in addition to verbal forms inflected for mood, tense, person, and number, pronouns are analyzed into several categories, such as direct, indirect, disjoint, reflexive, and so on.

5.2. FROM FEATURES TO TAGS

Once the morphological analysis is performed, one needs to translate the feature analysis into tags. We use an abbreviation of the features of the

[3] It is important to notice that this marking does not introduce any ambiguity with the French apostrophe, since apostrophes always occur after a consonant whereas the accent marks always occur after vowels.

word as its tag. For example, the tag `BD3S` stands for a third person (3) singular (S) personal pronoun (B) direct object (D).

This offers several advantages: first, it allows organization of the different categories by their syntactic feature, i.e. `verb, noun`, etc; second, the tag reflects an interesting feature hierarchy. For example, `VIP3S` which is third person present indicative verb, can be viewed in a feature hierarchy representation where verb is on top of mood, tense, number, and person. Third and consequently, rule operations can be done on any part of the structure hierarchy. For example, one can express generalizations on the tag paradigm, which simplifies the rule writing. In the following example, one can replace the tense by a metacharacter [$*$]: [`V3SPI,V3SFI,V3SSI,V3SII`] $\Rightarrow$ `V3S*I`. The rule will apply to every verb in the indicative mood (I), for every tense (*) which is in the third person (3) singular (S). The first set of tags represents the detailed morphological analysis; it corresponds to the large set of tags, i.e. 253 tags. Natural language systems, depending on what they try to achieve, vary in the number of tags they require as well as in the choice of tags. To address this issue, we left flexibility for the user so that any set or subset of tags that is desired in connection with the particular task in hand can be defined. The large set of tags can be redefined by any subset of the same tag(s) using a many-to-one mapping mechanism. In our current tagging scheme, the 253 tags are collapsed at the end of the tagging process to form a smaller set of 67 tags.

5.3. NEGATIVE CONSTRAINTS

Linguistic knowledge has been integrated in the system in the form of negative rules. Several transformational rules specify for bigrams, trigrams, and larger *n*-gram units that a particular sequence of tags is not legal for a French sentence. These rules are tightly dependent on morphological analysis. For example, the following negative constraints that list two continuous tags not admitted in French, are introduced:

- **BS3 BI1**. A BS3 (3rd person subject personal pronoun) cannot be followed by a BI1 (1st person indirect personal pronoun). In the example: “il nous faut” (*we need*) – “il” has the tag BS3MS and “nous” has the tags [BD1P BI1P BJ1P BR1P BS1P]. The negative constraint “BS3 BI1” rules out “BI1P”, and thus leaves only 4 alternatives for the word “nous”.
- **N K**. The tag N (noun) cannot be followed by a tag K (interrogative pronoun); an example in the test corpus would be: “... fleuve qui ...” (...river, that...). Since “qui” can be tagged both as an “E” (relative pronoun) and a “K” (interrogative pronoun), the “E” will be chosen by the tagger since an interrogative pronoun cannot follow a noun (“N”).

- **R V**. A word tagged with R (article) cannot be followed by a word tagged with V (verb): for example "l' appelle" (calls him/her). The word "appelle" can only be a verb, but "l"' can be either an article or a personal pronoun. Thus, the rule will eliminate the article tag, giving preference to the pronoun.

Negative constraints are examples of deterministic knowledge introduced in the system. They express linguistic relationships between the features of the words in a given n-gram, therefore performing some contextual diambiguation over word strings. These relationships could perhaps be discovered through statistical procedures, but since they are available to the human without significant effort, they are easy to implement. Each of the linguistic constraints is applied several times over the words that have only one tag. This iterative filtering process generates words with unique tags, which serve as anchors in the corpus. In this incremental fashion, anchors can create new anchors and thus enlarge the islands of disambiguated words.

6. Statistical Learning

We manually tagged a set of three corpora, containing 10,000, 30,000, and 36,000 words respectively, one from the ECI corpus – extracted from the newspaper "Le Monde"–, and two others from other news articles. Two additional corpora of 1,000 and 1,500 words were tagged for testing purposes. The test corpora were extracted from both sources to reflect the two different text styles.

A statistical model based on n-gram probabilities was implemented to find the best tag candidate for a given genotype. If **t** is a tag and **T** a tag genotype, the question is to find $P(t|T)$, so that the most likely tag for a given word can be selected. Bigram probabilities were computed in estimating the sequence of two tags given the two genotypes, i.e. $P(t_i, t_{i+1}|T_i, T_{i+1})$ and trigram probabilities, i.e. $P(t_i, t_{i+1}, t_{i+2}|T_i, T_{i+1}, T_{i+2})$. Notice here that for bigrams and trigrams, the model does not estimate a single tag occurrence but the sequence of tags.

Table 7 shows the best decisions that were made with n-gram probabilities. For a given genotype (1st column), the decision that was made over the 10,000 words training corpus (2nd column), the frequency of this case occurence (3rd column), and the strength of the decision (4th column) as explained below.

We use a *strength* score for each statistical rule based on the frequency, f, of the decision among n observations of the tag genotype. For instance, Table 7 gives f = 195 and n = 199 for the decision RDM from the tag genotype [BD3S,RDM]. The strength score assumes that f results from a binomial distribution $B(p, n)$. This is the distribution which results when n

genotype	**decision**	**freq.** f/n	**strength**
NMP P	P	82/82	98.54
BD3S NMS RDF	RDF	172/173	98.44
BD3S RDM	RDM	195/199	96.70
DMS NMS NXP RIMS W	RIMS	107/109	96.30
P RP	P	768/793	96.16
NMS pMS	pMS	30/30	96.09
NXP W	W	90/92	95.63
NMP V2SPI V2SPS	NMP	25/25	95.33

TABLE 7. Best decisions that can be made according to unigram distributions.

independent trials are made, each having probability p of the decision (and probability $1 - p$ of any other member of the tag genotype). We do not know p, but must make an estimate from the data. When $\hat{p}$ is estimated as the proportion f/n of the decision in the tag genotype, then the theory of the binomial distribution (Moore and McCabe, 1989) gives :

$$sd(\hat{p}) = \sqrt{p(1-p)/n}$$

We estimate

$$\hat{p} = \frac{f+0.5}{n+1}$$

so that neither $\hat{p}$ nor $(1-\hat{p})$ will be zero. This procedure is explained in Box (1973). We can estimate the uncertainty of $\hat{p}$ by:

$$\sqrt{\hat{p}(1-\hat{p})/n}$$

and we use the value

$$strength = (\hat{p} - \sqrt{\frac{\hat{p} * (1-\hat{p})}{n}}) * 100$$

to quantify the strength of the decision. This score represents the estimate of the probability less the estimate of the uncertainty. Notice in the above table that 25/25 has a lower strength than 30/30 which in turn has a lower strength that 82/82. The strength measure is designed to give lower values for the same f/n the smaller n is.

The interesting aspect in this model, as demonstrated in Section 3.2, is that probabilities are computed only on the genotypes, and not on the words. In using only unigram probabilities after the application of negative rules, the system disambiguates over 91% of the text. In applying bigram probabilities, system performance goes up to 93%.

7. Implementation and results

Each component of the system – tokenization, morphological analysis, deterministic rules, and statistical learning – is implemented as a stand-alone program to preserve the modularity of the system. In this fashion, it is possible to use all the modules in any desired order, except for the preprocessing and the morphology modules that are applied in a fixed order (see Figure 1 for a representation of the system). The system can be viewed as a series of operators, each of which performing some level of disambiguation of the morphological analysis. In the following sections, operators or modules are studied in different orders, so that one can scientifically test the relevance of the best operator order. The final output of the system contains all words from the original corpus grouped in three categories: (1) the correctly tagged words, (2) the incorrectly tagged words, (3) and those that are still ambiguous. The last group is particularly important; it means that the evidence for disambiguating a certain word is not sufficient at this point. When tagging text relies on the availability of training corpora but the amount of tagged text is small, leaving words without a decision seems to be better than making a decision without a strong enough level of confidence. Moreover, human taggers do not always agree with one another, and it gives the user the choice of picking the desired one.

We used the system modularity and combined the operators in many different ways. This resulted in several experiments to figure out the best path or the best order of module applications. In varying the parameters of the different modules as well as the ordering of the modules, a total of 43 plausible tagging schemes was considered, testing different orderings of (a) the deterministic stage, (b) the statistical learning with different confidence thresholds[4], (c) the application of unigram decisions, (d) and the tagset reduction.

7.1. PREVIOUS EXPERIMENTS

Three experiments were achieved in the previous version of this paper (Tzoukermann, Radev, and Gale, 1995). The first one explored the modu-

[4] We rank all possible unigram decisions according to their strength. In the different tagging schemes, we vary the strength threshold in order to achieve optimal results.

lar aspect of the system; several tagging combinations were performed in order to figure out which modular path gives the best results. Applying the operators in different order presents variants of the correct/incorrect/ambiguous tagging path. The scheme that achieves the largest percentage of correct tags is the one that applies sequentially Morphology (M), Negative Constraints (with 3 iterations) (D), Statistical Decisions with maximal coverage (A_5), and Tag Reduction (T). At this point, the system performance was 90.4% correct, 8.4% incorrect, and 1.2% ambiguous. At each iteration during the deterministic stage, the system tries to find anchors (i.e. words, which at that moment have been disambiguated). These words are used to propagate negative constraints to their neighbors. We found empirically that after 3 iterations, the number of anchors does not increase due to the large number of inherently unambiguous words in the texts.

In the second experiment, we varied the threshold so that we could analyze the effect on the performance. It turned out that a lower value of the threshold represents more (90.4% correct) but possibly incorrect (8.4% incorrect) statistical decisions, whereas a higher value gave fewer (83.4% correct) but more reliable decisions (3.9% incorrect). The last one explored the impact of the two different tagsets on the tagger performance. Interestingly, the reduction from the large tagset to the small one does not improve much (only .09%) the performance of the tagger. This is mainly due to the fact that the reduction is done mainly inside the main part-of-speech categories (i.e. verbs, nouns, etc.) and not accross the categories.

More recently, we have been working on re-implementing the part-of-speech tagger using only a finite-state machine framework. We have used a toolkit developed by Pereira et al. (1994) which manipulates weighted and unweighted finite-state machines (acceptors or transducers). Using these tools, we have created a set of programs which generate finite-state transducers from descriptions of negative constraints, as well as other transducers for the statistical learning. Statistical decisions on genotypes are represented by weights. Negative constraints are assigned the highest cost. In the earlier framework, conflicting n-gram decisions were handled in an arbitrary fashion - choosing the ones closer to an anchor. With the finite-state tools, we are able to prefer one n-gram decision over another based on their cost. For example, in order to trace the disambiguation stages, the word “moyenne” used in Table 1 was observed (genotype and resolution are represented in bold characters). Table 8 contains three steps of resolution. The first one, unigram probabilities, exhibits the genotype in the left column with the resolution in the second column and the frequency in third one. At this point, the verb tag “v3s” is in first position (with 41.54%). The second one, bigram probabilities, shows the first two genotypes providing a score of 100% and 75% in favor of the tag “jfs” for the word “moyenne”.

The third line indicates that for the bigram genotype [[nms],[jfs nfs v1s v2s v3s]], the resolution is 100% in favor of "v3s". In the context of a preceeding word being a masculine singular noun ("nms"), it is understandable that the verb tag ("v3s") is more likely to be correct. The trigram probabilities have two resolutions, one with a masculine noun on the left ("nms"), the other with a feminine noun also on the left ("nfs"), and the two cases still favor the adjective tag "jfs" for "moyenne". It shows that if the preceeding word is a noun, no matter its gender, the genotype [jfs nfs v1s v2s v3s] will be resolved by picking the adjective tag ("jfs"). This is another persuasive example of the power of n-gram genotype resolution that doesn't require lexical probabilities.

unigram probabilities		
genotype	**resolution**	**prob.**
[jfs nfs v1s v2s v3s]	**v3s**	41.54%
[jfs nfs v1s v2s v3s]	**jfs**	35.38%
[jfs nfs v1s v2s v3s]	**nfs**	23.08%

left bigram probabilities		
bigram genotype	**resolution**	**prob.**
[nfs], **[jfs nfs v1s v2s v3s]**	[nfs, **jfs**]	100.00%
[nfs nms], **[jfs nfs v1s v2s v3s]**	[nfs, **jfs**]	75.00%
[nms], **[jfs nfs v1s v2s v3s]**	[nms, **v3s**]	100.00%

trigram probabilities		
trigram genotype	**resolution**	**prob.**
[nfs nms], **[jfs nfs v1s v2s v3s]**, [a b p]	[nms, **jfs**, p]	50.00%
[nfs nms], **[jfs nfs v1s v2s v3s]**, [a b p]	[nfs, **jfs**, p]	50.00%

TABLE 8. An example of genotype resolution.

Table 9 presents the current performance of the tagger according to the different n-gram probabilities and the application of the linguistic rules. The results are based on the small training corpus of 10,000 words and we believe the performance will get better in using the entire training corpus. The figures shown in Table 9 reflect the percentage of words that are disambiguated correctly by the tagger. The remainder to 100% consists of words that have been incorrectly disambiguated.

	unigrams	**bigrams**
10K-word training corpus	92.0%	93.0%

TABLE 9. Tagger performance with n-gram probabilities and negative constraints.

7.1.1. *Comparative study of the two tagsets*

The goal in building a flexible tagset is to allow the user to pick whatever set is necessary for a particular application. Even though the reduction in size from the large to the small set is of 74%, this gain was not reflected in the tagged text. In order to understand better this phenomenon, we observed the distribution of tags in a corpus of over 200,000 words. We collected the frequencies of the tags in the two different sets and the top 35 most frequent tags appear in Table 10. The large tagset represents the morphological features obtained by the morphological analyzer. In constructing the small set, we eliminated a large number of morphological features that are of relatively little use at the syntactic level; for example, mood and tense for verbs, reflective, disjoint, and subject position for personal pronouns. Additionally, auxiliaries and verbs – a total of 93 different forms – were collapsed in fewer categories and only the person and the number were kept, which resulted in a total of 13 different tags. All personal pronouns were collapsed and the numbers went from 79 to 9. Column 2 of Table 10 shows the large tagset with the number of tag occurrences (column 1), and the tag meaning (column 3). Column 4, 5, 6 show similar information for the small tagset. As an example, we highlighted the different occurences of third person singular verbs on the left of the table that correspond to a single tag occurence in the small tagset on the right side.

Any subset of this large tagset can be (re)defined by the user with a very simple mapping mechanism. This is an important feature of the system design since it makes the tagger adaptable to different NLP applications requiring different sets of tags or morphological variants.

8. Related Research

A number of taggers and tagging methods are available. For the last decades, part of speech tagging systems have generally followed either a rule-based approach (Klein and Simmons, 1963), (Brill, 1992), (Voutilainen, 1993), or a statistical one (Bahl and Mercer, 1976), (Leech, Garside, and Atwell, 1983), (Merialdo, 1994), (DeRose, 1988), (Church, 1989), (Cutting et al., 1992). Statistical approaches often use Hidden Markov Models for estimating lexical and contextual probabilities, while rule-based systems capture

Num. of occ.	**Large Tagset**	**Description**	**Num. of occ.**	**Short Tagset**	**Description**
31562	P	prep.	31562	p	prep.
19567	.	punct.	19567	x	punct.
12398	NFS	fem. sg. n.	12398	nfs	fem. sg. n.
11792	NMS	masc. sg. n.	11792	nms	masc. sg. n.
8650	A	adv.	8650	a	adv.
7445	U	proper n.	8169	**v3s**	**3rd sg. verb/aux.**
7375	RDF	fem. def. art.	7445	u	proper n.
6975	^	beg. of sent.	7375	rf	fem. def. art.
6975	$	end of sent.	7074	r	indef. art.
4631	W	numeral	6975	^	beg. of sent.
4467	**V3SPI**	**3rd sg. pres. ind. verb**	6975	$	end of sent.
4363	NMP	masc. pl. n.	6208	b	pers. pron.
4171	CC	coord. conj.	4631	z	numeral
4002	i	inf. verb	4444	v	inf. verb/aux..
3958	NFP	fem. pl. n.	4363	nmp	masc. pl. n.
3726	RDM	masc. def. art.	4171	cc	coord. conj.
3471	QSMS	masc. sg. past part.	3958	nfp	fem. pl. n.
3379	JMS	masc. sg. adj.	3910	qsms	masc. sg. past part.
3299	RDP	partitive art.	3726	rm	masc. def. art.
2883	JXS	sg. inv. (gender) adj.	3379	jms	masc. sg. adj.
2597	CS	subord. conj.	3275	js	sg. adj.
2563	NMX	masc. inv. (number) n.	2898	v3p	3rd pl. verb
2469	JFS	fem. sg. adj.	2597	cs	subord. conj.
1995	RIMS	masc. sg. indef. art.	2571	nm	masc. n.
1979	JMP	masc. pl. adj.	2469	jfs	fem. sg. adj.
1743	E	rel. pron.	1979	jmp	masc. pl. adj.
1739	RIFS	fem. sg. indef. art.	1976	jp	pl. adj.
1511	BS3MS	3rd sg. subj. pers. pron.	1618	bms	masc. pers. pron.
1449	**&3SPI**	**3rd sg. pres. ind. aux.**	1251	jfp	fem. pl. adj.
1392	BR3S	3rd sg. refl. pers. pron.	1088	jms	masc. sg. adj.
1319	JXP	pl. inv. (gender) adj.	876	qp	pres. part.
1251	JFP	fem. pl. adj.	811	qsmp	masc. sg. past part.
1129	V3PPI	3rd pl. pres. ind. verb	746	h	acronym
1063	&3PPI	3rd pl. pres. ind. aux.	744	qsfs	fem. sg. past part.

TABLE 10. Most frequent 35 tags in a corpus of 200,000 words for large and small tagsets.

linguistic generalities to express contextual rules. Most of these approaches have benefited from large tagged corpora mentioned above, which make the training and testing procedures feasible.

Brill and Marcus (1992) and Brill (1992) proposed a simple and powerful corpus-based language modeling approach that learns a series of transformational rules that are then applied in sequence to a test corpus to produce predictions. The learning approach combines a large training corpus, a baseline heuristic for selecting initial default values, and a set of rule templates defining classes of transformational rules that use particular neighborhood characteristics as the grounds for changing a particular current value. For example, in the part of speech tagging application, the baseline heuristic might be to assign to each ambiguous word whatever tag is most often correct in the training corpus, and the templates, defined here over a window that includes a context of two tags on each side, will apply the rules if the tag transition needs to be changed. It seems clear, though, that the performance of Brill's tagger is contingent on the availability of a large training corpus. Since rules are acquired automatically, a small training corpus cannot provide enough empirical data for the acquisition of a large number of rules.

Chanod and Tapanainen (1995) compare two tagging frameworks for French, one that is statistical, built on the Xerox tagger (Cutting et al., 1992), and another based on linguistic constraints only. The constraints can be 100% accurate or describe the tendency of a particular tagging choice. The constraint-based tagger is proven to have better performance than the statistical one, since rule writing is more handlable or more controllable than adjusting the parameters of the statistical tagger. The tagset used is very small (37 tags), including a number of word-specific tags (which reduces further the number of tags), and does not account for several morphological features, such as gender, number for pronouns, etc. Moreover, categories that can be very ambiguous, such as coordinating conjunctions, subordinating conjunctions, relative and interrogative pronouns tend to be collapsed; consequently, the disambiguation problem is too simplified (therefore giving high performance) and results cannot be compared since the ambiguities do not lie at the same word level.

Merialdo (1994) makes comparisons among different tagging schemes using classic Viterbi algorithms on the one hand, and Maximum Likelihood Estimation tagging on the other. In (Merialdo, 1994), results show that the estimation of the model parameters counting the relative frequencies of a large quantity of hand-tagged corpus gives better results than training using Maximum Likelihood.

To contrast with these different approaches, our work has attempted to go in more nuances and refinements of the linguistic subtleties. Therefore, the disambiguating task is rendered more complex. In the spectrum of parts of speech taggers as described above, the originality of our work lies in the linguistic nuances and subtleties that are encoded in the system. Part of speech taggers could be divided in two categories, these which discriminate simple part of speech categories, such as (Church, 1989) and (Chanod and Tapanainen, 1995) and those like (Voutilainen, 1993), which detect noun phrases. Like Voutilainen's system, ours provides more morpho-syntactic information and therefore tackles more of the linguistic ambiguities. With its numerous encoded morphological features, not only is it more flexible to different NLP applications, but it can be also viewed as the first step towards a shallow parsing system.

9. Remarks and Future Work

This paper presents techniques for assigning the most appropriate tag among all the ones generated by the morphological analysis for each French word in unrestricted texts. We explored different strategies to capture both morphological and syntactic variants. With a restricted amount of tagged data, lexical probabilities were shown to be limited in their predictive ability. Even large amounts of training data would not solve the problem of sparseness in training data. The solution to this problem was to select the set of tags associated with a word – the genotype – and apply linguistic knowledge and statistical learning on this unit. This approach exhibits also an elegant way of smoothing probabilities, in giving estimates of unseen words in tagging. A tagger for unrestricted text was then built that took into account the limitation of training data, and combined empirical and symbolic methods to disambiguate word parts of speech. Among others, the contribution of this work resides in the successful approach of using the genotype to estimate n-gram probabilities. We are in the process of improving the system performance and are exploring the portability of the system to other languages, such as Spanish.

Acknowledgments

We wish to thank Ido Dagan, Judith Klavans, and Diane Lambert for their valuable comments throughout this work.

References

Bahl, Lalit R. and Robert L. Mercer. 1976. Part-of-speech assignement by a statistical decision algorithm. *IEEE International Symposium on Information Theory*, pages 88–89.

Box, G.E.P. and G.C. Tiao. 1973. *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, Mass.

Brill, Eric. 1992. A simple rule-based part of speech tagger. In *Third Conference on Applied Computational Linguistics*, Trento, Italy.

Brill, Eric and Mitch Marcus. 1992. Tagging an unfamiliar text with minimal human supervision. In *Proceedings of the AAAI Symposium on Probabilistic Approaches to Natural Language*, pages 10–16. American Association for Artificial Intelligence.

Chanod, Jean-Pierre and Pasi Tapanainen. 1995. Tagging French – comparing a statistical and a constraint-based method. In *EACL SIGDAT Workshop*, Dublin, Ireland. Association for Computational Linguistics - European Chapter.

Church, Kenneth W. 1989. A stochastic parts program noun phrase parser for unrestricted text. In *IEEE Proceedings of the ICASSP*, pages 695–698, Glasgow.

Church, Kenneth W. 1992. Current practice in part of speech tagging and suggestions for the future. In Simmons, editor, *Abornik praci: In Honor of Henry Kučera*. Michigan Slavic Studies.

Cutting, Doug, Julian Kupiec, Jan Peterson, and Penelope Sibun. 1992. A practical part-of-speech tagger. Trento, Italy. Proceedings of the Third Conference on Applied Natural Language Processing.

DeRose, Stephen. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics*, 14(1):31–39.

Duval et al., Alain. 1992. *Robert Encyclopedic Dictionary (CD-ROM)*. Hachette, Paris.

Francis, W. Nelson and Henry Kučera. 1982. *Frequency Analysis of protectEnglish Usage: Lexicon and Grammar*. Houghton Mifflin Company, Boston, Massachusetts. with the assistance of Andrew W. Mackie.

Johansson, Stig. 1980. The LOB Corpus of British English Tests: presentation and comments. *Association for Literary and Linguistic Computing*, 1:25–36.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Antilla. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin, New York.

Klein, S. and R. F. Simmons. 1963. A grammatical approach to grammatical tagging coding of English words. *JACM*, 10:334–347.

Leech, Geoffrey, Roger Garside, and Erik Atwell. 1983. Automatic grammatical tagging of the LOB corpus. *ICAME News*, 7:13–33.

Merialdo, Bernard. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–172.

Moore, D.S. and G.P. McCabe. 1989. *Introduction to the Practice of Statistics*. W. H. Freeman, New York.

Pereira, Fernando, Michael Riley, and Richard Sproat. 1994. Weighted rational transductions and their application to human language processing. In *ARPA Workshop on Human Language Technology*, pages 249–254. Advanced Research Projects Agency, March 8–11.

Tzoukermann, Evelyne and Mark Y. Liberman. 1990. A finite-state morphological processor for Spanish. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, Helsinki, Finland. International Conference on Computational Linguistics.

Tzoukermann, Evelyne, Dragomir R. Radev, and William A. Gale. 1995. Combining linguistic knowledge and statistical learning in French part-of-speech tagging. In *EACL SIGDAT Workshop*, Dublin, Ireland. Association for Computational Linguistics - European Chapter.

Voutilainen, Atro. 1993. NPtool, a detector of English noun phrases. Columbus, Ohio. Proceedings of the Workshop on very large corpora.

INDEX