Combining Linguistic Knowledge and Statistical Learning in French Part-of-Speech Tagging

Evelyne Tzoukermann, Dragomir R. Radev, and William A. Gale
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974–0636
evelyne,s_radev,gale@research.att.com

Abstract

This paper presents a new part-of-speech tagger that takes into account both linguistic knowledge and statistical learning. Its novelty relies in several aspects: (a) a fully modular architecture that allows the user flexible use of each independent module, (b) an expanded tagset that gives the user the flexibility to use it directly or use any individually defined tag subset, (c) the exportability of the system to other languages, and (d) the use of a mixed linguistic and statistical approach. Preliminary results are provided, as well as directions for future use of the model.

1 Introduction

Tagging part-of-speech on unrestricted texts for a language other than English raises a certain number of questions: if the language has a rich morphology, one needs to analyze the words in the sentence since the inflected forms are not in the dictionary; an electronic dictionary is necessary to write a morphological analyzer. At the morphology level: what should be the level of morphological refinement, what should be the morphological features selected to analyze words? These questions have important consequences on the development of the part-of-speech tagger, mainly, what kind of grammatical tags should be chosen. Additionally, as the tagger will be used in a natural language (NL) application, it would be good if it were not limited by a fixed size set of tags. This paper attempts to address these questions and answer them. We developed a finite-state morphological analyzer for French, solved many problems related to tokenization, and implemented a tagger that proceeds to part-of-speech disambiguation through the application of constraint rules and statistical learning. Finally, we present results of the tagger in progress and discuss ways to improve its performance.

2 Corpora and tokenization

2.1 Data collection

The input to the system is any unrestricted text from French. In the past couple of years, we have accumulated a fair number of French corpora. Each corpus has been encoded in a scheme similar to the one used for English (Church, 1991), (Yarowsky, 1992). The encoding scheme allows one to query the text using words in a Unix regular expression mode. One can extract any type of collocation, or any type of pattern occurring in a particular corpus or in all corpora. Data were encoded in 8-bit format to handle French accents. A sample of the text processed for this paper is presented in Table 1 [1], 900,000 words purchased from a CD-ROM of Le Monde (LEM, 1992).

Ce chorégraphe fait partie de cette génération montante dont les oeuvres sont teintées de spiritualité, voire d'un certain sens du sacré ou de religiosité. Catherine Diverrès le suit immédiatement avec Fragments, esquisse d'une morale qui se forgerait dans une orgie pasolinienne comme dans les sermons de Bossuet. Bigre! Duroure et Diverrès attirent à eux les meilleurs danseurs. Leur univers politique, souvent raffiné, pèche encore par excès d'images, d'intentions.

Table 1: Corpus sample of "Le Monde" newspaper

.

The collection of French corpora is of 50.2 million French words, including 5.5 million of Le Monde newspaper provided by the ECI [2] , 1.6 million of French daily news, 18.8 million words from

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

[2] We would like to thank the ECI (European Community Initiative) for letting us use these data.

the French side of the HANSARD Canadian parliament proceedings, and others. For the development of the tagger, a small 10,000-word subset of the ECI as well as a 1,000-word subset from another newswire corpus were extracted and tagged by hand.

2.2 Tokenization

The first step in tagging involves a series of text preprocessing modules that are applied to the corpus so that a text like in Table 1 can be tokenized:

- **Sentence boundaries:** places where sentences begin are identified and replaced by appropriate SGML 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 the number of proper nouns that appear in the corpus that needs to be tagged and which are 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_NAMES 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”, “top secret”, or “raz de marée” will be treated as unique lexical entries.
- **Personal pronoun:** 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 the other processes are achieved, the corpus is split into lexemes and translated from 8-bit characters to 7-bit ascii characters. Accents are marked with diacritic symbols following the accented letter. Example: “co^te's” for “côtés”.

3 Linguistic knowledge

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

3.1 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 is originally built in using some of the entries of the Robert Encyclopedic dictionary (Duval et al., 1992) and increased through acquisition of lexical information from unrestricted texts.

The FST used in the morphological stage of the tagger can consist of up to 4 distinct sub-FST's: the main non-proper-noun FST, the main proper-noun FST, the proper-noun FST which is dynamically generated from the learning corpus, and 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, ajectives, as well as uninflected forms, such as adverbs, pronouns, 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, disjunct, reflexive, and so on.

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

This offers several advantages: first, it allows to organize the different categories by their main part-of-speech, 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 such as in Figure 1.

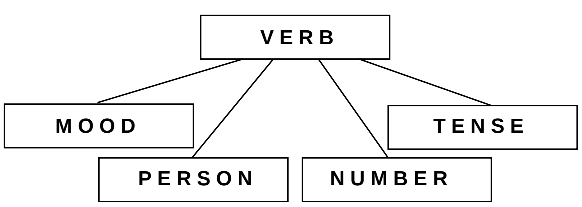

Figure 1: Feature hierarchy

.

Third and consequently, rule operations can be done on any part of the structure hierarchy. For example, one can express generalizations on one feature of the tag, which simplifies the rule writing. In the following example, one can replace the tense by a star: [`V3SPI,V3SFI,V3SSI,V3SII`] $\Rightarrow$ `V3S*I`. The rule will apply to every verb at the indicative mood (I), for every tense (*) which is at the third person (3) singular (S). The first set of tags represent the detailed morphological analysis; it corresponds to the large set of tags, i.e. 253 tags. NL 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 complete flexibility to the user and let him/her define any set or subset of tags that is desired in connection with the particular task in hand. The large set of tags can be redefined by any subset of the same tags using a many-to-one mapping. 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.

3.3 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 are introduced:

- **BS3 BI1** (3rd person subject personal pronoun; 1st person indirect personal pronoun). In the example: "il nous faut" (we need) – the morphological tags are BS3MS for "il" and (BD1P BI1P BJ1P BR1P BS1P) for "nous". The negative constraint "BS1 BI1" rules out "BI1P", and thus leaves only 4 alternatives for the word "nous".
- **N K** (noun; interrogative pronoun): an example in the test corpus would be: "... fleuve qui ..." (...river, that...). Since "qui" can be tagged as both 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** (article; 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 by linguists. They express linguistics relationships between the features of the words in a given n-gram, therefore performing some contextual disambiguation over word strings. These relationships could be perhaps discovered through statistical procedures, but as they are available to the human without effort, they are easy to implement.

Each of the linguistic constraints is applied several times over the anchors of the corpus. This way anchors can create new anchors and thus enlarge the islands of disambiguated words.

4 Statistical Learning

A minimal training corpus of 10,000 words was manually tagged from the ECI corpus. An additional 1,000 words were tagged for testing purposes. The test corpus was extracted from another newswire source so that the style of the text would be different from the training one, in order for the testing to be as fair as possible.

A statistical model based on unigram probabilities was implemented to find the best tag candidate for a given series of tags. If $\mathbf{t}$ is a tag and $\mathbf{T}$ a tag genotype, i.e. all possible tags that can be assigned to a word from a morphological viewpoint, the issue is to find $P(\mathbf{t}|\mathbf{T})$, so that the most likely tag for a given word can be selected. The model was implemented using the training corpus and the following data were provided:

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 2: Best decisions that can be made according to unigram distributions

Table 2 shows for a given genotype (1st column), the decision that was made over the training corpus (2nd column), the frequency of this case occurence (3rd column), and the power of the decision (4th column) explained below.

We use a cautious *strength* score for each statistical rule based on the frequency, f, of the decision among n observations of the tag genotype. For instance, Table 2 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 binary distributiuon $B(p, n)$. This is the distribution which results when n 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, page 398) 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 and Tiao, 1973, pages 34–36). We can estimate the uncertainty of $\hat{p}$ by:

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

and we use the strength

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

for the decision. This score represents our estimate of the probability less our 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 to notice in this model is that a) only the unigram probabilities are computed, and b), they are computed only on the tags. With this simple model, and after the application of negative rules, the system disambiguates over 91% of the text. Bigram and trigram probabilities should be estimated to improve the system performance. Attempts have been done to compute them from the 10k training corpus, but the small amount of text does not suffice to estimate reliable probabilities.

5 Implementation and results

Each module of the system is implemented separately 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. Also, each module has a single and well-defined task which is clearly distinct from the others.

Our system can be viewed as a series of operators, each of them performing some level of disambiguation of the morphological analysis. In the following sections, operators 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 training corpus and that the amount of tagged text is of a small size (10,000 French words as opposed to 1 million English words of the Brown corpus (Francis and Kučera, 1982) for example), leaving words without decision seems to be better than making a decision without a good enough level of confidence.

It was interesting to make use of the modularity and try to combine the operators in many possible ways. This resulted in several experiments that were led in order to figure out the best path or the best order of module applications. A total of 43 plausible tagging schemes were considered, testing different orderings of (a) the deterministic stage, (b) the statistical learning with different confidence thresholds, (c) the application of unigram decisions, (d) and the tagset reduction.

5.1 Analysis by tagging schemes

Table 3 compares several tagging schemes. M represents the morphological operator, D, the deterministic (negative constraint) stage, Ax, the application of unigram decisions with a threshold of x, and T, the tagset reduction.

tagging scheme	%correct	% incorrect	%ambiguous
(M)	53.5	1.0	45.7
(M)D	60.9	1.2	38.1
(M)DA5	89.3	9.4	1.3
(M)DT	64.7	1.3	34.1
(M)T	57.3	1.1	41.7
(M)DA5T	90.4	8.4	1.2
(M)DA5DT	90.2	8.7	1.1
(M)A90DT	74.7	1.4	23.9
(M)TA5	90.0	8.9	1.1
(M)DA90DT	78.4	1.8	19.9

Table 3: Results of the different tagging schemes

The table demonstrates that at the end of the morphological stage, 53% of the corpus has a single, unique, and correct tag, 1% of the words are incorrectly tagged and 45.7% are still ambiguous. The deterministic stage increases the percentage of correct tags by almost 7%, while the statistical stage with the maximum coverage (i.e. 5) provides almost 90% correct tags. Finally, the reduction of the tagset from the large set to the small one, improves further the performance of the tagger, which disambiguates correctly 90% of the tags. In fact, the number is, at first, surprisingly low; when the tagset is reduced from 253 tags to 67 tags, the tagger performance goes only from 89.3% to 90.4%. These aspects are discussed below in the section on tagset.

Applying the operators in different orders 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 (A5), and Tag Reduction (T).

5.2 Analysis by threshold

Table 4 reflects the differences in performance of the tagger when only the threshold of the statistical operator varies. A lower value of the threshold represents more (possibly) incorrect statistical decisions; a higher value – fewer (but more reliable) decisions.

tagging scheme	%correct	% incorrect	%ambiguous
(M)DA5T	90.4	8.4	1.2
(M)DA30T	89.8	8.3	1.9
(M)DA45T	89.1	7.9	3.0
(M)DA60T	83.4	3.9	12.7
(M)DA75T	81.8	2.7	15.6
(M)DA90T	76.8	1.6	21.7
(M)DA100T	53.5	1.0	45.7

Table 4: Analysis of statistical decisions

5.3 Analysis by tagset

Table 5 presents the different tagging schemes with the reduction to the small set of tags at different levels. Because of the large discrepancy in number between the large tagset (253) and the small one (67), one might have thought that there would be a significant difference each time the tagset was collapsed. The numbers in Table 5 do not verify this hypothesis, and rather show that the difference in performance is small when using different versions of the tagset. These differences are discussed in the following section.

tagging scheme	%correct	%incorrect	%ambiguous
(M)	53.5	1.0	45.7
(M)T	57.3	1.1	41.7
(M)A5	89.1	9.7	1.2
(M)A5T	90.2	8.7	1.1
(M)DA90	73.3	1.6	25.3
(M)DA90T	76.8	1.6	21.7
(M)A5D	88.7	10.3	1.0
(M)A5DT	89.8	9.3	0.9

Table 5: Comparison between the two tagsets

5.3.1 Comparison study between the two tagsets

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 were of no use at a syntactic level; for example, mood and tenses for the verbs, reflective, disjunct, and subject position for personal pronouns. Additionally, auxiliaries and verbs – a total of 93 different forms – were collapsed in the same category and only the person and the number were kept, which resulted in a total of 13 different tags. All the different personal pronouns were collapsed into one category and the numbers went from 79 to 9.

Even though the reduction from the large to the small set is of 74%, this gain, as noticed earlier, 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 6.

As said earlier, this comparison study is made between the large tagset which is of fixed size, and the small tagset which we defined of 67 tags. Nevertheless, it is clear that any subset of the large tagset can be (re)defined by the user using a very

Freq.	Lrage Tagset	Description	Freq.	Small Tagset	Description
31562	P	preposition	31562	p	preposition
19567	.	punctuation	19567	x	punctuation
12398	NFS	feminine singular noun	12398	nfs	feminine singular noun
11792	NMS	masculine singular noun	11792	nms	masculine singular noun
8650	A	adverb	8650	a	adverb
7445	U	proper noun	8169	v3s	verb 3rd person singular
7375	RDF		7445	u	proper noun
6975	$	sentence end	7074	r	indefinite article
4467	V3SPI	verb 3rd person singular present indicative	6975	$	sentence end
4363	NMP	noun masculine plural	6208	b	personal pronoun
4171	CC	coordinating conjunction	4631	z	numeral
4002	i	verb infinitive	4444	v	verb infinitive
3958	NFP	noun feminine plural	4363	nmp	noun masculine plural
3726	RDM	definite masculine article	4171	cc	coordinating conjunction
3471	QSMS	past participle masculine singular	3958	nfp	noun feminine plural
3379	JMS	masculine singular adjective	3910	qsms	past participle masculine singular
3299	RDP	definite partitive article	3726	rm	definite masculine article
2883	JXS	singular adjective invariable in gender	3379	jms	masculine singular adjective
2597	CS	subordinating conjunction	3275	js	singular adjective
2563	NMX	masculine noun invariable in gender	2898	v3p	verb 3rd person plural
2469	JFS	feminine singular adjective	2597	cs	subordinating conjunction
1995	RIMS	indefinite masculine singular article	2571	nm	noun masculine
1979	JMP	masculine plural adjective	2469	jfs	feminine singular adjective
1743	E	relative pronoun	1979	jmp	masculine plural adjective
1739	RIFS	indefinite feminine singular article	1976	jp	plural adjective
1713	*****	unknown word	1713	*****	unknown word
1511	BS3MS	personal pronoun subject masculine 3rd person singular	1618	bms	personal pronoun masculine
1449	&3SPI	auxiliary 3nd person singular present indicative	1251	jfp	feminine plural adjective
1392	BR3S	personal pronoun reflechi 3rd person singular	1088	jms	masculine singular adjective
1319	JXP	plural adjective invariable in gender	876	qp	present participle
1251	JFP	feminine plural adjective	811	qsmp	past participle masculine plural
1129	V3PPI	verb 3rd person plural present indicative	746	h	acronym
1063	&3PPI	auxiliary 3rd person plural present indicative	744	qsfs	past participle feminine singular

Table 6: Frequency of tag appearance in a corpus of 200,000 words

simple mapping mechanism. This is an important feature of the system design since it makes the tagger adaptable to different NLP applications requiring different number of tags.

6 Related Research

There have been a fair number of works in part-of-speech tagging for the last decades that 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 et al., 1983), (Merialdo, 1994), (DeRose, 1988), (Church, 1989), (Kupiec, 1993)). Statistical approaches often use Hidden Markov Models for estimating lexical and contextual probabilities, while rule-based systems capture linguistic generalities to express contextual rules. Most of these works have benefited from large tagged corpora, making feasible the training and testing procedures. In French for example, because no publicly available tagged corpus exists, other techniques must be discovered to tackle this problem.

7 Remarks and Future Work

We present here the first generation of a part-of-speech tagger for unrestricted French. The contribution of this works resides in a flexible tagset, a language independent system so that other languages can be tagged, and a combination of empiric and symbolic approaches. We are exploring the portability of the system to other languages, such as Spanish or other Romance languages where (a) morphological analysis is available and (b) the grammar of the language is closely related so that linguistic knowledge can be interchangeable. We are in the process of improving the system in several ways: first, in adding more training data, second, in adding more negative constraints, and third, in applying contextual probabilities (bigrams and trigrams). Finally, in order to reduce the amount of manual input, we are exploring ways to apply some unsupervised learning strategy to a large amount of corpus in order to automatically learn linguistic constraints as well as contextual probabilities.

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

References

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

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

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

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

Kenneth W. Church. 1991. Concordances for parallel text. Oxford, England. Seventh Annual Conference of the UW Centre for the New OED and Text Research.

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

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

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

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

Julian Kupiec. 1993. Augmenting a hidden markov model for phrase-dependent word tagging.

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

1992. L'histoire au jour le jour 1944-1991. In *version 1992.* CD-ROM.

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

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

Evelyne Tzoukermann and Mark Y. Liberman. 1990. A finite-state morphological processor for spanish. In *Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, Finland. International Conference on Computational Linguistics.

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

David Yarowsky. 1992. Conc: Tools for text corpora. In *11222-921222-29TM*, Murray Hill, N.J., USA. Technical Memorandum, AT&T Bell Laboratories.