*Behavior Research Methods, Instruments, & Computers*
*2000, 32 (3), 468-481*

# Automatic disambiguation of morphosyntax in spoken language corpora

CHRISTOPHE PARISSE and MARIE-THÉRÈSE LE NORMAND
*Institut National de la Santé et de la Recherche Médicale, Paris, France*

The use of computer tools has led to major advances in the study of spoken language corpora. One area that has shown particular progress is the study of child language development. Although it is now easy to lexically tag every word in a spoken language corpus, one still has to choose between numerous ambiguous forms, especially with languages such as French or English, where more than 70% of words are ambiguous. Computational linguistics can now provide a fully automatic disambiguation of lexical tags. The tool presented here (POST) can tag and disambiguate a large text in a few seconds. This tool complements systems dealing with language transcription and suggests further theoretical developments in the assessment of the status of morphosyntax in spoken language corpora. The program currently works for French and English, but it can be easily adapted for use with other languages. The analysis and computation of a corpus produced by normal French children 2–4 years of age, as well as of a sample corpus produced by French SLI children, are given as examples.

Automatic analysis of transcripts is not always as simple as it should be. Some of the tasks involved are quite tedious, although computer tools are already a great help. One of these tasks is the disambiguation of lexically tagged texts. In a language such as French or English, more than 70% of the words in a full adult lexicon (more than 100,000 words) are ambiguous. Smaller lexicons have fewer ambiguous words but will produce omission errors. When creating a lexicon, it is very difficult to decide in advance that a word is not going to be ambiguous in a given corpus. It is better to use a full child or adult lexicon and choose from the whole set of lexical possibilities. This task can be very time consuming when analyzing a large transcript. Fortunately, it is possible to make this process fully automatic by using an advanced part-of-speech program that can tag and disambiguate a corpus in a few seconds, with an accuracy rate that may be better than or about the same as human processing accuracy. Also, the adequacy of such automatic processing shows that the morphosyntax of child language is very consistent, in itself and in relation to adult language.

Morphosyntactic analysis involves the determination of the syntactic category and the morphological decomposition of a word. The categories used are the same as those found in a dictionary. When words are viewed in isolation, they are often ambiguous. For instance, in English, the word *back* is ambiguous between a noun referring to a part of the body, a preposition referring to a direction, a verbal particle referring to returning (e.g., *give back*), and a verb meaning to support something. Or, to take an example from French, it is necessary to determine whether the word *porte* corresponds to the feminine noun *porte* (door) or to the conjugated verb *porter* (to carry, wear) in the present tense, either in the first person or in the third person.

To demonstrate how important it is to be able to automatically resolve word ambiguities, it is interesting to give some specific figures for French. A 2-year-old child can use the word *la* in three different forms in French: *la* singular feminine article, *la* singular feminine object pronoun, and *là* adverb of place. Once the corpus has been transcribed in written format, the ambiguity between *la* and *là* is automatically resolved. Nevertheless, the ambiguity between the two forms of *la* still needs to be addressed. With large corpora, the number of ambiguous cases may be very high. In the child database described below, there are 95,000 words, with 66,800 cases of ambiguity. If manual tagging takes 10 sec per word, it will take at least 16,000 min or about 270 h to hand-disambiguate the corpus.

In this article, we describe a way of automating this process. The method uses a tagger called POST (part-of-speech tagger) based on a Markov model to aid linguistic analysis. Then, we give examples of the use of such software in French and show how it adapts to English.

## STATE OF THE ART

There is currently a wide variety of software tools designed to help the researcher, linguist, psychologist, or psycholinguist study spoken language corpora. For child language corpora (e.g., Baker-Van den Goorbergh, 1994; Long & Fey, 1995a, 1995b; MacWhinney, 1991; Mac-

This work was supported by Grant 4U009B from INSERM, France: Contrat de Recherche Inserm. The authors thank Isabelle Barrière for proofreading this text and Brian MacWhinney for his suggestions and considerable help that have clarified the presentation and improved the text. Correspondence should be addressed to C. Parisse, Laboratoire de neuropsychologie de l'enfant, Bâtiment Pharmacie, 3ème étage, Hôpital de la Salpêtrière, 47 Boulevard de l'Hôpital, 75651 Paris Cedex 13, France (e-mail: parisse@ext.jussieu.fr).

Whinney & Snow, 1985; Miller & Chapman, 1982), many of these systems perform some sort of morphological analysis. Examples include the morphological decomposition described by Miller and Chapman (1983) and Cappelli, Maccari, and Pfanner (1991) in order to compute the *mean length of utterance*. Other kinds of statistical automatic assessment of corpora of impaired speech have been carried out by Perkins (1994) and Perkins, Catizone, Peers, and Wilks (1997). More complex linguistic evaluations, such as the *language assessment, remediation, and screening procedure* (LARSP) created by Crystal, Fletcher, and Garman (1976), must be performed by hand. There have been some attempts to provide a software aid to the production of this assessment (e.g., Bishop, 1984), although the procedure was actually more suited for training people in the system than for carrying out the analysis per se.

For written discourse, there are many part-of-speech taggers that can be found on the Internet, including the Xerox Part of Speech Tagger (ftp://parcftp.xerox.com/pub/tagger/), the Brill Transformation Based Tagger (ftp://ftp.cs.jhu.edu/pub/brill/Programs/ or http://www.cs.jhu.edu/~brill/), QTAG (http://www-clg.bham.ac.uk/oliver/java/qtag/), the TreeTagger (http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html), CLAWS (http://www.comp.lancs.ac.uk/ucrel/claws/), XRCE Part of Speech Disambiguators (http://www.rxrc.xerox.com/research/mltt/Tools/pos.html), the MBT Tagger (http://ilk.kub.nl/~zavrel/tagtest.html), and LT POS (http://www.ltg.ed.ac.uk/software/pos/index.html).

## Basic Principles of Part-of-Speech Tagging

Part-of-speech taggers are rooted in the principle that the syntactic category of a word can be determined by the context of the syntactic categories of the adjacent words. This applies to any word; thus, the categories of the adjacent words are determined by the categories of the words adjacent to them, the categories of these adjacent words by the categories of the following words, and so on. In order to determine the word categories, it is thus necessary to consider the whole sentence. The whole succession of categories, from the first to the last word of a sentence, can be rated as syntactically correct or not. In a real implementation, the analysis of a word relies on hypotheses stemming from the previous words and is not validated before reaching the end of the sentence. For example

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| . | the | single | can | put | him | at | ease | . |
| pct | det | adj | n | v:past | pro | prep | n | pct |
| pct | det | n | v:modal | v:inf | pro | prep | n | pct |

Two analysis of this sentence are possible, and there is a mutual interdependence between the possible syntactic categories of the second, third, and fourth words. In the first analysis, *put* can be a verb in the past tense only if *can* is a noun and *single* is an adjective. In the second analysis, *put* can be the base form of a verb only if *can* is a modal and *single* is a noun. Another example is

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| . | the | man | was | a | prime | suspect | . |
| pct | det | n | v:aux | det | adj/n | n | pct |
| . | the | number | was | a | prime | . | |
| pct | det | n | v:aux | det | adj/n | pct | |

Both sentence beginnings allow *prime* to be either an adjective or a noun. For the second sentence, the fact that the sentence ends just after *prime* forbids the possibility of an adjectival interpretation of *prime*. For technical reasons, it is very cumbersome to deal with the beginnings and ends of sentences in a specific way. It is easier to include the full stop in the lexicon and consider the words at the beginnings and ends of sentences as having a context of full stop, either to the left (for the beginning of the sentence) or to the right (for the end of the sentence)—hence, the full stop at the beginning of the sentence in the example above. To categorize a word, it is not necessary to know all its contexts since the beginning of the sentence. Chaining contexts and hypotheses word after word obtains the same result. For example, the following chaining involves contexts of only one word. This is called a *binary matrix analysis* (the syntactic categories are displayed between brackets):

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| . | the | man | was | a | prime | suspect | . |
| (pct | det) | (n | v:aux) | (det | adj) | (n | pct) |
| | (det | n) | (v:aux | det) | (adj | n) | |

Chaining contexts of one-word length is not equivalent to chaining contexts of two-word length (called *ternary matrix analysis*), as can be shown by the example below:

One-word length context:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Sentence: | . | the | number | was | a | prime | . |
| Contexts: | (pct | det) | (n | v:aux) | (det | adj) | |
| | | (det | n) | (v:aux | det) | (adj | pct) |
| | (pct | det) | (n | v:aux) | (det | n) | |
| | | (det | n) | (v:aux | det) | (n | pct) |
| Result: | pct | det | n | v:aux | det | adj/n | pct |

Two-word length context:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Sentence: | . | the | number | was | a | prime | . |
| Contexts: | (pct | det | n) | (v:aux | det | n) | |
| | | (det | n | v:aux) | (det | n | pct) |
| | | | (n | v:aux | det) | | |
| Results: | pct | det | n | v:aux | det | n | pct |

In the one-word-length context, two results are possible, due to the fact that an adjective can be the last word of a sentence (e.g., *This is heavy*). A one-word context does not reveal the fact that the last word of a sentence cannot be an adjective if the penultimate word is a determiner. Thus, the chaining of the analysis does not reject the category Adjective for *prime* in the one-word-length context. However, this was rejected by the two-word-length context analysis, because a determiner cannot be followed by an adjective that is followed by the end of the sentence.

The bigger the context, the better the determination of the syntactic category. Unfortunately, there is a tradeoff between the size of the context and the complexity of the mechanism necessary to handle it. Let us suppose that there are 30 possible categories for a word. For a one-word context, this means that there are 30 × 30 (900) possible configurations to be memorized. Some are grammatical, some are not; some are more probable than others. For a two-word context, there are 30 times more configurations to be memorized (27,000). This becomes difficult to handle manually but is perfectly manageable if automatic training takes place using either pretagged corpora or raw corpora. For a three-word context, there are 810,000 configurations to be stored. This number is obviously impossible to handle manually, and it is also difficult to handle automatically because corpora of the required size are not readily available. It is considered reasonable to use a corpus of 100,000 to 1 million words to train models using a two-word context. For a three-word context, these numbers would be greater by two or three orders of magnitude. Because such quantities of training data are not available for spoken language, reliable training of three-word contexts is not yet feasible.

## Training Corpus for Part-of-Speech Taggers

As demonstrated above, although it is technically possible to describe all syntactic contexts by hand, it is too long a task to be performed easily. So procedures that can train part-of-speech taggers automatically have been implemented for a long time. These procedures can be divided into two main categories: supervised and unsupervised.

The most common principle for supervision is to rely on pretagged corpora (i.e., the same type of data that a part-of-speech tagger produces after training). If no pretagged corpus is available, it is necessary to build one from scratch. To avoid having to manually tag several dozens of thousands of words, a semiautomatic process can be implemented. A short corpus is first tagged manually. This corpus is then used for training the part-of-speech tagger. Then, a bigger corpus is tagged automatically using the short training data. This bigger corpus is corrected manually and used for retraining the tagger. Such an iterative process will lead to bigger tagged corpora until the size necessary for correct training of the tagger is reached. This procedure is very useful for training the tagger on a new language and for dealing with a new type of linguistic data (e.g., children's data), because few pretagged corpora are available for children's language. The sizes of the training corpora vary. The minimum size in recent implementations is 50,000, but older part-of-speech taggers used to work correctly with only 5,000 words. The variability of the training corpus is an important factor in reducing the amount of training necessary. Theoretically, the longer the training set, the better the results. In practice, there is a plateau in the improvement of part-of-speech tagging. When new training material is added, the quality of part-of-speech tagging gets better, however, after a certain amount of training, it reaches a maximum and later decreases. This is due to noise and small errors in the training material (Adda, 1987). The best amount of training should be chosen on a case-by-case basis, because it depends on the type of algorithm used and the quality of the training corpus. This also explains why the theoretically most powerful methods do not always obtain the best results.

Unsupervised tagging can be used by only certain specific part-of-speech tagger implementations. In some types of taggers (e.g., the *hidden Markov model*, HMM, implementations; see below), it is not necessary to have a pretagged corpus as input to the training process. Only a word lexicon with, for each entry, the list of possible syntactic categories is needed. Unfortunately, this principle requires very fine tuning of the initial weights in the model, and using as much supervised data as possible improves the results (Merialdo, 1994). It is also possible to devise mechanisms to automatically cluster groups of words into similarly behaving words and use these clusters to train the part-of-speech tagger (Schütze, 1995, 1997). This method is reserved for very large corpora and does not yet obtain very good results, although it may develop greatly in the future.

## Implementation of Part-of-Speech Taggers

Part-of-speech taggers are usually categorized as rule-based versus stochastic and as supervised versus unsupervised. However, a classification based on the actual algorithms used in the implementations is also helpful, since some rule-based algorithms can also be stochastic and unsupervised training can be combined with supervised training. The major algorithms used are the following: rule-based, nonstochastic, manual training (Chanod & Tapanainen, 1995); rule-based, nonstochastic, supervised, automatic computation of rules (Brill, 1995); precedence matrixes, bigrams or trigrams, standard Markov model, stochastic, supervised, automatic computation of matrixes (Church, 1988; Fluhr, 1977); rule-based, stochastic, binary, or ternary rules (Andreewsky, Debili, & Fluhr, 1980; Andreewsky & Fluhr, 1973); HMM, stochastic, supervised (and unsupervised), lexicon known in advance (Chanod & Tapanainen, 1995; Cutting, Kupiec, Pedersen, & Sibun, 1992; Merialdo, 1994); fully unsupervised (Schütze, 1995); and neural-network, supervised (Nakamura, Maruyama, Kawabata, & Shikano, 1990; Schmid, 1994).

For a detailed mathematical description of part-of-speech taggers using Markovian models, refer to Charniak, Hendrickson, Jacobson, and Perkowitz (1993) and Charniak (1997).

**Table 1**
**Binary Rules Retrieved for the Utterance *I play with her book***

| | | | | |
|---|---|---|---|---|
| <I * play> | ⇨ | <pro> * <v, n> | ➔ | pro * n, pro * v |
| <play * with> | ⇨ | <v, n> * <prep, adv> | ➔ | v * prep, n * prep, v * adv, n * adv |
| <with * her> | ⇨ | <prep, adv> * <pro, det:poss> | ➔ | prep * det:poss, prep * pro |
| <her * book> | ⇨ | <pro, det:poss> * <v, n> | ➔ | det:poss * n |
| <book * .> | ⇨ | <v, n> * <.> | ➔ | n * . |

Note—The symbol ⇨ means access through the lexicon, and the symbol ➔ means access through the dictionary of binary rules.

## SPECIFIC IMPLEMENTATION OF POST

The morphosyntactic analyzer used for POST has been developed to deal with such languages as French, English, German, Dutch, Italian, Spanish, and Greek, which use various forms of positional word order patterns. POST derives from much older work (Andreewsky et al., 1980; Andreewsky & Fluhr, 1973; Fluhr, 1977) and is grounded in a Markov model of binary rules. Its goal is to resolve ambiguities. Binary rules have the following format:

$$<C^1_1, C^1_2, . . , C^1_m> * <C^2_1, C^2_2, . . . , C^2_n>$$
$$\rightarrow (R^1_1 * R^2_2), . . . , (R^1_p * R^2_p).$$

$C^x_y$ and $R^z_t$ are syntactic categories (for any $x$, $y$, $z$, $t$), and "*" should be read as "followed by."

The two elements of the left-hand part of a rule correspond to the list of all syntactic categories for a word out of context. For example, given the words *one* and *book*, *one* can be either a numeral or an indefinite pronoun out of context, whereas *book* can be a noun or a verb. An example of the left-hand part is thus <num,pro:indef> * <n,v> (see Table 3 for the meaning of the symbols for syntactic categories). The elements of the right-hand part of a rule are two syntactic categories that resolve the ambiguity of the left-hand part of the rule. This means that, when a word that has the possible categories $<C^1_1, C^1_2, . . , C^1_m>$ is followed by a word that has the possible categories $<C^2_1, C^2_2, \ldots , C^2_n>$, then the first word will have the category $R^1$ and the second word the category $R^2$—that is, $(R^1 * R^2)$. Since more than one solution may exist, the right-hand part of a rule is made up of as many resolution pairs as necessary. With the previous example, when the pair of words *one book* occurs in context, two solutions are possible: *one* is a numeral and *book* is a noun, or *one* is a pronoun and *book* is a verb; so the right-hand part would be (num * n, pro:indef * v). An example in French is *la porte* (*the door*): article or pronoun followed by noun or conjugated verb. This binary rule resolves into two possibilities, so there are two pairs on the right-hand side of the rule: article followed by noun, and pronoun followed by verb, depending on the context. Learning these rules is easy with a pretagged corpus. First, all pairs of words are extracted from this corpus with their correct respective categories—that is, $(\{w^1, r^1\}, \{w^2, r^2\})$, where $w^x$ is the word's lexical form and $r^x$ is the category of the word in context. The category part of these pairs $(r^1, r^2)$ corresponds to the right-hand part of the rules to be generated. It is then possible to get the set of all possible categories for the words $w^1$ and $w^2$ from the lexicon. This produces the left-hand part of the binary rule, $w^1$ providing $<c^1_1, c^1_2, . . , c^1_m>$ and $w^2 <c^2_1, c^2_2, . . ., c^2_n>$. If one drops the information about the words, only the binary rule remains, $<c^1_1, c^1_2, . . , c^1_m> * <c^2_1, c^2_2, . . . , c^2_n> \rightarrow (r^1 * r^2)$. Lexical information is lost in this process; however, for a word that presents a specific ambiguity pattern—that is, its corresponding ambiguous class $<c_1, c_2, . . . , c_p>$ is unique—then some lexical information will be kept hidden in the dictionary of rules (Adda, 1987). This explains why binary rules have a greater theoretical potential than precedence matrixes.

The analysis is performed at utterance level. For each word of an utterance, the list of all the categories that the words can have out of context is generated. Then, proceeding pair by pair, all possible binary rules are extracted from the dictionary of rules. An example of the result for the five-word sentence *I play with her book* is given in Table 1.

Only the lists of resolution pairs are necessary to compute the part-of-speech analysis of the utterance. It is easier to understand the process by presenting these lists as in Table 2.

For each word that is neither the first nor the last word of the utterance, the intersection between the right-hand part of the resolution of the binary rule associated with the previous word (e.g., $r_1$) and the left-hand part of the resolution of the binary rule associated with the following word (e.g., $l_2$) is computed as shown in Table 2. The syntactic categories that belong to the intersection are kept and used for building the output of the tagger. Conventionally (this is not represented in Table 2), it is considered that each sentence begins with a period, which allows us to apply a binary rule to the first element of an utterance and to take into account the fact that a sentence does not usually begin with any type of syntactic category. There may be more than one solution. If this is the case, they are sorted on the basis of the frequencies of the binary rules that were used for constructing the solutions.

### Specificity of Training

The use of the program includes two main steps: training and analysis. These two steps are used in the iterative construction of a tagged corpus as described above. This procedure is efficient but has one pitfall. When the corpus becomes very large, the work needed to check it is very time consuming and conflicts with the initial goal of saving work and speeding up processing time. Even simple

Table 2
**Sketch of the Analyzing Process: Intersection of Resolved Pairs**

| I | * | play | play | * | with | with | * | her | her | * | book | book | * | . |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $l_1$ | | $r_1$ | $l_2$ | | $r_2$ | $l_3$ | | $r_3$ | $l_4$ | | $r_4$ | $l_5$ | $r_5$ | |
| pro | * | n | v | * | prep | prep | * | det:poss | det:poss | * | n | n | * | . |
| pro | * | v | n | * | prep | prep | * | pro | | | | | | |
| | | | n | * | adv | | | | | | | | | |
| | | | v | * | adv | | | | | | | | | |
| ⇩ | | ⇩ | | | ⇩ | | | ⇩ | | | ⇩ | | ⇩ | |
| $l_1$ | | Intersection between $r_1$ and $l_2$ | | | Intersection between $r_2$ and $l_3$ | | | Intersection between $r_3$ and $l_4$ | | | Intersection between $r_4$ and $l_5$ | | $r_5$ | |
| pro | | v, n | | | prep | | | det:poss | | | n | | | . |

Note—The words are presented only for easier understanding of the algorithm. Only categories are needed at this point in the analysis.

checking of a corpus can be time consuming. In order to avoid this shortcoming, it is necessary, at a certain point, to decide that the quality of the automatic tagging is sufficient and that an exhaustive control is no longer necessary. To facilitate the training phase, the analyzer may be used for pointing out the words on which it is likely to perform poorly. This allows the user to avoid an exhaustive examination of the whole corpus. Typically, difficulties arise in new syntactic situations (situations that appear in the analyzed text and did not appear in the training text). This will be the case when (1) a type of ambiguity has not been encountered before, (2) the string of resolutions cannot provide global disambiguation, or (3) there is more than one global resolution. Case 1 does not arise very often if the training corpus is large enough. However, when it does occur, it has to be examined by hand. Case 3 is frequent but does not necessarily require a full manual verification, because probabilistic solutions give good enough results to consider the average quality as satisfactory. Case 2 requires a more thorough examination, but systematic correction is often possible, once the error has been analyzed by hand. In this case, there is a clear lack of training, and these errors will be repeated. It is then possible to look at all identical contexts and automatically correct the error.

The use of training corpora makes it possible to create a grammatical system specific to a corpus, a human being, or a level of language. On the other hand, training specific to a given corpus is not always adequate for dealing with each new situation in new texts. In particular, children's corpora have many isolated words, which is not the case in the adult language written texts we used to create the initial morphosyntactic database. Isolated words can also be found in adult spoken language databases, and the problem of determining the lexical class of isolated words is the same for children and adults. Of course, it is not possible to build sophisticated context rules with utterances of only one word. Since the only context is punctuation (full stop, exclamation mark, or question mark) to the left and the right, no rule can solve the ambiguities. The cases of ambiguity between verb and noun were solved by hand, one by one, using the context of other sentences. One example is that of *un* (*a*/*one*). In French, *un* represents the number 1 and the indefinite article. We considered that, when used in isolation, it was the number. In fact, when checking every occurrence of *un* in isolation, we found that, in one case, it was actually an article used by an adult to suggest a word to a child. This circumstance is exemplary in two ways: It shows that automatic analysis cannot fully replace manual examination of data and that we must use the latter to study some very specific and localized situations; it also shows that there are always some "agrammatical" utterances that are justified by the pragmatics of the discourse, and no software will be able to deal with these in the near future.

This problem of isolated words is handled by POST using local word frequency information. If there is some ambiguity about an isolated word, then POST will choose the category that this word had most often in its previous occurrences. This allows POST to choose the categories of isolated words according to their local usage. Word frequency was not used in other syntactic situations because this would skew the analysis toward the repetition of the same patterns. The use of the syntactic context provides a more open set of solutions.

### Interface With CLAN and MOR

The tagging procedure and software presented here are available in the public domain. The software can already be used as a plug-in to the CLAN program (MacWhinney, 1995; http://childes.psy.cmu.edu) for tagging CHAT files. Within the CLAN programs, there are several methods for converting open text or files in other formats to CHAT. The CHILDES system already provides a program for morphological analysis, the MOR program. POST is not an alternative to MOR, but a complement. MOR provides a full decomposition of words into morphemes, which POST does not, because it uses only full-form lexicons. When one uses MOR, much manual work is still required to solve numerous ambiguities. POST provides a way to tackle this problem (see the Appendix).

### Evaluation for the French Language

**Lexical categories**. The choice and the size of the set of syntactic categories have some important consequences for the results of the tagger. If there are few cat-

**Table 3**
**List of 25 Morphosyntactic Categories Used by Children from Age 2 to Age 4**

| Tag for the Class | No. of Occurrences at 2 Years Old | No. of Occurrences for SLI Children | No. of Occurrences at 4 Years Old | Description of the Morphosyntactic Class |
|---|---|---|---|---|
| adj | 107 | 30 | 2,475 | Adjective |
| adv | 159 | 110 | 4,023 | Adverb |
| adv:neg | 130 | 119 | 3,255 | Adverb of negation |
| adv:place | 273 | 202 | 3,585 | Adverb of place |
| adv:voilà | 100 | 10 | 838 | Locution *voici, voilà* |
| co | 241 | 132 | 2,207 | Interjection |
| co:act | 253 | 65 | 1,590 | Interjection or exclamation |
| conj | 31 | 37 | 1,915 | Conjunction |
| det | 181 | 203 | 8,977 | Article |
| det:gen | 9 | 1 | 1,171 | Generalized article |
| n | 847 | 386 | 13,909 | Noun |
| n:prop | 21 | 93 | 509 | Last name, proper name |
| num | 6 | 17 | 503 | Number |
| prep | 15 | 69 | 3,657 | Preposition |
| prep:art | 42 | 48 | 1,599 | Preposition article |
| pro | 303 | 271 | 14,980 | Pronoun |
| pro:dem | 122 | 60 | 2,411 | Demonstrative pronoun |
| pro:rel | 72 | 34 | 2,148 | Relative or interrogative pronoun |
| pro:y | 38 | 27 | 1,009 | Pronouns *y, en* |
| v | 169 | 208 | 8,999 | Verb |
| v:aux\|avoir | 87 | 46 | 2,753 | Verb *avoir* (*to have*) |
| v:aux\|être | 246 | 129 | 4,745 | Verb *être* (*to be*) |
| v:inf | 115 | 115 | 4,558 | Infinitive |
| v:pp | 198 | 88 | 2,323 | Past participle |
| v:prog | – | – | 25 | Present participle |
| Total No. of Occurrences | 3,765 | 2,500 | 79,251 | |

egories, confusions are less likely to occur, and the tagger will give better results. However, the resultant structure provides less information than is needed for many types of linguistic analysis. On the other hand, a greater number of syntactic categories captures a greater amount of information, but confusions and ambiguities are more common. POST uses 25 general categories (see Table 3; punctuation at the end of sentences is excluded from this table). No tagging work has been done with regard to the problem of gender and number because these categories are not very prominent in the corpora of children between the ages of 2 and 3 years and do not require a very complex linguistic analysis. The lexical information provided by the MOR program is generally sufficient for these purposes. For example, only 1,707 nouns out of 38,164 have an undetermined gender. This means that only 4.4% of all nouns have to be checked manually. Furthermore, in a given transcript, these words will usually be used only with a specific meaning. This makes it possible to quickly go through every utterance of a word with ambiguous gender as soon as the first utterance has been identified. The choice of lexical tags reflects the distributional analysis of the French language. The difference between the various pronouns (personal, possessive, demonstrative, or relative) corresponds to the different contexts in which they can appear. If we wanted to use a greater number of categories, it would be necessary to verify that each newly created category could be differentiated using only its context.

**Material.** A systematic evaluation of the development of lexical categories in young children has been done using a new database (see Table 4) created with a technique of direct observation of behavior samples (Le Normand, 1986). It uses direct spontaneous speech data produced during symbolic play, in the same standard situation, video-recorded openly, always by the same observer. The recordings were made in this play situation to let the children comment on their own actions, talk about real or imaginary events, and converse with a familiar adult partner. The strictly standardized material consists of a toy house with five characters (two adult figurines, two child figurines, and one baby), one dog, 11 pieces of furniture (two tables, four chairs, two armchairs, and three beds), and five figurative objects (stairs with a mobile door, garage with a sliding door, and a front door bell).

For the data gathering, the technique of full sampling of behaviors was used. Child speech was segmented into utterances using the criteria defined by Rondal, Bachelet, and Pérée (1985), in accordance with the CHAT system (MacWhinney, 1995). The corpora presented here range from the age of 2 years to the age of 4 years. The children have a normal pattern of linguistic development. We have also added to this test a smaller database drawn from children with specific language impairment (SLI). We included 10 SLI children (8 boys and 2 girls, ranging in age from 4 years to 4 years 6 months), in accordance with the following criteria: (1) no hearing impairment or history of recurrent middle ear pathology, (2) no mental retarda-

**Table 4**
**List of the Characteristics of Corpora According to Age**

| Age (year.month) | Number of Children | Mean MLU | Minimum MLU | Maximum MLU | Number of Utterances | Mean Number of Utterances | Minimum Number of Utterances | Maximum Number of Utterances |
|---|---|---|---|---|---|---|---|---|
| 2.0 | 27 | 1.63 | 1.10 | 2.88 | 2,157 | 79.89 | 27 | 187 |
| 2.3 | 24 | 2.04 | 1.15 | 3.71 | 2,156 | 89.83 | 46 | 161 |
| 2.6 | 30 | 2.62 | 1.28 | 3.79 | 3,149 | 104.97 | 41 | 283 |
| 2.9 | 24 | 3.33 | 1.67 | 4.74 | 3,300 | 137.50 | 41 | 567 |
| 3.0 | 19 | 3.72 | 1.67 | 4.98 | 2,085 | 109.74 | 52 | 220 |
| 3.3 | 23 | 3.82 | 2.68 | 4.66 | 3,450 | 150.00 | 48 | 305 |
| 3.6 | 23 | 4.11 | 1.88 | 6.88 | 2,884 | 125.39 | 50 | 260 |
| 3.9 | 20 | 4.42 | 3.49 | 5.47 | 2,192 | 109.60 | 34 | 217 |
| 4.0 | 28 | 5.39 | 2.60 | 10.55 | 4,024 | 143.71 | 25 | 603 |
| SLI | 10 | 2.64 | 2.26 | 3.21 | 962 | 96.2 | 45 | 173 |

tion, (3) a severe expressive language disorder demonstrated by very low scores (more than 2 *SD* below the mean for the child's chronological age) in expressive subtests of a French language screening battery *Épreuves pour l'examen du langage 4–8 ans* (EEL; Chevrie-Muller & Decante, 1981), (4) no motor-speech problems, and (5) mean length of utterance (MLU) within the range of 2.26 to 3.21 (average MLU = 2.64). The MLU of these SLI children is about the same as that of normal children 2 years 6 months of age. The characteristics of the corpora are listed in Table 4.

## RESULTS

The first tagged corpus was the one for the 2-year-olds. The whole set of words, sorted by alphabetical order, was first tagged using a classical computerized dictionary, and all unknown words were tagged by hand (usually, these were interjections, exclamations, or distorted utterances). The first syntactic training of the tagger was carried out using a corpus we already had from previous work (Parisse, 1989). The 2-year-old children's corpus was then analyzed automatically and later corrected by hand. When the number of new syntactic occurrences became great enough, a supplementary training phase was carried out using the corrected part, and a new automatic analysis was performed. The same procedure was followed for each corpus, starting from the age of 2 years and moving up to the age of 4 years (with the younger children's corpora taken into account each time a new corpus was processed).

Even with corpora as small as in these studies (regarding the absolute number of lexical entries), there were still numerous ambiguities to be solved. Table 5 presents the number of ambiguities, first for the vocabulary of the children of the same age and then for the whole French lexicon. In the last line of Table 5, figures for English show that the number of ambiguities is not language specific.

The detailed analysis of every ambiguous case at the age of 2 years is easy to perform, because the number of ambiguous words is small, and a fine-grained control of the analysis is possible. The detailed list of ambiguities encountered at the age of 2 years is given in Table 6. This table also presents some examples of ambiguities encountered in the discourse of the SLI children. There are three types of ambiguities:

1. Classical (normal) ambiguities. Some are already functional at 2 years of age (e.g., *l'*); others are just beginning to appear (e.g., *la*).

2. Nonclassical ambiguities revealed by the principles of morphosyntax—for example, *au dodo* (*to bed*) versus *dodo*! (*bed!*). In the first case, *dodo* is a noun, whereas, in the second case, it is an interjection. Classically, these two cases correspond to the same category, but, because the distributional structures are different, it is better to tag the two occurrences of *dodo* differently.

3. Errors or contentious cases (cited in quotation marks in Table 6). It is not clear whether *place* in *le chien place* is an imperative verb, a verb in the present tense, or a noun. The analysis depends on what the child meant. It may be necessary to go back to the original recording and take into account the prosody and context of the utterance.

The errors are problematic because they could derive from incomplete or incorrect sentences. Fortunately, they are few in number, and the decision "made" by the analyzer permits identification of this type of problem.

Qualitative evaluation was carried out to check the quality of the automatic analysis. The differences between the automatically analyzed files and the later files corrected by hand were assessed. The overall results are given in Table 7. The four variables presented in each column provide various types of information. The first variable is the actual percentage of errors made by the analyzer. This goes from 3% up to 11% in the worst case, with a general average percentage of 5%, which is fairly accurate and similar to the results obtained by different taggers in the literature. The goal of this study was not to devise the best possible tagger but to construct the tagger most suitable for our purpose, which is to be able to easily check and correct the processed files. This was the main reason why we developed a new tagger and did not use those discussed in the literature. The POST tagger is able to signal when it encounters a difficulty (e.g., signaling unknown words or signaling where the analysis could not be completed). The latter situation occurs in two

**Table 5**
**Ambiguity Rate According to Age, Both Relative and Absolute**

| | | Relative Ambiguity (Relative to Children of the Same Group) | | Absolute Ambiguity (Relative to Adult) | |
|---|---|---|---|---|---|
| Age (year.month) | No. of Words | No. of Ambiguous Words | Rate of Ambiguity per Word | No. of Ambiguous Words | Rate of Ambiguity per Word |
| 2.0 | 3,765 | 534 | 1.15 | 2,433 | 2.30 |
| 2.3 | 4,542 | 725 | 1.17 | 3,144 | 2.44 |
| 2.6 | 8,506 | 1,572 | 1.19 | 5,897 | 2.41 |
| 2.9 | 11,510 | 2,417 | 1.21 | 7,979 | 2.43 |
| 3.0 | 7,871 | 1,493 | 1.20 | 5,335 | 2.36 |
| 3.3 | 13,180 | 2,389 | 1.18 | 8,859 | 2.37 |
| 3.6 | 12,153 | 2,424 | 1.20 | 8,049 | 2.35 |
| 3.9 | 9,742 | 2,184 | 1.23 | 6,490 | 2.33 |
| 4.0 | 22,899 | 5,612 | 1.27 | 14,947 | 2.38 |
| SLI | 2,501 | 376 | 1.16 | 1,707 | 2.40 |
| Adult | 132,982 | 63,707 | 1.79 | 87,691 | 2.52 |
| English | 417,711 | – | – | 313,194 | 2.56 |

Note—When not specified, these figures refer to the French language. The figures for English cover a mix of child and adult language.

different contexts: when there is a lack of rules and the analyzer finds no rule corresponding to the input words, or when the analyzer finds some rules but cannot chain a coherent analysis from the first to the last word of the utterance.

To improve the results of the analysis, the first and easiest task involves pretagging the unknown words of a new corpus. Unknown words are often text transcription errors that should be corrected before further analysis is conducted. In many other cases, unknown words may be interjections, which have a very free syntax that cannot be predicted by POST and thus generate a lot of tagging errors. The variable in the fourth line in Table 7 gives the percentage of errors that did not occur in a situation signaled as potentially erroneous by the analyzer. This is the most important variable. It means that the use of the analyzer makes it necessary to check only 8%–15% of the text to be tagged (those cases signaled as potentially erroneous; two thirds of these are correct, so they are quickly processed), and the final result will have less than 1% of errors (those not signaled by the analyzer as potentially erroneous), which may represent less than the number of errors a human would make when tagging a large corpus.

In Table 7, various results obtained from speech samples of both French-speaking normally developing 4-year-old children and French-speaking language-impaired

**Table 6**
**Examples of Ambiguities at the Age of 2.0 and for SLI children**

| | | | Example | | |
|---|---|---|---|---|---|
| Word | Ambiguous Classes | No. of Occurrences | 1 | 2 | 3 |
| | | | Normal 2-Year-Old Children | | |
| autre (other) | n, adj | 17, 3 | autre chaise | l'autre | |
| "balance" | n, v | 3, 1 | "balance le cheval" | "la balance" | ? |
| boum | n, i, adj | 2, 26, 2 | oh boum | un autre boum | c'est boum |
| bébé (baby) | n, np | 24, 80 | le bébé | oh bébé! | |
| dodo (sleep) | n, i | 25, 47 | au dodo | dodo! | |
| fait (does/do) | v, pp | 7, 4 | fait dodo | c'est fait | |
| l' (the [masc./fem.]) | prn, art | 55, 18 | où l'est | l'école | |
| la (the [fem.]) | prn, art | 1, 54 | ouvrir la porte | c'est la dame | |
| le (the [masc.]) | prn, art | 4, 65 | y a le chien | le voilà | |
| maman (mummy) | n, propn | 2, 35 | la maman | maman! | |
| "petit" (little) | n, adj | 2, 15 | "tout petit" | "les petits enfants" | |
| "place" | n, v | 2, 1 | "le chien place" | "place" | |
| qui (who) | prn, prn-r | 3, 3 | c'est qui | qui c'est | |
| tout (all) | prn, art-g | 3, 6 | c'est tout | tout ça | |
| un (a/one) | prn, art | 3, 21 | un lit | encore un | |
| | | | SLI Children | | |
| autre (other) | n, adj | 2, 2 | un autre | ah un autre fauteuil | |
| fait (do, done) | v, pp | 7, 3 | i fait noir au garage | i s'est fait mal | |
| un (one, a) | prn, art, nb | 3, 28, 1 | encore un ici | un fauteuil | un, deux, trois |
| l' (the [masc./fem.]) | art, prn | 14, 8 | oui l'a encore mangé | l'auto papa attend | |
| la (the [fem.]) | art, prn | 75, 3 | elle va la pousser la maman | | |
| dodo (sleep) | n, i | 11, 8 | un dodo | dodo! | |

Note—The first number under "No. of occurrences" refers to the first class name under "Ambiguous Classes"; the second number to the second class name; and so on. Examples cited in quotation marks are errors or contentious cases.

**Table 7**
**Rates of Errors**

| | French<br>Normal Developing Children | | | | French<br>Language-Impaired Children | | |
|---|---|---|---|---|---|---|---|
| | 2.0-a | 2.3–4.0 | 4.0-ae | 4.0-e | a | e | ae |
| Total % of errors | 6.73 | 3.27 | 4.75 | 4.71 | 11.1 | 3.01 | 5.07 |
| % of errors due to unknown words | 2.92 | 0.79 | 1.28 | 3.70 | 3.95 | 2.12 | 1.74 |
| % elements to be checked | 15.4 | 13.5 | 17.0 | 15.5 | 18.6 | 8.6 | 17.0 |
| % of nonsignaled errors | 1.07 | 0.45 | 0.74 | 0.23 | 3.32 | 0.20 | 0.83 |

Note—2.0 = 2 years 0 months of age. 2.3 = 2 years 3 months of age. 4.0 = 4 years 0 months of age. The following represent the corpora used during training: a, adult oral corpus; e, younger children's corpora; ae, combination of adult oral corpus and younger children's corpora.

children are presented. They correspond to different training sets. For 4-year-old children, the first training set consists of spoken adult language, and the second consists of child language (using the texts from younger children, from the ages of 2 years to 3 years 9 months). As the results show, although the global number of errors did not change (4.75% vs. 4.71%), the nature of the errors per se was not the same. With the adult training set, the percentage of unknown words was small, and the percentage of nonsignaled errors was 0.74%. With the child training set, the percentage of unknown words was higher,.and the percentage of nonsignaled errors was only 0.23%. This could mean that there is a difference between adult and child syntax that the analyzer is able to demonstrate. An alternative explanation may account for the differences outlined above: There are some residual errors in the training corpora, and the analyzer cannot cope very well with that problem. Indeed, the analyzer still requires improvement, not only in order to obtain better and more reliable results but also to identify problems in corpora that have already been tagged. Finally, we present the results obtained for the speech samples of language-impaired children. In this case, the nature of the training corpus is most relevant and interesting. Three types of training corpora were tested: spoken adult language only, child language only, and a mix of both. In every case, the percentage of unknown words did not change; this means simply that impaired children use some unknown words specific to their own situation. Still, there are half as many errors involving unknown words if we use only child training corpora. What is much more interesting is that the global error rate is three times smaller when one uses a child corpus for training the tagger than when one uses an adult corpus and that the rate of nonsignaled errors is 10 times lower (with 0.20%, it is our best result). This is very interesting because it not only shows that the use of an analyzer specific to a given age is necessary but also that the morphosyntax of some language-impaired children is very close to that of young children. It also demonstrates that it is possible to perform an automatic analysis on the speech of certain types of language-impaired children if the analyzer used is adapted to the speech that is to be analyzed.

## Evaluation for English

**Training corpora and syntactic categories.** POST was evaluated for English using the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 1999) from the CHILDES database. This corpus has been fully tagged for parts of speech; thus, it is a good candidate to train POST. The corpus is very large, with 627,645 utterances (1,979,221 words), and this makes it even more suitable for automatic training. In order to use POST, it is first necessary to substitute in the files of the Manchester corpus the tier name "%mor:" with "%trn:". This entails a simple global replacement of the string "%mor:" with the string "%trn:". Then, the MOR command of the CLAN software is used to add a new ambiguous "%mor" line to all of the files. The result is files where each utterance is accompanied by a "%mor:" tier in which none of the ambiguities have been resolved and a "%trn:" tier in which all the ambiguities have been resolved. Examples of English training data are given in the Appendix. There is no difference in format between the training data and the final result of POST. Only the names of the tiers in the CHAT files are changed to emphasize the function of these tiers.

This format is used as input to the training part of the POST software. Some preprocessing was necessary to unify some notation and to add all unknown words in the MOR lexicon. Out of the Manchester corpus, 78,011 utterances could not be used for training. In most of the cases (68,510), the problem is either a difference in format between the output of MOR and the tagged line of the Manchester corpus, a simple typographic error, or the presence of unintelligible words ("xxx" words in the CHAT format). In 9,501 cases, the difference was more serious, since the "%mor" and "%trn" tiers were tagged differently. Some differences were the result of errors, some corresponded to different choices of notation, and some displayed differences in syntactic interpretation (such as *done*, which is tagged as a communicator in the corpus and considered only as a verb form by MOR). It is a fact that both interpretations are possible in such utterances as *all done*, *well done*, or *done*.

**Results and comparison with other taggers.** Ninety percent of the checked Manchester corpus was used for training POST, and the remaining 10% was used to test the tagging (50,989 utterances, 149,255 words). Correct tagging occurs in 96.0% of the words. There is no unknown word in this test. The errors arising in this analysis were usually quite systematic and could be corrected globally. When checking the results, every time an error is

**Table 8**
**Characteristics of Various Part-of-Speech Taggers Available on the Internet**

| Tagger | URL | No. of Errors | % Correct |
|---|---|---|---|
| POST | | 9 | 95.2 |
| XEROX | www.rxrc.xerox.com/research/mltt/Tools/pos.html | 11 | 94.1 |
| CLAWS | www.comp.lancs.ac.uk/ucrel/claws/trial.html | 11 | 94.1 |
| MBT | ilk.kub.nl/~zavrel/tagtest.html | 19 | 89.8 |
| QTAG | tagger@clg.bham.ac.uk | 22 | 88.2 |
| LT POS | www.ltg.ed.ac.uk/software/posdemo.html | 13 | 93.0 |

Note—For QTAG, it is only necessary to send a text to the e-mail address above, and the tagged text will be sent back automatically.

encountered, it is important to check whether or not this error repeats itself in the remainder of the corpus. For example, *there* as a demonstrative pronoun can be confused with *there* as an existential pronoun. In our test, *there* was tagged erroneously as a demonstrative pronoun 276 times by the parser out of a total of 2,508 occurrences of *there*. A systematic search of all cases in which *there* as a demonstrative pronoun is followed by the verb *be* allows one to quickly correct these errors. Another example of a highly systematic error is the confusion between *did* as a verb and *did* as an auxiliary. This error occurred 263 times out of 717. Another example is *right* tagged as a communicator instead of an adverb, which occurred 155 times out of 1,297. A last example is *again* tagged as a particle instead of an adverb. Interestingly, *again* is never tagged as a particle in the Manchester corpus. This would mean that *again* should always be considered as an adverb, and it would be easy to change all occurrences of *again* as a particle into *again* as an adverb. However, there is also the possibility that always tagging *again* as an adverb was an error and that it should sometimes be tagged as a particle. This may lead one to revise the pretagged training corpus.

It is possible, in principle, to compare POST with other part-of-speech taggers, since many of these have demonstrations available on the Internet. These demonstration versions are unfortunately not designed to conduct large tests, so the test has to be done using a small corpus. To do this, we used 100 utterances (187 words) that were randomly extracted from a sample of speech uttered by a child 2 years 6 months of age. The small size of the sample decreases the significance of the test. Still, this sample appears to be representative since the percentage of correct tagging obtained using POST is very close to the percentage obtained with the big test corpus. Results for five different taggers including POST are presented in Table 8.

For the other taggers, it was impossible to determine which words were unknown. Therefore, we simply assumed that all interjections and compound words were potentially unknown, and we excluded errors involving these forms from the comparison.

A comparison with the percentage of correct tagging claimed by the various part-of-speech taggers shows that the results obtained with POST are comparable with the best results of the other taggers. POST obtained 96% accuracy with child language data, and this level is not obtained by all taggers. Few taggers claim a higher percentage, 97% or 98%. For example, the CLAWS tagger claims between 96% and 98% accuracy, depending on the corpus, and the Xerox tagger achieves 96% (Cutting et al., 1992). This comparison attests to the robustness of the binary rules algorithm implemented by POST. However, a more exact comparison will require that these various parsers be trained and tested on the same child language corpora, using the same treatment of unknown words.

## DISCUSSION

It may seem questionable whether the use of an automatic part-of-speech tagger is really necessary, especially since these tools do not provide a 100% quality analysis. To give a better idea of the real amount of work needed to tag a corpus using the morphosyntactic analyzer, we would like to present figures and comments about its use when tagging the corpus of French SLI children. There were 2,501 words to be analyzed (3,490 including punctuation). The time needed to run the software is very short, a few seconds on a computer with an Intel Pentium II 400-MHz microprocessor under Windows. The full corpora of normally developing children were used as a training corpus. After running the procedure, all three types of situations in which the software is expected to encounter problems (see the Specificity of Training section above) were examined. There were 8 examples of Case 1, in which an ambiguity had not been encountered before, 14 examples of Case 2, in which there was no global resolution, and 182 examples of Case 3, in which there was more than one global resolution. There were 60 other cases requiring verification, including those in which a word was unknown. We found 8 transcription errors and 77 syntactic errors during these verifications. Eight other errors were discovered during a subsequent, more thorough verification of the rest of the corpus. All this took approximately 2 h including a complete check. There are undoubtedly a few residual errors (0.20%; see the Evaluation section above), and an even more precise check of the whole corpus will require additional effort. But such a check will take far less time than the full manual tagging of the corpus. There were 1,707 words out of the 2,501 words that were ambiguous out of context. This meant 1,707 words had to be tagged assisted with a simple dictionary (about 2 days of hard work, with some remaining errors, so that a more thorough check such as the one suggested above is not out of order).

The use of automatic tagging does not make all intervention by the investigator unnecessary, but it speeds up the work in a nonnegligible way. Thus, going back to the example of normal children, the 7 weeks needed to process a corpus of 95,000 words can be reduced to 2 weeks or less, depending on the type of corpus (variable or even) and the needs of the user (general statistics or detailed analysis). If the user just wants a rough idea of the figures that can be drawn from a corpus, then it will only take a few minutes to analyze the text. And a surface checking of the results may take only a few days. For example, the verification done on the SLI children was more thorough (see above) because, on the one hand, we wanted a good-quality database and, on the other hand, we expected more errors because of these children's language deficit. In fact, this was not the case, probably because many "agrammatical" utterances were the same as those of very young normal children. The reduced database tagging time significantly changes the usual techniques of corpus study because it makes it possible to obtain, in a reasonable time, a global view of linguistic phenomena and provides a precise evaluation of the relative impact of occurrences considered to be exceptional or frequent.

## CONCLUSION

POST, a public-domain morphosyntactic analyzer that processes spoken language corpora such as dialogs between parents and their children, achieves around 96% correct tagging. This program can be found as a plug-in to the CLAN programs at the CHILDES Website (http://childes.psy.cmu.edu/). POST is intended for professional users. However, its use by the computer neophyte should not be problematic, especially in the fully automatic mode and as interfaced with CLAN. It also could prove a good addition to other software, such as SALT (Miller & Chapman, 1982), CLEAR (Baker-Van den Goorbergh & Baker, 1991), or CP (Long & Fey, 1995b).

Because POST is a part-of-speech tagger that uses local syntax, it is not required to organize information about the global structure of the sentence. However, given this aim, it offers increased speed of database processing, which is the main goal of numerous current computer applications. Part-of-speech tagging is usually considered the first step in any automatic processing of natural language. If this is true for natural language processing in general, it should be true for child language studies and for other studies of spoken language corpora. Thus, use of POST can open up new directions in developing tools for corpus analysis.

## REFERENCES

Adda, G. (1987). *Reconnaissance de grands vocabulaires: Une étude syntaxique et lexicale* [Recognition of large vocabularies: A syntactic and lexical study]. Unpublished doctoral dissertation, Université de Paris-Sud, Orsay.

Andreewsky, A., Debili, F., & Fluhr, C. (1980). Apprentissage—Syntaxe, sémantique lexicale [Training—syntax, lexical semantics]. *Revue du palais de la découverte*, **9**(83), 17-40.

Andreewsky, A., & Fluhr, C. (1973). *Apprentissage—Analyse automatique du langage, application à la documentation* [Training—automatic language analysis, application to data-retrieval] (Vol. 21). Paris: Dunod.

Baker-Van den Goorbergh, L. (1994). Computers and language analysis: Theory and practice. *Child Language Teaching & Therapy*, **10**, 329-348.

Baker-Van den Goorbergh, L., & Baker, K. (1991). *1991: Computerised language error analysis report* (CLEAR). Kibworth, U.K.: FAR Communications.

Bishop, D. V. M. (1984). Automated LARSP: Computer-assisted grammatical analysis. *British Journal of Disorders of Communication*, **19**, 78-87.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, **21**, 543-565.

Cappelli, G., Maccari, A., & Pfanner, L. (1991, May). *A system for semiautomatical treatment of child morphology*. Paper presented at the 4th Annual Sentence Processing Conference (CUNY, Rochester).

Chanod, J. P., & Tapanainen, P. (1995, March). *Tagging French—Comparing a statistical and a constraint-based method*. Paper presented at the Seventh Conference of the European Chapter of the Association for Computational Linguistics, Dublin.

Charniak, E. (1997). Statistical techniques for natural language parsing. *AI Magazine*, **18**, 33-44.

Charniak, E., Hendrickson, C., Jacobson, N., & Perkowitz, M. (1993). *Equations for part-of-speech tagging*. Paper presented at the Eleventh National Conference on Artificial Intelligence, Menlo Park.

Chevrie-Muller, C. S. A. M., & Decante, P. (1981). *Épreuves pour l'examen du langage* [Tools for language assessment]. Paris: Editions du Centre de Psychologie Appliquée.

Church, K. W. (1988, April). *A stochastic parts program and noun phrase parser for unrestricted text*. Paper presented at the Conference on Applied Natural Language Processing, Trento, Italy.

Crystal, D., Fletcher, P., & Garman, M. (1976). *The grammatical analysis of language disability*. London: Edouard Arnold.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992, April). *A practical part-of-speech tagger*. Paper presented at the conference on Applied Natural Language Processing, Trento, Italy.

Fluhr, C. (1977). *Algorithmes à apprentissage et traitement automatique des langues* [Learning algorithms and automatic language processing]. Unpublished thesis, Université de Paris-Sud Orsay, Orsay.

Le Normand, M. T. (1986). A developmental exploration of language used to accompany symbolic play in young, normal children (2–4 years old). *Child: Care, Health & Development*, **12**, 121-134.

Long, S. H., & Fey, M. E. (1995a). Clearing the air: A reply to Baker-Van den Goorbergh (1994). *Child Language Teaching & Therapy*, **11**, 185-192.

Long, S. H., & Fey, M. E. (1995b). Computer applications: Computerized profiling (1993). *Child Language Teaching & Therapy*, **11**, 209-216.

MacWhinney, B. (1991). *The CHILDES project—Computational tools for analyzing talk*. Hillsdale, NJ: Erlbaum.

MacWhinney, B. (1995). *The CHILDES project: Tools for analyzing talk* (2nd ed.). Hillsdale, NJ: Erlbaum.

MacWhinney, B., & Snow, C. E. (1985). The Child Language Data Exchange System. *Journal of Child Language*, **12**, 271-296.

Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics*, **20**, 155-172.

Miller, J. F., & Chapman, R. S. (1982). *SALT: Semantic Analysis of Language Transcripts*. Language Analysis Laboratory, Waisman Center on Mental Retardation and Human Development, University of Madison, Wisconsin.

MILLER, J. F., & CHAPMAN, R. S. (1983). Using microcomputers to advance research in language disorders. *Theory Into Practice*, **XXII**, 301-307.

NAKAMURA, M., MARUYAMA, K., KAWABATA, T., & SHIKANO, K. (1990, August). *Neural network approach to word category prediction for English texts.* Paper presented at the COLING—90, Helsinki.

PARISSE, C. (1989). *Reconnaissance de l'écriture manuscrite: Analyse de la forme globale des mots et utilisation de la morpho-syntaxe* [Machine recognition of handwriting: Global analyses of word shapes and morpho-syntactic evaluation]. Unpublished doctoral dissertation, Université de Paris-Sud, Orsay.

PERKINS, M. (1994). Repetitiveness in language disorders: A new analytical procedure. *Clinical Linguistics & Phonetics*, **8**, 321-336.

PERKINS, M., CATIZONE, R., PEERS, I., & WILKS, Y. (1997, June). *Clinical computational corpus linguistics: A case study.* Paper presented at the 6th Annual Conference of the ICPLA, Nijmegen.

RONDAL, J. A., BACHELET, J. F., & PÉRÉE, F. (1985). Analyse du langage et des interactions verbales adulte-enfant [Analysis of adult-child language and verbal interactions]. *Bulletin d'Audiophonologie*, **5**, 507-536.

SCHMID, H. (1994, August). *Part-of-speech tagging with neural networks.* Paper presented at the COLING—94, Kyoto.

SCHÜTZE, H. (1995, March). *Distributional part-of-speech tagging.* Paper presented at the Seventh Conference of the European Chapter of the Association for Computational Linguistics, Dublin.

SCHÜTZE, M. (1997). *Ambiguity resolution in language learning.* Stanford, CA.

THEAKSTON, A. L., LIEVEN, E. V. M., PINE, J. M., & ROWLAND, C. F. (1999). The role of performance limitations in the acquisition of "mixed" verb-argument structure at stage 1. In M. Perkins & S. Howard (Eds.), *New directions in language development and disorders.* New York: Plenum.

## APPENDIX
## Examples of English Training Data

Use of the Analyzer Inside the CLAN System

1) Find the name of the file containing the training database to be used: french.db
2) Find the name of file (in CHAT format) to be analyzed: example.cha
3) Start the CLAN software and go to command mode (do not forget to set up the default MOR files option to 'fralex' for the lexicon name and 'fra' for the grammar name).
4) Run the MOR command on 'example.cha': the result is the file 'example.mor.cex'
5) Run the POST command on the resulting file 'example.mor.cex': the result is the file 'example.mor.pst'
6) Edit the resulting file 'example.mor.pst' with an editor or the CED software. The result is presented in a special tier called '%pos' that may complement the '%mor' tier of the CHILDES system.

Sample of Results of POST in French

```
*PHI: non . (no)
%mor:   adv:neg|non^n|non&MASC .
%pos:   adv:neg|non .

*PHI:   les cheminées (the chimneys)
%mor:   det|les^pro:obj|les&PL^pro|les v:pp|cheminées^n|cheminée&FEM-PL
%pos:   det|les n|cheminée&FEM&PL

*PHI:   y en a une qui est tombée (there is one of them which is down)
%mor:   adv|y^pro:y|y^n|y&MASC
adv|en^pro:y|en^prep:art|en^prep|en^pro|en
        v:aux|avoir&PRES&3S^n|a&MASC num|une^num|une^det|une^pro|une
        pro:rel|qui^pro:int|qui^pro|qui^conj|qui
        v:aux|être&PRES&3S^n|est&MASC n|tombée&FEM^v:pp|tomber&FEM&SING
        %pos:   pro:y|y pro:y|en v:aux|avoir&PRES&3S det|une pro:rel|qui v:aux|être&PRES&3S
        v:pp|tomber&FEM&SING

*PHI: xxx je ferai une autre maison . (xxx I will make another house)
%mor:   pro:subj|je&1S^pro|je v|faire&FUT&1S
        num|une^num|une^det|une^pro|une
        det:gen|autre^pro|autre^n|autre^adj|autre n|maison&FEM .
%pos:   pro:subj|je&1S v|faire&FUT&1S det|une adj|autre n|maison&FEM .
```

**APPENDIX (Continued)**

```
*PHI:   ça s' appelle des jetons . (it is called chips)
%mor:   pro:dem|ça pro:refl|se&3^pro|s'^conj|s'
        v|appeler-SUBJV:PRES&3S^v|appeler-SUBJV:PRES&1S^v|appeler-IMP&2S^v|appeler
        PRES&3S^v|appeler-PRES&1S
        det|des^prep:art|des^prep|des
        n|jeton&MASC-PL^v|jeter-IMP&1P^v|jeter-PRES&1P .
        %pos:  pro:dem|ça pro:refl|se&3
        v|appeler&SUBJV:PRES&3S^v|appeler&SUBJV:PRES&1S^v|appeler&IMP&2S^v|appeler&PRES&3
        S^v|appeler&PRES&1S
        det|des n|jeton&MASC&PL .

*PHI:   là c' est quoi ? (this, what's that?)
%mor:   adv:place|là pro|ce/ces v:aux|être&PRES&3S^n|est&MASC
        pro:int|quoi^pro|quoi^conj|quoi^pro:rel|quoi ?
%pos:   adv:place|là pro|ce/ces v:aux|être&PRES&3S pro:rel|quoi ?

*PHI:   si on prend ça . (if one takes that)
%mor:   adv|si^conj|si^pro:rel|si^n|si&MASC pro:subj|on&IMP^pro|on
        v|prendre&PRES&3S pro:dem|ça .
%pos:   pro:rel|si pro:subj|on&IMP v|prendre&PRES&3S pro:dem|ça .

*PHI: de la confiture . (some jam)
%mor:   adv|de^det|de^prep:art|de^prep|de
        det|la^pro:obj|la&SING&FEM^pro|la^n|la&MASC n|confiture&FEM .
%pos:   prep|de det|la n|confiture&FEM .

*PHI:   ouais . (yeah)
%mor:   co|ouais .
%pos:   co|ouais .

*PHI:   voilà là, c' est des miettes ? (here it is, that some crumbs)
%mor:   adv:voici|voilà^prep:voici|voilà adv:place|là pro|ce/ces
        v:aux|être&PRES&3S^n|est&MASC det|des^prep:art|des^prep|des
        n|miette&FEM-PL ?
%pos:   adv:voici|voilà adv:place|là pro|ce/ces v:aux|être&PRES&3S
det|des n|miette&FEM&PL ?

*PHI:   monté dans le camion le gros . (got up in the truck the big one)
%mor:   v:pp|monter&MASC&SING prep:art|dans^prep|dans
        det|le^pro:obj|le&SING&MASC^pro|le n|camion&MASC
        det|le^pro:obj|le&SING&MASC^pro|le
        adj|gros&MASC&SINGPL^n|gros&MASC&SINGPL^adv|gros .
        %pos:  v:pp|monter&MASC&SING prep|dans det|le n|camion&MASC det|le
        adj|gros&MASC&SINGPL .
```

Sample of Training Data for POST in English

```
*CHI:   there's that one on there , look .
%mor:   pro:dem|there^n:prop|there^pro:exist|there^pro:dem|there
        n-cl|v:aux|be&3S^n-cl|v|be&3S^n-cl|POSS
        wh:pro|that^wh:adv|that^pro|that^pro:indef|that^adv|that^wh:rel|
that^wh:rel|that^det|that^conj:subor|that^pro:dem|that
        pro|one^pro:dem|one^num|one^det|one^pro:indef|one^num|one    adv|on^ptl|on^prep|on
        pro:dem|there^n:prop|there^pro:exist|there^pro:dem|there    v|look^n|look^co|look .
```

**APPENDIX (Continued)**

%trn: pro:dem|there n-cl|v|be&3S det|that pro:indef|one prep|on
pro:dem|there v|look .

*CHI: just leave you on there a minute .
%mor: adv|just^adj|just v|leave^n|leave pro|you^co|you^pro|you adv|on^ptl|on^prep|on
pro:dem|there^n:prop|there^pro:exist|there^pro:dem|there
det|a^det|a n|minute^adj|minute .
%trn: adv|just v|leave pro|you prep|on pro:dem|there det|a n|minute .

Sample of Results of POST in English

*CHI: oh .
%mor: co|oh .
%pos: co|oh .

*CHI: I have them this morning .
%mor: pro|I v|have^v:aux|have pro:dem|them^pro|them pro:indef|this^det|this^pro:dem|this
n|morning^co|morning .
%pos: pro|I v|have pro|them det|this n|morning .

*MOT: did you ?
%mor: v|do&PAST^v:aux|do&PAST pro|you^co|you ?
%pos: v:aux|do&PAST pro|you ?

*MOT: oh right .
%mor: co|oh pro:dem|right^co|right^adv|right^adj|right .
%pos: co|oh co|right .

*CHI: he's in back .
%mor: pro|he n-cl|v:aux|be&3S^n-cl|v|be&3S^n-cl|POSS co|in^adv|in^ptl|in^prep|in^n|in
prep|back^ptl|back^v|back^n|back^adv|back^adj|back .
%pos: pro|he n-cl|v|be&3S prep|in **adv**|back .

*CHI: Mummy .
%mor: n:prop|Mummy^n|Mummy .
%pos: n|Mummy .

*CHI: why did he do that ?
%mor: wh:rel|why^wh:adv|why v|do&PAST^v:aux|do&PAST pro|he n|do^v|do^v:aux|do
wh:pro|that^wh:adv|that^pro|that^pro:indef|that^adv|that^wh:rel
|that^wh:rel|that^det|that^conj:subor|that^pro:dem|that ?
%pos: wh:adv|why v:aux|do&PAST pro|he v|do pro:dem|that ?

*CHI: just Sukie like-es them .
%mor: adv|just^adj|just n:prop|Sukie v|like-3S pro:dem|them^pro|them .
%pos: **adj**|just n:prop|Sukie v|like&3S pro|them .

*CHI: Sukie just bite me as+well .
%mor: n:prop|Sukie adv|just^adj|just v|bite^n|bite co|me^pro|me n|as+well^co|as+well^adv|as+well^n
comp|as+well .
%pos: n:prop|Sukie adv|just v|bite pro|me co|as+well .

Note—Errors are presented in bold face.

(Manuscript received January 2, 1999;
revision accepted for publication February 9, 2000.)