Automatic disambiguation of morphosyntax in spoken language corpora

Christophe Parisse, Marie-Thérèse Le Normand

Fully automatic disambiguation of the morphosyntax in young children: the POST software

Christophe PARISSE and Marie-Thérèse LE NORMAND

Institut National de la Santé et de la Recherche Médicale (INSERM), Paris, France

Abstract

The use of computer tools does not always speed up the analysis of young children's transcripts. Although it is now easy to lexically tag every word in a corpus, you still have to choose between numerous ambiguous forms, especially with languages such as French or English, where nearly 50% of the words are ambiguous. Computational linguistics now offer well-developed part of speech labeling which permits fully automatical disambiguation of lexical tags: the tool presented here (POST) can tag and disambiguate a large text in a few seconds. This could form a complement to many systems dealing with language transcript, and also suggests further theoretical developments about the assessment of the status of morphosyntax in child language. The program works for French but is open to other languages such as English. The analyses and computation of a corpus produced by normal children aged two to four, as well as of a sample corpus produced by SLI children are given as examples.

Introduction

Automatic analysis of transcript is not always as simple as it should be. Some of the tasks it involves are quite fastidious although computer tools are already of great help. One of these tasks is the disambiguation of lexically tagged texts. In a language like French or English, half the words are ambiguous, given a full adult lexicon (more than a hundred thousand words). A reduced lexicon size reduces the number of ambiguous words, but also leads to more errors. It is very difficult when creating a lexicon to decide in advance that a word is not going to be ambiguous in the analyzed corpus. It would be better to use a full child and adult lexicon and to choose within the whole set of lexical possibilities. This task can be very time consuming when analyzing a large transcript. Fortunately, it is possible to render this fully automatic using an advanced part-of-speech program than can fully tag and disambiguate a corpus in a few seconds, with a error rate that may be inferior to or about same as human processing error rates. Also, the adequacy of such automatic processing shows that the morphosyntax of child language is very consistent, in itself and with regard to that of the adult.

Morphosyntactic analysis consists of looking for the syntactic tag and the morphological decomposition of a word. The tags used in that kind of analysis match those one could find in a dictionary, that is, the word class without any semantic or pragmatic context. In this situation, a word can have an ambiguous tag: it could be a homophone or a homograph and then more than one lexical input share the same phonemic or graphemic shape. The analysis has to rely on context in order to determine which word class it really belongs to. For instance, in

French, it is necessary to determine if the string of letters ‘p o r t e’ corresponds to the feminine substantive ‘porte (door)’ or to the conjugated verb ‘porter (to open)’ in the present tense, either in the first person or in the third person. This style of morphosyntactic analysis is well suited to the study of young children language because it takes into account three significant linguistic components: morphology, syntax and distributional analysis.

Concerning the reality of the ambiguous form problem, it is interesting to give some specific figures about the French language. A two 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. If the analyzed corpus has been transcribed from oral format to written format, there is no more oral ambiguity to deal with (between 'la' and 'là' who share the same pronunciation). Nevertheless, the graphemic ambiguity still needs to be addressed. With large corpora, the number of ambiguous cases may be very large. In the 95,000 occurrences in the child database described below, 2,900 occurrences of 'la' have to be disambiguated between article and pronoun (in the case of written text), 4,600 occurrences between article, pronoun and adverb (in the case of spoken text). For this corpus of 95,000 words, a manual tagging which takes 10 seconds per word will take 7 weeks. It may take over a year to complete this task. A quasi-automatic morphosyntactic analysis would facilitate research progress and should offer new prospects for corpus study. In this article, we propose the use of an automatic parser based on a Markov model to aid linguistic analysis and give examples of the use of such software.

State of the art

Many software tools currently exist which are intended to help the researcher, linguist, psychologist or psycholinguist in their study of linguistic corpora. These tools are general and may be complemented by other tools more specific to linguistic processing such as tools to cut a text into words, to split words into morphemes, to perform a grammatical analysis of a full sentence, either a relatively simple morphosyntactic one, or some more complex decomposition into groups or trees (e.g., Miller and Chapman, 1983; MacWhinney and Snow, 1985; Baker-Van den Goorbergh, 1994; and Long and Key, 1995a, 1995b). However few systems truly perform a thorough linguistic analysis. Examples are the morphological decomposition described by Miller and Chapman (1983) and Cappelli, Maccari and Pfanner (1991) in order to compute the Mean Length of Utterance, and the last implementation of the MOR program of the CHILDES system that now provide some automatic disambiguation tools. 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. Some attempts to provide a software aid to the production of this assessment have been done (e.g., Bishop, 1984) although actually the procedure was better used to train people in the system than to carry out the analysis per se.

Morphosyntactic analysis can be carried out on natural speech samples. Automatic processing has expanded

over the past 10 years because on the one hand such an analysis is often used as a starting point for more complex linguistic processes, and on the other hand, the techniques used to build such automatic systems are now well known. This kind of part-of-speech parser may be found freely on the Internet, e.g., the Xerox Part of Speech Tagger (ftp://parcftp.xerox.com/pub/tagger/) or the Brill Transformation Based Tagger (Brill, 1994; ftp://blaze.cs.jhu.edu/pub/brill/Programs/), or with a fee from specialized linguistic suppliers. There are different systems of lexical categories that can be covered by automatic morphosyntax processing. It varies with the type of application considered or the use it is planned for, and also with the morphosyntactic nature of the language processed. Thus, some parsers do not deal with gender or number agreement, either because the planned software does not need that information, either because gender and/or number information is not morphological in a language or does not modify the structure of the sentences (these two remarks apply to English, the second one applies most often to French). Subject and object identification is not very often solved by morphosyntactic parsers unless the processed language includes case markers that directly solve the problem. In the languages that do not use case makers, word order is often strict and should, at first glance, make it possible to tackle the question. Unfortunately, outside some simple structures, the structural ambiguities compel one to resort to semantics and push subject and object identification outside of the morphosyntactic field for that language. This shows the limits of morphosyntax which works in a very limited context and does not use any semantic knowledge. Morphosyntax is part of a full linguistico-cognitive system and is closely related both to lexical processing and to syntactic or sometimes even semantic processing.

Methodology

The morphosyntactic analyzer we used has been created to deal with different European positional or semi-positional languages like French, English, German, Dutch, Italian, Spanish and Greek to help reduce the complexity of automatic recognition of script or speech (Parisse, 1989). The goal of POST (Part Of Speech Tagger) is to automatically tag texts following an automatic training period which is as short as possible. It is grounded in a Markov model of ambiguous bi-class rules succession resolution. When there is a choice between some classes following another choice between other classes – for example, ‘la porte’ (the door): 'article' or 'pronoun' followed by 'substantive' or 'conjugated verb'; ‘her book’: 'personal pronoun' or 'possessive adjective' followed by 'verb' or 'noun' – it is an ambiguous bi-class (this will also be the case if only one of the two parts is ambiguous). The ambiguity has to be resolved, that is one has to decide what the two real classes in the current context are – in the previous example, in ‘la porte’, only article followed by substantive and pronoun followed by verb are possible and the context will decide between the two possibilities; in ‘her book’, only possessive adjective followed by noun is possible. Using the succession of ambiguous bi-classes and their resolution, the analyzer provides some sentence-level resolution. There may be more than one solution: in that case, they are sorted on the basis of the probabilities of grammatical bi-class occurrences. The morphosyntactic analyzer is well suited to solve the lexical ambiguity problems which are

very numerous in French, as in many other languages.

The use of the program included two main steps: training and analysis. Data about French training is already available, as will be explained below. Once training is done, the analysis of a text takes a few seconds. It may be necessary to check the result of the analysis, depending on what quality one is looking for. The analyzer may be tuned to point out the words where it is likely to poorly perform. An evaluation of the quality of the analyzer will be given later in the text.

Before any analysis, it is necessary to have a training corpus. Thus, one must tag by hand some hundreds of words which are going to form the starting point of a cyclical process. These words represent the first training corpus. With it, one can learn morphosyntactic rules and later analyze a larger corpus that will be manually checked and later will form the new training corpus. Step by step, it enables one to control and deal with larger and larger corpora. This procedure has one pitfall. When the corpus becomes very large, the work needed to check it is very time-consuming and tends to be the opposite of the initial goal of saving work and speeding up processing time. Even simple 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. The analyzer will still have difficulties which arise in new syntactic situations (situations that appear in the analyzed text and did not appear before in the training text). This will be the case when: 1) an ambiguous bi-class has not been encountered before, 2) the string of resolutions cannot provide a global resolution (an analysis of the full sentence without any break), 3) there is more than one global resolution. Case 1) does not happen very often if the training corpus is large enough, but has to be checked by hand. Case 3) is frequent but does not necessarily imply a full manual verification because probabilistic solutions give good enough results to consider the average quality as satisfactory. Case 2) requires a more thorough control, 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 repeat themselves. It is only necessary to look at all identical contexts and a global correction can be carried out.

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 enough to face every 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 started with as the initial morphosyntactic database. These isolated words can also be found in adult spoken language databases and the question of lexical class determination of isolated words is the same for children and adults. Of course it is not possible to build sophisticated context rules with sentences of only one word. The only context being punctuation (full stop, exclamation or question marks) at the left and at the right, no rule can solve the ambiguities. The cases of ambiguity between verb and noun have been solved one by one by hand using the context of other sentences. One example is that of 'un' (a/one). In French, 'un' stands for the number 1 and for the indefinite article. We considered that, when used in an isolated way, it was the number. In fact,

when checking every occurrence of 'un' when it is an isolated word, we can see that in one case, it is in fact the article that is 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 a manual examination of data and that we have to use the latter to study some very specific and localized situations; it shows that there are always some 'agrammatical' utterances which are justified by the pragmatics of the discourse and that no software will be able to deal with these in the near future.

Lexical categories

The choice of lexical tags generated by the software is that of very general categories (see Table 1 – punctuation at the end of sentence are excluded from this table). No tagging work has been done with regard to the problem of gender and number because these categories are not too prominent in corpora of children aged between 2 and 3 and do not require a very complex linguistic analysis. In general, this choice of lexical tags reflects the distributional analysis of the French language. Thus, the difference between the various pronouns, personal, possessive, demonstrative or relative corresponds to the different contexts where they can appear. If we wanted to use a greater number of categories, it would be necessary to verify that the newly created classes could be differentiated using only their context.

insert Table 1 about here (list of 25 morphosyntactic categories used for the child from 2 to 4)

Material

A systematic evaluation of the development of lexical categories in the young child has been done using a database created with a technique of direct observation of behavior sample (Le Normand, 1986). This uses direct spontaneous speech data produced during symbolic play, in the same standard situation, video-recorded in an open way to the child and always by the same observer. The recordings were done in this play situation to let the child comment on his/her own actions, to tell about real or imaginary events, and to have some exchange with a familiar adult partner. The strictly standardized material consist of a toy house with five characters (two adult figurines, two child figurines and one baby), one dog, eleven 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 were used and the child speech was segmented into utterances using the criteria defined by Rondal, Bachelet and Pérée (1985), which allow a standard transcription and the computation of linguistic parameters as described in the corpus processing system CLAN (Child Language Analysis, version 2.01, MacWhinney and Snow, 1985, 1995). The corpora presented here range from the age of 2.0 to the age of 4.0. These children have a normal linguistic development pattern. We have also added to our test a smaller database drawn from children with specific language impairments (SLI). Ten SLI children (8 boys and 2 girls, ranging in age from 4 years to 4 years 6 months) were selected in accordance with the following criteria: (1) no hearing impairment or history of recurrent middle ear pathology, (2) no mental retardation, (3) a severe expressive

language disorder demonstrated by very low scores (more than 2SD below the mean for the child's chronological age) in expressive subtests of a French language screening battery ‘Épreuves pour l'examen du language 4-8 ans’ (EEL, Chevrie-Muller, Simon and Decante, 1981), (4) no motor-speech problems, (5) and 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 children of two years and a half. The characteristics of the corpora are given below in Table 2.

insert Table 2 about here (list of the characteristics of corpora according to age)

Results

The first tagged corpus was of two-year-old children. 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, it was interjections, exclamations, or children's distorted utterances). The first syntactic training was carried out using a corpus we already had from a previous work (Parisse, 1989). The corpus of the two-year-old children was then analyzed automatically and later corrected by hand. When the number of new syntactic occurrences became great enough (especially because of isolated words or small or incomplete exclamation sentences specific to the language of the young child), a complementary training phase was carried out using the corrected part and a new automatic analysis performed. The same procedure has been followed at each age for the children (with the previous younger children's corpora taken into account each time a new corpus is processed).

Even with corpora as small as these studies (regarding the absolute number of lexical entries), there were still numerous ambiguities to be solved. Table 3 presents the number of ambiguities, first compared with the vocabulary of the children of the same age (a word is ambiguous at a given age, if it appears in more than one class at that age), then compared with the whole French lexicon (a word is ambiguous at a given age, if it appears in more than one class in the French language, as an adult or a linguist would interpret the word without knowing the child’s precise vocabulary).

insert Table 3 about here (ambiguity rate according to age, either in a relative way, either in absolute)

The detailed analysis of every ambiguous case at the exact age of two is still quite easy to perform, because the number of ambiguous words is small and a fine-grain control of the analysis is still possible. The detailed list of ambiguities encountered at the age of two is given in Table 4. In this table are also some examples of ambiguities encountered in the discourse of the SLI children. In this list, there are three kinds of situations displayed: classical (normal) ambiguities – already functional as is the case for ‘l’’, just beginning to appear as is the case for ‘la’ –, non-classical ambiguities relative to the principle of morphosyntax – ‘c'est qui’ (it's who?) vs. ‘qui c'est’ (who is it?): here ‘qui’ does not have the same function; ‘au dodo’ (to bed) vs. ‘dodo!’ (bed!), where the distributional structures are different –, and errors or contentious cases, written between quotes in Table 4. These last cases are problematic because they could be incomplete or incorrect sentences. They are few in number, and the decision “taken” by the

analyzer permits identification of this kind of problem.

insert Table 4 about here (examples of ambiguities at the age of 2 years 0 month)

Evaluation

Qualitative evaluation was carried out to check the real quality of the automatic analysis of POST. In order to do this, the differences between automatically analyzed files and the later files corrected by hand were assessed. The overall results are given in Table 5. 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 other results obtained by different taggers in the literature. The goal of this study was not to devise the best tagger, but the 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 an original tagger and did not use those mentioned in the literature. To assist us in our verification of the results, the analyzer can provide clues. This is relevant for the following reason: when one has to tag some thousands of words, if it is possible to know in advance what the elements to be checked exactly are, one can drastically prune the resulting work. The first and easiest task which is to be carried out consists in pre-tagging the unknown words of a new corpus. There are three reasons do to this. First, the analyzer still performs poorly on unknown words (no more than 30% of good labeling – the second variable in the columns of Table 5 gives the percentage of errors due to unknown words in a text.) Secondly, the ratio of type-token being on average 2, there is half the work to be done if one tags the unknown words before the analysis (one tags the types and not the tokens). Thirdly, unknown words are often errors of text transcription and it is advisable to carry out a first verification. The third variable in the columns of Table 5 gives the percentage of values which are marked as ambiguous by the analyzer. They correspond to the above mentioned three cases described in the Methodology section. The last variable in the columns give the percentage of errors that are not signaled by the analyzer. This is the most important variable. It means that the use of the analyzer makes it possible to check only 8 to 15% of the text to be tagged (and two thirds of these are correct, so they are quickly processed) and the final result will have less than one percent of error, which may be less than the number of human errors when tagging a large corpus.

insert Table 5 about here ‘rates of errors (total number of errors, errors induced by unknown words, elements to be verified by hand and non-signaled errors)’

In Table 5, various results obtained from speech samples of both French-speaking Normally Developing (FND) children at age 4 years 0 month and for French Language Impaired (FLI) children are presented. These correspond to different training sets. For children at age 4, the first training set consists of spoken adult language and the second consists of child language (using the texts coming from younger children, aged from 2 years to 3 years and 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 while the percentage of non-signaled errors were 0.74%. With the children training set, the percentage of unknown words was higher while the percentage of non-signaled errors was only 0.23%. This could mean that there is a difference in the syntax between adult and children that the analyzer is able to demonstrate. An alternative explanation may account for the differences outlined above: namely, that 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 which 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 kinds 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. But still, there are half as many errors in unknown words if we use only child training corpora. What is much more interesting is that the global error rate is a third in the case of child only training corpora, and the rate of non-signaled errors is ten times less (with 0.20%, it is our best result). This is very interesting for the following reason: it does 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 use an automatic analysis on certain type of language impaired children if the analyzer used is adapted to the speech which is to be analyzed.

Conclusion

POST, a morphosyntactic analyzer that process the language of the children is completely operational (see Figure 1 for an example of use in analysis mode). Its use may as an option require user supervision because the choices of the analyzing software are not always the good ones. This control can be skipped or limited to a small percentage of the text with a remaining small error rate. In addition, errors are usually systematic, in contrast to the human errors, and may be globally corrected. 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 SLI children. There were 2,501 words to be analyzed (3,490 with the punctuation). The time needed to run the software is very short, some few minutes. The full corpora of normally developing children has been used as a training text. After running the procedure, all three points (see the Methodology section above) where the software is supposed to meet difficulties have been verified: case 1) ‘an ambiguous bi-class has not been encountered before’ occurred 8 times; case 2) ‘the string of resolutions cannot provide an global resolution’ occurred 14 times, case 3) ‘there is more than one global resolution’ occurred 182 times. There were 60 other cases requiring verification, those where the word was unknown. It enabled us to correct 8 transcriptions errors. 77 errors were found in these verifications and 8 more in the more thorough verification of the

rest of the corpus. All this took approximately 2 hours 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 take additional effort. But such check will take less time than the full manual tagging of the corpus. There were 1,707 words out of the 2,501 words that were ambiguous outside context. This means 1,707 words to finally be tagged assisted with a simple dictionary (about two days of hard work, with some remaining errors, so that a more thorough check like the one suggested above is not out of order).

The number of ambiguous words increases slowly with age. This advance can be attributed to the increasing size of the child's lexicon as well as of our corpora. But for the linguist or the psychologist, even the two-year-old child requires thorough work because there is already a potential for ambiguous words of 40%. For example, even if ‘porte’ is not yet ambiguous in our corpus, the word has to be processed as if it were and this actually raises two tricky cases: ‘ça porte’ and ‘porte là’. These two cases, brought to the fore by the syntactic analysis which suggest the class verb (V) for ‘porte’, are going to be thoroughly checked with, if possible, a backcheck to the original audio or video recordings. The use of automatic tagging does not cancel every intervention of the investigator, but it speeds up the work in a non-negligible way. Thus, going back to the example of normal children, the time needed to process a corpus of 95,000 words can be cut by as much as two weeks to one month, depending on the type of corpus (variable or even) and the needs of the user (general statistics or detailed analysis). If the user just want to have a rough idea of the figures that can be drawn from a corpus, then it will only need 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 the language deficit of these children. In fact that 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 phenomenon and provides a precise evaluation of the relative impact of occurrences considered as exceptional or frequent.

The fact that it was relatively easy to use morphosyntax to analyze our corpora of child language suggests certain theoretical developments about syntax acquisition in the child which are yet outside the goal of this paper and should require further specific studies. Our very first analyses, realized with an initial training done on adult corpora of written texts, did perform well, as if the syntax of the child was nearly the same as that of the adult. Only special cases like the ones put into brackets in Table 4, the analysis of the very numerous exclamations and the utterances reduced to one word, have led us to undertake further training. It is true that the analyzer we used was especially suited to the processing of fragments of sentence or incomplete utterances, but this implies that the language of the child is built of correct fragments of the adult language. Also, the analyzer makes it possible to easily identify the

real ‘agrammatical’ utterances – and to evaluate this agrammatism; that is, to see which elements (words) are missing (or incorrect) in the discourse of the child.

The tagging procedure and software presented here is destined to be given to the public domain. It is a program directly resulting from research and intended for professional users but its use by the computer neophyte should not be problematic, especially in the fully automatic mode. Rather, it may prove a good addition to existing programs like SALT (Miller & Chapman, 1982-95), CLEAR (Baker-Van den Goorbergh & Baker, 1991) or Computerized Profiling (Long & Fey, 1995b). It can already, for example, be used to tag CLAN files. The CHILDES system already provides a program for morphological analysis: MOR. The current analyzer should not be seen as an alternative of MOR, but as an complement. MOR gives full decomposition of words into morphemes, which POST does not. But MOR has only some simple rules for automatic disambiguation, as explained by MacWhinney (1994, 1996) pp. 423-424. It needs a future program called PARS to solve the ambiguity problem. As it is, a lot of manual work is still required to solve numerous ambiguities. This work is made easier by the latest version of CED, but long texts take a long time to be processed. POST provides a way to tackle this problem and to use natural tagged text as a training database, which is the most efficient and clear system devised till now. The linguistic performance of POST is still relatively simple because it provides only coarse-grained lexical tags, no fine-grained morphological analysis of word structure, and does not yet retrieve information about the global structure of the sentence. But, even with these limits, it already offers increased speed of database processing, which is the main and/or only goal of numerous current computer applications. Moreover, it permits the research to obtain supplementary and new information which is not only related to quantity but also to quality – that is innovative in comparison to other software – and this suggests an original contribution in identifying the appropriateness of its principles (morphosyntax and local context) in child language.

References

Baker-Van den Goorbergh, L. (1994). Computers and language analysis: theory and practice. Child Language Teaching and Therapy, 10, 3, 329-348.

Baker-Van den Goorbergh, L., & Baker, K. (1991). 1991: Computerised Language Error Analysis Report (CLEAR) (Kibworth, Leics: FAR Communications).

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

Brill, E. (1994). Some advances in transfromation-based part of speech tagging. In AAAI-94, Proceedings of the 12th National Conference on Artificial Intelligence, Cambridge: MIT Press.

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

Chevrie-Muller, C., Simon, A. M., & Decante, P. (1981). Epreuves pour l'examen du langage (Paris, France: Editions du Centre de Psychologie Appliquée).

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

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 and 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 and Therapy, 11, 185-192.

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

MacWhinney, B. (1994). New horizons for CHILDES research. In J. L. Sokolov & C. E. Snow (Eds.), Handbook of research in language development using CHILDES. Hillsdale: Lawrence Erlbaum Associates.

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

MacWhinney, B. (1995). The CHILDES project: Tools for analyzing talk (2nd edition) (Hillsdale: Lawrence Erlbaum Associates).

MacWhinney, B. (1996). American Journal of Speech-Language Pathology, 5, 5-14.

Miller, J. F., & Chapman, R. S. (1982-95). SALT: Semantic Analysis of Language Transcripts (Wisconsin: Language Analysis Laboratory, Waisman Centre 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, 4, 301-307.

Parisse C. (1989). Reconnaissance de l'écriture manuscrite: analyse de la forme globale des mots et utilisation de la morpho-syntaxe (Unpublished PhD thesis, Université de Paris-Sud, Orsay, France).

Perkins M. (1994). Repetitiveness in language disorders: a new analytical procedure. Clinical Linguistics and Phonetics, 8, 321-336.

Perkins M. (1995). Corpora of disordered spoken language. In G. Leech, G. Myers and J. Thomas (Eds.), Spoken English on Computer: Transcription, Mark-up and Application. (London: Longman).

Perkins, M., Catizone, R., Peers, I., & Wilks, Y. (1997). Clinical computational corpus linguistics: a case study (Paper presented at the 6th Annual Conference of the ICPLA, Nijmegen, Holland).

Rondal, J. A., Bachelet, J. F., & Pérée, F. (1985). Analyse du langage et des interactions verbales adulte-enfant. Bulletin d'Audiophonologie, 5/6, 507-536.

Acknowledgments

This work was supported by a grant from INSERM, France: ‘Contrat de Recherche Inserm (4U009B)’. Grateful thanks to Isabelle Barrière for proof reading this text.

Point of contact

Christophe PARISSE or Marie-Thérèse LE NORMAND
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

Table 1
List of 25 morphosyntactic categories used for the child from age 2 to age 4

Tag for the class	Number of occurrences at 2 years old	Number of occurrences for SLI children	Number of occurrences at 4 years old	Description of the morphosyntactic class
A	87	46	2,753	Verb 'to have'
ADJ	107	30	2,475	Adjective
ADV	159	110	4,023	Adverb
ADV-l	273	202	3,585	Adverb of place
ADV-n	130	119	3,255	Adverb of negation
ART	181	203	8,977	Article
ART-g	9	1	1,171	Generalized article
COJ	31	37	1,915	Conjunction
E	246	129	4,745	Verb 'to be'
I	241	132	2,207	Interjection
I-e	253	65	1,590	Interjection of exclamation
NB	6	17	503	Number
PN	21	93	509	Last name, proper name
PP	198	88	2,323	Past participle
PREP	15	69	3,657	Preposition
PREP-a	42	48	1,599	Preposition article
PRN	303	271	14,980	Pronoun
PRN-d	122	60	2,411	Demonstrative pronoun
PRN-r	72	34	2,148	Relative or interrogative pronoun
S	847	386	13,909	Noun
V	169	208	8,999	Verb
V-inf	115	115	4,558	Infinitive
V-ppre	--	--	25	Present participle
VOICI	100	10	838	Locution 'voici', 'voilà'
Y	38	27	1,009	Pronouns 'Y', 'EN'
Total number of occurrences	3,765	2,501	94,164	

Table 2
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

Table 3
Ambiguity rate according to age, either in a relative way, either in absolute

Age Year Month	Words number	Relative ambiguity (relative to children of the same group) Ambiguous words number	Rate of ambiguity per word	Absolute ambiguity (relative to adult) Ambiguous Words Number	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	63707	1.79	87,691	2.52

Table 4
Examples of ambiguities at the age of 2 years 0 month and SLI children

Word	Ambiguous classes	Number of occurrences	Examples #1	#2	#3
Normal children of two years 0 month					
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/m. f.)	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 çà	
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/m. f.)	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 !	

Table 5
Rates of errors (total number of errors, errors induced by unknown words, elements to be verified by hand and non-signaled errors).
2.0-a means that the age of the test was 2 years 0 month and the training used an adult oral corpus. Column 2.3-4.0 includes the average values of tested age ranging from 2 years 3 months to 4 years 0 month, with a training corpus made of the collapsing of an adult oral corpus and younger children corpora. 4.0-ae is the same as 2.3-4-0 but with tested children aged only 4 years 0 month. 4.0-e means that the tested age was 4 years 0 month and with a training of younger children corpora only. All i-(a,e,ae) mean that the tested corpora came from children who display severe language impairment. In the case of i-a, the training corpus was of adult oral language. In the case of i-e, the training corpora was of young children language only. In the case of i-ae, the training corpus was composed of a mix of adult oral corpus and young children corpora.

	French normal developing children				French language-impaired children		
Age	2.0-a	2.3-4.0	4.0-ae	4.0-e	i-a	i-e	i-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 non-signaled errors	1.07	0.45	0.74	0.23	3.32	0.20	0.83

Figure 1
Presentation of the analyzer at work

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) Choose the name of the output file (in CHAT format): result.cha
4) Use at the DOS prompt the following command: 'ANALYZE french.db example.cha result.cha'
5) Edit the resulting file (result.cha) with an editor or the CED software. The result is presented in a special field called '%ms' that may complement the '%mor' field of the CHILDES system.

Sample of a result file:

*PHI: non. (no)
%ms: non ADV-n . PCT
*PHI: les cheminées. (the chimneys)
%ms: les ART cheminées S . PCT
*PHI: y en a une qui est tombée. (one of them which is down)
%ms: y Y en Y a A une ART/PRN qui PRN-r est E tombée PP . PCT
*PHI: xxx je ferai une autre maison. (xxx I will make another house)
%ms: xxx I/S je PRN ferai V une ART autre ADJ maison S . PCT
*PHI: ça s' appelle des jetons. (it is called chips)
%ms: ça PRN-d s' PRN appelle V des ART jetons S*/ADJ*/... . PCT
*PHI: là c' est quoi? (this what's that?)
%ms: là ADV-l c' PRN est E quoi PRN-r ? PCT
*PHI: si on prend ça. (if one takes that)
%ms: si ADV on PRN prend V ça PRN-d . PCT
*PHI: de la confiture. (some jam)
%ms: de PREP la ART/PRN confiture S*/ADJ*/... . PCT
*PHI: ouais. (yeah)
%ms: ouais I . PCT
*PHI: voilà là, c' est des miettes? (here it is, that some crumbs)
%ms: voilà VOICI là ADV-l , PCT c' PRN est E des ART miettes S*/ADJ*/... ? PCT

(the '/' marks the unsolved ambiguities with the most probable one in first position, the '*' marks unknown words).