# TEXT ANALYSIS SOFTWARE TO HELP LEARNERS WRITE IN FRENCH.

AUDRAS, ISABELLE
GANASCIA, JEAN-GABRIEL

## ABSTRACT

New text analysis software developed thanks to research in areas such as Machine Learning and Natural Language Processing is also useful in language theory and research. Littératron is a new data-processing tool for automatic syntactic pattern extraction that was designed at the LIP6 laboratory by Jean-Gabriel Ganascia. By syntactic pattern we mean an association of coherent linguistic units. More exactly, the inputs of Littératron are syntactic analysis trees, provided by a linear text analyzer, and its outputs being recurrent syntactic patterns. In addition, Littératron is able to compare several texts in order to detect which syntactic pattern is present in one text and absent from another. It is this kind of discrimination which issued to help build the characteristics of learner writing. Littératron performs a learner cognitive diagnosis by analysising the linguistic style of the written French of learners of French as a foreign language.

Our experiments were based on certification in French as a foreign language following the guidelines laid down by ALTE (Association of Language Testers in Europe). The narrative framework of writing tests is postcards, friendly letters and personal statement. Because we work with written productions with strong textual models, comparing these productions allows us to detect syntactic particularities (linguistic and stylistic mistakes for example).

The acquisition of French as a foreign language is studied here by comparing the nature and frequency of extracted syntactic patterns taken from the written production of learners and native speakers. The learner may come from a heterogeneous group (different language levels and different mother tongues) or from a homogeneous group (only one language level and one mother tongue, here Arabic). We found decisive patterns in these two groups of learners.

In the aim of computer-assisted language learning, a cognitive diagnosis can be carried out using Littératron to study stylistic figures that are counted using recurrent patterns. The diagnosis is built from the stylistic mistakes taken from the learners' written work. Our software extracts a set of characteristic figures concerning noun expansion for instance, or the use of adjectives, adverbs, punctuation, etc. It then analyses the mistakes or the over-use of certain expressions using these figures.

Future developments might be to transfer the data to the inference module, which is

based on a rule database in order to establish the learner profile. This profile could be used as a diagnostic tool by the remote tutor. From the information provided by the system about the characteristics of the learner's mistakes, the tutor can make choices and point the learner to the learning activities that are best adapted to the learner's needs, as well as show him how to build a representation of his own learning and needs.
This approach can be of interest in three fields: language teaching, on a purely educational basis; computational linguistics; computer-assisted learning (as a tutoring tool).

## PRESENTATION

The approach to foreign language teaching could be transformed considerably by the use of Natural Language Processing tools. This idea might seem strange at first sight, because automatic language processing aims, among other things, to remove language barriers thanks to the use of the computers, thus making foreign language acquisition less necessary. In this context the teaching of foreign languages would not be transformed, it would simply disappear... But we can look at it in another way and that is what we are doing here. Thus, from out point of view, it is not a question of eliminating foreign language teaching, but of facilitating it by benefiting from the knowledge obtained thanks to automatic language processing tools. The acquisition of writing skills in French as a foreign language is of scientific interest through the concept of a style produced by learning. In a communicative context, the style, understood as the horizontal organization of the words, is defined by the choice of the linguistic expression from among all those known by the learner. Working on style in a foreign language class helps enrich grammar and vocabulary while introducing the subtleties and distinctions for a correct and varied form of expression. Each foreign language learner makes mistakes during the learning process but they are constructive and allow the learner to progress if they are understood and corrected. In addition, the style of the errors acts as an indicator for the tutor, who is able to establish a remedial programme on the basis of these error types (nature and frequency of the error). It should be added that French as a foreign language is assessed by a certification system called DELF (Diploma of Studies in French Language), based on ALTE. The instructions for the written tests, A1, A2, B1, are postcard, friendly letter, personal statement.
The object of this article is to present the possibilities for the use of automatic language procesing tools in the acquisition of writing skills, in terms of comprehension, evaluation and learning.

## 1. THEORETICAL POSITIONING

Research on the acquisition of writing in a foreign language is recent, and has profited from the results of research about the acquisition of writing skills in one's mother tongue. Writing distances the user from the language and helps memorize it. The cognitive virtues of learning to write no longer need to be proved (Mangenot, 1998). Moreover, results of research in text linguistics also apply to the teaching of writing skills. Text coherence and pragmatic concepts, which come within text linguistics, require the development of a textual competence, which gives the learner the rhetorical tools necessary to construct a discourse. In other words, "the formal textual structures (...) guide the writer in the construction of a text and the reader in his comprehension" (Scardamalia & Bereiter, 1986). According to Tuffs (Tuffs, 1993), working on different text types facilitates the acquisition of foreign languages. Generally speaking, learning to write in a language class means obeying strict instructions within a narrative framework dictated by the kind of text or a precise communicative objective. This necessitates functional objectives whose morphosyntactic and lexical expression are taught in the language class. This linguistic content, found within a communication situation, is automated during re-employment, and this is always true if the context is similar. In addition, the analysis of communicative needs from the narrative framework helps the learner to adapt in a new situation (Tagliante, 1994). Wether in a classroom situation or during the DELF test, the level of learning is validated in terms of the ability of the learner to express a written message through a model that has been learned and recognized rather than in terms of the learner's grammatical competences. It is for these reasons that the aim is to use the current techniques of Natural Language Processing to locate common written errors in a learner population, so as to make it possible to correct them, during class. Errors can be detect either in absolute terms, by detecting the syntactic errors, or by comparing language use between learner and native speakers, in a precise narrative framework. It is this second approach which we have adopted, and we know that role of the language teachers is not to teach a perfect and abstract language but to transmit the different uses of a language. More exactly, the work presented here uses automatic stylistic analysis tools to establish the linguistic characteristics of learners, according to their level, and to distinguish them from native speakers. Empirical studies have been done with three different learner populations to validate the approaches suggested: one adult learner population is made up of people with different mother tongues in an endolinguistic learning context, the second population is made up of Arabic-speaking adults and teenagers in an exolinguistic learning context and the third is made up of Creole-speaking in a second language learning context. In the last part, we shall consider the possibilities

for CALL research to use the linguistic profiles produced by Littératron.

## 2 DATA-PROCESSING TOOLS USED

Two data-processing tools are necessary to extract the syntactic patterns, which are characteristic of various learner populations. A syntactic pattern is defined as an association of coherent linguistic units. Here is an example of an extracted pattern from the analyzers with the syntactic structure (preposition + personal pronoun + verb in the infinitive form): "de vous adresser", "de m'investir", "de vous donner". These three patterns were extracted together from the same group of personal statement learner writers.

The first data-processing tool is a French morphosyntactic analyzer, which builds syntactic trees from written texts. We used Vergne, a linear analyzer with partial dictionary, which was created by Jacques Vergne, University of Caen, in 1998 (Vergne, 2001). The second is Littératron, the stylistic analyzer developed at the LIP6 by Jean-Gabriel Ganascia (Ganascia, 2001), which outputs recurrent syntactic patterns from stratified ordered trees (SOT). More exactly, the Vergne analyzer associates a tag to each word or word group; a structured ordered tree is thus a partition of tags whose classes depend on the depth of the node in the tree analysis. Littératron calculates a similarity measurement between several SOT, based on the distance-edition concept, and generates a similarity graph which records the sub-trees closest to the SOT. It is this similarity graph, which is then used as input for the classification algorithm of Littératron, called "centre-stars", which builds similar classes of patterns and gives them a meaningful name. The "centre-star" algorithm initially evaluates all the centered stars on the various nodes and then takes, for each one, the sum of the similarity values of the nodes of each star in the centre. Once the value of each star has been computed, the "centre-stars" algorithm keeps the one with the highest value. The nodes which belong to this first star are marked and then same algorithm is applied recursively to the unmarked nodes, until no unmarked node remains. In short, any star is a subgraph of the similarity graph, centered on a node. For each class thus built, the algorithm chooses the most similar pattern to the centre of the star, to illustrate the significance of the star. It also indicates the source text covered by each pattern. Here we have the example of a centre-star, illustrated in figure 1:

( PREP ( "de" ) ) + ( GN ( ART ( "la" ) ) + ( NAME ( "forêt" ) ) ) (text: "de la forêt"), with which the 5 following syntactic patterns are associated:

- (PREP ("à")) + (GN (ART ("l")) + (NOM ("auberge"))) (texte : "à l\'auberge ");
- (PREP ("d\")) + (GN (ART ("un")) + (NOM ("hiver"))) (texte : "d\'un hiver");
- (PREP ("dans")) + (GN (ART ("le")) + (NOM ("monde"))) (texte : "dans le monde");

- (PREP ("avec")) + (GN (ART ("les")) + (NOM ("chiens"))) (texte : "avec les chiens");
- (PREP ("depuis")) + (GN (ADJ ("quelques")) + (NOM ("jours"))) (texte : "depuis quelques jours").

This means that the similarity measurement between the first pattern ("de la forêt") and one of the derived trees from the syntactic trees of each one of these five nominal groups is higher than a certain threshold. These five patterns belong to the same star whose centre is of the form: ( PREP ( "de" ) ) + ( GN ( ART ( "la" ) ) + ( NAME ( "forêt"))).

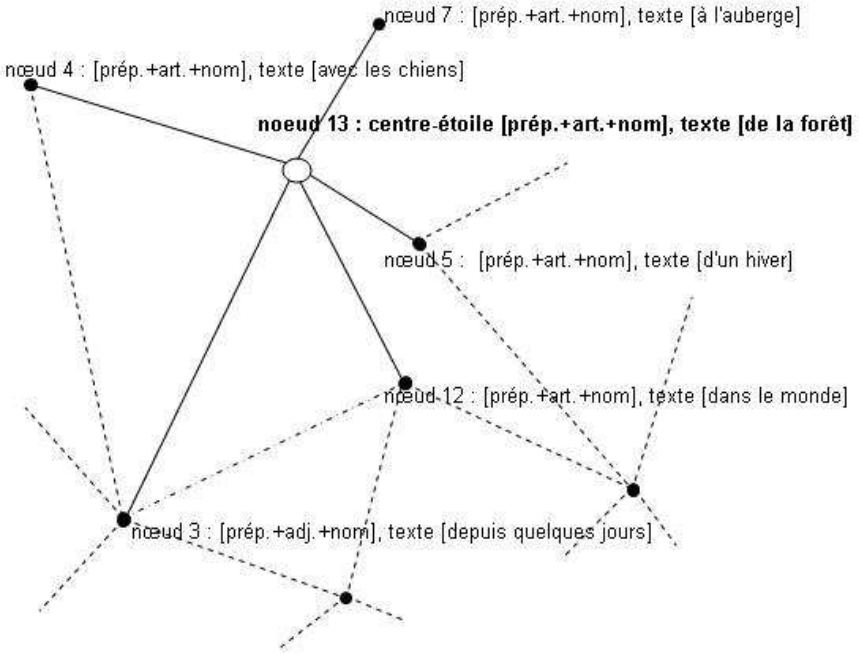


Figure 1 - Graph of the centre-star given in the example.

In addition to building the star and extracting pattern, Littératron carries out a second type of operation which consists in comparing stars resulting from several texts in order to locate stars which are present in one and absent from the other. This makes it possible to identify, among the patterns present in a written text, those which distinguish it from other written

text. It is starting from this type of identification that the syntactic characteristics of learner populations will be built. In figure 2, we see the result of the Littératron analysis of three different personal statements corresponding to the three different categories of learner: personal statement of learners with different mother tongues (a); personal statement of Arabic-speaking learners (b); personal statement of French-native-speakers (c).

Pattern N°46

Complete sentence

S

Pp 1 s

V

COD

N c f s

Examples[1]:

(a): Je parle la langue anglaise et française

(b): J'apprends la presse à l'université de Naplouse

(c): Je maîtrise la mise en place de l'organisation de l'archivage

Figure 2 - Result of the Littératron analysis of three different personal statements.

This example describes the syntactic structure of the extracted pattern. We have a complete sentence whose form is: subject (S) + verb (V) + direct object (COD), where the subject is a singular first person personal pronoun, (Pp 1 S) and the COD is a feminine singular common noun (Nc F S).

## 3. FIRST TYPE OF EXPERIMENT: LEARNERS HAVE DIFERENT MOTHER TONGUE.

The aim of this research is to compare the written work of learners of French as a foreign language at different levels with the written work of French-speaking people, within the

[1] Literal translations: (a): I speak the French and English Language; (b): I learn the press at the University of Nablus; (c): I master the implementation of the organization of the archiving.

same communicative context. The French native-speakers have a high level education: at least 4 years higher education. Other studies could look at French native-speakers with a lower level of, or even adults who are beginning to learn to write (Morais & Kolinsky, 2001). In other words, the criterion of level of education may be important.

3.1. Written learner texts and experimental methodology

Four types of text were selected: postcard (PC), friendly letter (FL), personal statement (PS), description (Des). Each written text corresponds to a different learning level in French as a foreign language. For the description, each learner, whatever their level, has to describe the same picture (village square, primitive art). Table 1 has a dual function. Firstly, it recapitulates the experiments by type of text. For example, for the "post cards" (PC), the texts written by the beginners and the French native-speakers were introduced simultaneously into the analyzers. Secondly, it details the total number of texts for each type.

| | Learners | | | French native-speakers (FNS) |
|---|---|---|---|---|
| | Beginners (A1) | Intermediate (A2) | Advanced (B1) | |
| Post cards (PC) | 6 | | | 6 |
| Friendly letter (FL) | | 4 | | 4 |
| Personal statement (PS) | | | 6 | 6 |
| Description (Des) | 5 | 5 | 5 | 5 |

Table 1 - Type and number of written texts.

3.2. Results and comments

The results obtained are statistical, to which we have added linguistic comments on the extracted pattern.

| | PC Beg. | PC FNS | FL Inter. | FL FNS. | PS Adv. | PS FNS | Des Beg. | Des Inter. | Des Adv. | Des FNS |
|---|---|---|---|---|---|---|---|---|---|---|
| Number of centre-stars | 6 | 10 | 2 | 5 | 6 | 6 | 2 | 3 | 3 | 13 |
| % text | 50 | 50 | 60 | 30 | 25 | 17 | 33 | 33 | 35 | 14 |

Table 2 - Number of centre-stars and percentage of texts represented by the centre-stars.

Table 2 gives the numerical results of the statistical calculations carried out by the analyzer. Littératron uses the same parameters for all the texts, in particular the thresholds of the centre-star algorithm and the similarity graph. In other words, the number of stars detected is thus a good indicator of stylistic richness: the more stars there are, the richer the style, i.e. the fewer automatisms there are. The same applies to the percentage of text represented by the stars: the lower the percentage, the greater the variety of the patterns, which means that the style is richer. Note that this concept of stylistic richness is relative; after all, a great writer could well use a limited range of patterns, while a writer with no style might use many. In spite of these reservations, in the particular case of language teaching which interests us here, we compare the richness of a text (or a set of texts) to the number of syntactic figures used.

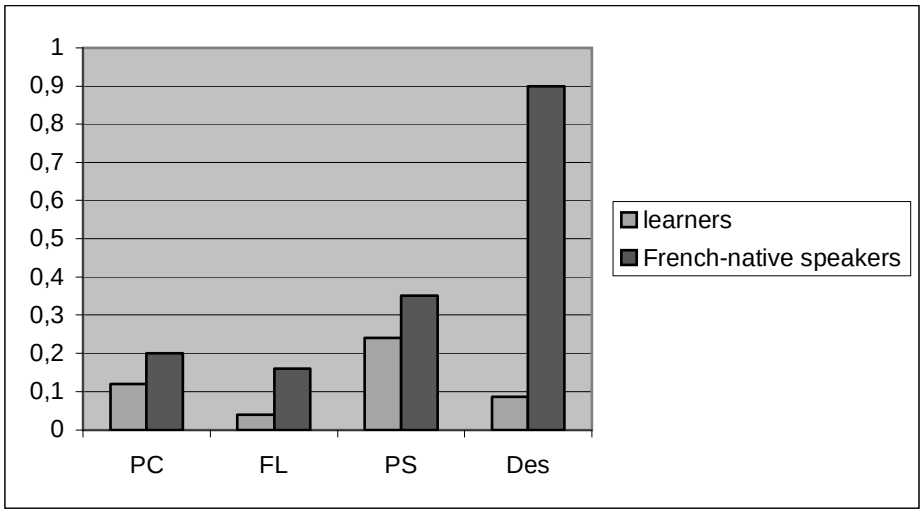


Figure 3 - Variability index according to the type of text.

In figure 3, the variability index, for each type of text, is defined as the ratio of the number of stars detected to the percentage of texts used by the application. What can be concluded from the number of recurring syntactic patterns and the percentage of texts

covered by these patterns ? First, the statistical results represented by the variability index show us that for the same kind of written text, the syntactic patterns extracted by the application from texts produced by French native-speakers are greater in number and more varied and than those produced by learners. Second, the percentage of text not represented by the recurring syntactic patterns is twice as high in PC, PA and FL texts and nine time as high in Des texts produced by French native-speakers than by even the most advanced learners. The part of the text where Littératron does not detect any recurring patterns could be used in the future to identify the category of writer or even the individual writer. This analysis shows automatisms in writing within different types of text, both those produced by learners and those produced by French native-speakers. There are thus objective and measurable models of how to write postcards, friendly letters or personal statements.

## 4 SECOND TYPE OF EXPERIMENT: LEARNERS HAVE SAME MOTHER-TONGUE

The aim of the experiment is to detect a specific of French syntax within a homogeneous class.

### 4.1. Arabic-speaking learners

#### 4.1.1. Presentation of the written texts

The texts that where analysed correspond to the Civilisation and History examination paper of (argumentative discursive framework), of students of a French department in an Arabic-speaking university. English is their first foreign language, French their second. There are ten texts.

The experimental methodology is the same as for the first experiment: each type of text is introduced successively into the morpho-syntactic analyzer then into Littératron, as are texts written by French native-speakers who were given the same instructions. The French native-speakers have at least 4 years higher education.

#### 4.1.2. Results and comments

The same three syntactic patterns appear systematically in the written work of the Arabic-speakers. These patterns cover 25% of the analysed text: two noun groups and one verb group. The two noun groups: preposition DE + adjective + noun, as in the examples: "de choses magnifiques" and "d'autres villages".

This syntactic pattern shows a massive use of noun groups after the verb starting with the preposition DE, to the detriment of other determiners and other prepositions. The student does not know the correct construction of a verb, direct or indirect, or the right use of the

partitive article, the definite article or the indefinite article and systematically uses a preposition to introduce the object. Learning does not seem to help use prepositions and determiners correctly. This experiment shows that for a homogeneous class of Arabic-speaking learners, there are particular aspects of French syntax, which cannot be learned without special help. Let us take one of the two often-repeated extracted patterns: "de choses magnifiques". This pattern appears in several sentences and was produced by different learners. Here are two examples, taken from two learners:

- learner 1: "J'ai visité *de* choses magnifiques" [1]

- learning 2: "J'ai vu *de* choses magnifiques"

These two occurrences show an erroneous use of the determiner DE which introduces the object after the verb "visiter" in the first example and "voir" in the second.

4.2. Créole-native speakers
An experiment in progress, based on the same instructions given to Créole native-speaking learners when producing written texts seems to converge towards the same conclusion, with an obvious recurrence of patterns of the form: de + adjective + noun.

## 5. CONCLUSION AND FUTURE RESEARCH

Used in research into learning how to write in French as a foreign language, Littératron is able to carry out a linguistic diagnosis of learning based on written texts within the constraint of a narrative framework. This diagnosis is used to help learn how to write in French as a foreign language. The examples given above show that Littératron is able to

[1] The correct version would be: « J'ai visité des choses magnifiques »; « J'ai vu des choses magnifiques »

model the learner's expression on the basis of style errors identified in the learner's written production. By identifying the errors in the written production, Littératron shows objectively some of the morpho-syntactic characteristics present in these texts. If we consider the information extracted from learners written work in the light of language teaching theory, based on computer-assisted learning and evaluation of a style in a second language, we see that Littératron is able to build a linguistic learning profile. This profile could be used as a diagnosistic tool for the distant tutor who can use the information provided by the system on the characteristics of the learning errors in order to help the learner and to suggest activities according to the learner's observed needs.
Two future lines of research are necessary and complementary: technical and experimental. Here we have limited ourselves to decomposing syntagms and studying the structure of the sentence with respect to this decomposition. The problem is it restricts the type of pattern detected and it will be necessary to implement a richer decomposition which will take into account the propositional structure. The pattern extraction algorithm is identical, but the syntactic analysis is different; and in particular the tree resulting from this analysis will have to be enriched considerably. Second, the results obtained from Arabic- and Créole- speaking learners encourage us to continue studying he specific differences between learners who come from different parts of the world.

## REFERENCES


Audras I. & Ganascia, J.-G. (2005). Analyses comparatives de productions d'apprenants du français et de francophones, à l'aide d'outils d'extraction automatique du langage. *Acquisition du Langage et Système d'Information et de Communication (A.L.S.I.C), 8*, 81-94. Available: http://alsic.u-strasbg.fr/v08/audras/alsic_v08_16-rec10.htm.

Fayol, M. (1996). La production du langage écrit. In J. David & S. Plane (Eds.), L'apprentissage de l'écriture de l'école au collège (pp. 9-36). Paris : Presses Universitaires de France.

Ganascia, J.-G. (2001). Extraction automatique de motifs syntaxiques. In *Actes du colloque Traitement Automatique du Langage Naturel 2001 (TALN 2001).* Available: http://www.li.univ-tours.fr/taln-recital-2001/index1.html

Ganascia, J.-G. (2001). Extraction of Recurrent Patterns from Stratified Ordered Trees. In *Procedings of the 12th European Conference of Machine Learning 2001 (ECML 2001).* Freiburg: Springer-Verlag, pp. 167-179.
Mangenot, F. (1998). Outils textuels pour l'apprentissage de l'écriture en L1 et en L2**.** In Souchon, M. (dir.), Pratiques discursives et acquisition des langues étrangères (pp. 515-525)

Actes du X^e colloque international Acquisition d'une langue étrangère : perspectives et recherches. Besançon : Université de Franche-Comté.

Morais J. & Kolinsky R. (2001). The literate mind and the universal human mind. In Dupoux E. (ed.). Language, brain and cognitive development : Essays in Honor of Jacques Mehler (pp. 463-480). Cambridge, Mass.: MIT Pr.

Pery-Woodley, M. -P. (1993). Les écrits dans l'apprentissage. Clés pour analyser les productions des apprenants. Paris : Hachette FLE.

Scardamalia, M. & Bereiter, C. (1986). Research on written composition. In Wittrock, M.C. (dir.). Handbook of research on teaching (pp 778-803) New York : McMillan.
Tagliante, C. (1994). La classe de langue. Paris : CLE International.
Tuffs, R., (1993), A genre approach to writing in the second language classroom: the use of direct mail letters. In *Revue belge de philologie et d'histoire, 71, 3*, 691-721.

Vergne, J. (1999). Etude et modélisation de la syntaxe des langues à l'aide de l'ordinateur. Analyse syntaxique non combinatoire. Synthèse et résultats. Available: http://users.info.unicaen.fr/~jvergne/#HDR.

Vergne, J. (2001). Analyse syntaxique automatique de langues : du combinatoire au calculatoire. In *Actes de Traitement Automatique du Langage Naturel 2001 (TALN 2001)*. Available: http://users.info.unicaen.fr/~jvergne/Taln2001FR_JV.pdf

We would like to thank Rosalind Greenstein, Université Paris I for the English version of this paper.

## APPENDIX/APPENDICES

## BIODATA

Isabelle Audras is a French as Foreign Language teacher. She is a PhD student in the field of Cognitive Sciences at the *Laboratoire d'Informatique de Paris 6*, where she works with Jean-Gabriel Ganascia about the acquisition of written French with text analysis software. She is although attached to the Department of Language Sciences (French as Foreign

Language section) at the *Université de Franche-Comté* for teaching and research (in the field of Distance Learning).

## CONTACT

Isabelle Audras

Université Pierre et Marie Curie, Paris 6
LIP6
8 rue du Capitaine Scott
75015 Paris
France

Université de Franche-Comté
Laseldi
30-32 rue Mégevand
25000 Besançon
France

Isabelle.Audras@lip6.fr