CEFLE and Direkt Profil: a new computer learner corpus in French L2 and a system for grammatical profiling

Granfeldt, Jonas; Nugues, Pierre; Persson, Emil; Thulin, Jonas; Ågren, Malin; Schlyter, Suzanne

Published in:
Proceedings of the 5th International Conference on Language Resources and Evaluation

2006

CEFLE and Direkt Profil: a New Computer Learner Corpus in French L2 and a System for Grammatical Profiling

Jonas Granfeldt*, Pierre Nugues†, Malin Ågren*, Jonas Thulin†, Emil Persson*, Suzanne Schlyter*

*Centre for languages and literature – Lund university
Box 201, S-221 00 Lund, Sweden
{Jonas.Granfeldt, Malin.Agren, Suzanne.Schlyter}@rom.lu.se, Emil.Persson@telia.com
† Department of Computer science – Lund Institute of Technology
Box 118, S-221 00 Lund, Sweden
Pierre.Nugues@cs.lth.se, f00jt@efd.lth.se

Abstract

The importance of computer learner corpora for research in both second language acquisition and foreign language teaching is rapidly increasing. Computer learner corpora can provide us with data to describe the learner's interlanguage system at different points of its development and they can be used to create pedagogical tools. In this paper, we first present a new computer learner corpora in French. We then describe an analyzer called *Direkt Profil*, that we have developed using this corpus. The system carries out a sentence analysis based on developmental sequences, i.e. local morphosyntactic phenomena linked to a development in the acquisition of French as a foreign language. We present a brief introduction to developmental sequences and some examples in French. In the final section, we introduce and evaluate a method to optimize the definition and detection of learner profiles using machine-learning techniques.

1. Introduction

The importance of computer learner corpora (CLC) for research in both second language acquisition and foreign language teaching is rapidly increasing. As pointed out by Granger (2004), CLCs can serve different purposes in the research process. They can provide us with data to describe the learner's interlanguage system at different points of its development and they can be used to create pedagogical tools. CLCs might also be used indirectly to improve classroom practice.

In this paper, we first present a new CLC in French, the CEFLE corpus. We then describe an analyzer called *Direkt Profil*, that we have developed using this corpus. The system carries out a sentence analysis based on developmental sequences, i.e. local morphosyntactic phenomena linked to a development in the acquisition of French as a foreign language. The objective of the program is to establish a learner profile based on the grammatical features of the input text. We present a brief introduction to developmental sequences and some examples in French. We also present and evaluate some recent developments in Direkt Profil. In the final section, we introduce and evaluate a method to optimize the definition and detection of learner profiles using machine-learning techniques.

2. The CEFLE Corpus

The Lund CEFLE Corpus (*Corpus Écrit de Français Langue Étrangère*) is a written corpus of texts in French as a foreign language. This longitudinal corpus contains approximately 400 texts (100,000 words) written by Swedish learners of French with different levels of proficiency and by French native speakers in a control group. CEFLE is the result of a study that surveyed 85 learners of French in the Swedish high school throughout the academic year 2003/2004. During this period, each learner wrote four texts in French at two months intervals. The aim of this study was to analyze the morphosyntactic development in written production. The control group of 22 native speaking adolescents is completing this material.

The foreign language learners in the CEFLE corpus have Swedish as their mother tongue and they are advanced L2 learners of English. French corresponds to their second or third foreign language. They all learn French in a traditional instructional setting at the Swedish high school. The beginner learners are attending their first year of French when writing the first text. The most advanced learners started their fifth year of French at the beginning of the study.

CEFLE contains texts from four different tasks, which were created to elicit written data as spontaneously as possible from all kinds of learners. Two different task types were used: (1) story telling tasks based on picture sequences, (2) descriptive narratives based on personal experiences. The texts *L'homme sur l'île* 'The man on the island' and *Le voyage en Italie* 'The trip to Italy' are representing the first task type, while *Moi, ma famille et mes amis* 'Me, my family and my friends' and *Un souvenir de voyage* 'Memory of a journey' are representing the personal narratives. All texts were written on a computer using plain text formatting.

The texts from one of the four eliciation procedures, *Le voyage en Italie* 'The journey to Italy', has been used as a subcorpus receiving special attention in several respects: a cross-sectional linguistic analysis was carried out on this material (Ågren, 2005) and these texts are used in the work with Direkt Profil. Developmental sequences based on morphosyntactic criteria (Bartning and Schlyter, 2004) were used to place the learner texts on four levels of development: stage 1 (initial), stage 2 (post-initial), stage 3 (intermediate), and stage 4 (preadvanced). This part of the corpus is annotated for a specific set of lexical or syntactic phenomena using the XML format. A brief description of the linguistic levels in the subcorpus is presented in Table 1. Vocd is a measure of vocabulary diversity developed on the basis of the traditional type-token ratio by

CEFLE corpus			Subcorpus *Le voyage en Italie* (averages)			
Task name	Elicitation type	Words		Text length	Sent.Length	Vocd
Homme	Pictures	17,260	Stage 1 (N=10)	127	7.0	40.5
Souvenir	Pers. Narrative	14,365	Stage 2 (N=29)	175	8.4	53.5
Italie	Pics	30,840	Stage 3 (N=39)	276	9.9	60
Moi	Pers. Narrative	30,355	Stage 4 (N=17)	369	11.8	74
Total		92,820	Control (N=22)	334	9.7	104

Table 1: General description of the CEFLE corpus and the subcorpus *Le voyage en Italie*.

Malvern et al. (2004). A high Vocd value is interpreted as a rich vocabulary.

3. Developmental Sequences in French L2

Developmental sequences are features and constructions linked to a development over time in the grammar of the language learner. Beginning in the 1970s, the so-called *morpheme order studies* identified the order in which grammatical morphemes like *–ing*, *the*, *a* and *'s* simply appeared in English spoken by nonnative speakers (Dulay and Burt, 1974; Bailey et al., 1974). Some argued that these sequences were universal referring to them as the "natural order of development". More recently Pienemann (1998) framed the development of German L2 in six grammatically defined stages. The underlying rationale behind these kinds of proposals is the idea of the learner language as its own grammatical system, an *interlanguage* (Selinker, 1972), independent from the target language system. Theories differ however, when it comes to accounting for the observed development.

According to the *Processability Theory* (Pienemann, 1998), currently the most detailed account for sequences of L2 development, the learner can only produce structures that can be processed. In this theory, acquiring to produce linguistic structures is seen as a process of automatization where each step in the development builds on the previous one. Development is constrained by limits of the working memory. Automatizing the processing will free working memory capacity that will in turn enable the learner to process and produce increasingly more complex structures. Consider an example from our corpus: *Un soir les filles mange dans un restaurant*. In French, a language with rich subject-verb agreement in the written system, the writer must store in memory the features of the subject when producing the verb. In this case, the subject is in 3rd person plural (*les filles*) and the verb should be marked accordingly (*mangent*). In the example above, the learner has instead used a default form, the 3rd person singular (*mange*) thus producing a typical learner error. What is important for us, is that the theory of developmental sequences and developmental stages in the learning of foreign language predicts that learners in their production of S-V agreement and other features will show optionality. For the relevant features, optionality is developmental in nature and a characteristic of transitional stages. Therefore, we expect a lack of S-V agreement to be more important in initial stages of development, and perhaps with more variation, and we expect it to disappear in the advanced stage.

Developmental sequences have been studied in spoken L2 French by Bartning and Schlyter (2004). They defined about 25 different morphosyntactic features and proposed a definition of their development over time in adult Swedish learners of French. Taken together, these features shape six grammatical profiles – ranging from beginner to very advanced learners. Examples of features are given in Table 2. As the language learner moves towards an increasing automatization of the target language, the produced structures become increasingly more complex and more target-like. Developmental sequences describe in linguistic terms this process.

4. Direkt Profil

Direkt Profil (DP) is the system we are developing to identify, annotate, quantify, and display the specific linguistic constructions connected to a development over time in foreign language French. In other words, DP analyzes the learners' texts for structures occurring in developmental sequences (see Table 2). The CEFLE corpus (see above Section 2.) serves as a development and test corpus in the implementation of DP. The overall architecture of Direkt Profil was described in a previous paper (Granfeldt et al., 2005) and we will limit our presentation here to some recent developments.

Verb groups and noun groups represent the essential grammatical support of our annotation. The majority of syntactic annotation standards for French take such groups into account in one way or another. The PEAS annotation scheme (Gendner et al., 2004) is a consensual example that reconciles a great number of annotations. However, in their present shape, these standards are insufficient to mark up constructions of Table 2, many of which are specific to foreign language writers. On the basis of the linguistic constructions in Bartning and Schlyter (2004), we developed our own annotation scheme. The current version of DP, 1.5.4, detects four types of syntactic groups, nonrecursive noun groups, verb groups, prepositional groups, and conjunctions, that it annotates using the XML format.

The DP architecture is a cascade of five layers. The first layer corresponds to tokenization of the text. The second layer annotates prefabricated expressions or sentences (e.g. *je m'appelle* 'my name is'). These structures correspond to linguistic expressions learned "by heart" in a holistic manner. It has been shown that they have a great importance in the first years of learning French.

The third layer corresponds to a chunk annotation of the text, restricted to the phenomena to identify. This layer marks up the verb and noun groups. As in PEAS, the verb

Stages	1	2	3	4	5	6
% of finite forms of lexical verbs in obligatory contexts	50-75	70-80	80-90	90-98	100	100
% of 1st person plural S-V agreement (*nous V-ons*)	–	70-80	80-95	100	100	100
% 3rd pers plural agreement with irregular lexical verbs like *viennent, veulent, prennent*	–	–	a few cases	≈ 50	few errors	100
Object pronouns (placement)	–	SVO	S(v)oV	SovV app.	SovV prod	acquired (also *y* and *en*)
% of grammatical gender agreement	55-75	60-80	65-85	70-90	75-95	90-100

Table 2: Developmental sequences from Bartning and Schlyter (2004). Legend: – = no occurrences; app = appears; prod = productive advanced stage.

group incorporates the subject clitic pronouns. The XML element `segment` marks the groups. Table 3 presents an evaluation in precision and recall of the chunking layer.

The fourth layer uses a `tag` element with attributes to indicate the lemma, the part of speech, and the grammatical features. For the verb group, the sentence *Ils parlons dans la bar* extracted from the learner text above receives the following annotation:

```
<segment class="c5131"><tag pos="pro:
nom:pl:p3:mas">Ils</tag> <tag pos="ver:
impre:pl:p1">parlons</tag></segment>
dans la bar.
```

The `c5131` class is interpreted as "finite lexical verb no agreement".

The fifth layer counts structures typical of an acquisition stage. It uses the `counter` XML element,

```
<counter id="c5200" counter_name=
"passe_compose" rule_id="participe_4b"
value="1"/>.
```

The analyzer uses manually written rules and a lexicon of inflected terms. The recognition of the group boundaries is done by a set of closed-class words and the heuristics inside the rules. It thus follows an old but robust strategy used in particular by Vergne (1999), *inter alia*, for French.

Direkt Profil applies a cascade of three sets of rules to produce the four layers of annotations. The first unit segments the text in words. An intermediate unit identifies the prefabricated expressions. The third unit annotates simultaneously the parts of speech and the groups. Finally, the engine creates a group of results and connects them to a profile. It should be noted that the engine neither annotates all the words, nor all segments. It considers only those which are relevant for the determination of the stage. The engine applies the rules from left to right then from right to left to solve certain problems of agreement.

The current version of Direkt Profil is available online from this address http://www.rom.lu.se:8080/profil. This version of the system implements phenomena related to the verb phrase. In (Granfeldt et al., 2005) the performance of Direkt Profil version 1.5.2 was evaluated. The results showed an overall F-measure of 0.83.

5. Determining Profiles with a Machine Learning Approach

The linguistic constructions behind the profiling method are the result of systematic empirical observations and analyses of longitudinal corpora. The stages of development and the phenomena that make them up were presented in Bartning and Schlyter (2004). These are elaborated on the basis of more than 80 individual recordings. In all, some 25 phenomena are taken into consideration when establishing a learner profile and a learner stage. In the text classification step, we consider these phenomena as features that represent the learners' texts.

5.1. Optimizing Feature Selection

We manually classified the texts of the subcorpus *Le voyage en Italie* (see Table 1) according to the development stage they were reflecting. We developed a machine learning approach to optimize the profiles on the basis of this classification. Optimizing can be of at least two types. First, this approach will limit the need for manual parameter tuning. Using this technique, we expect to be able to narrow down percentage spans like those in Table 2. For example the span for nonfinite lexical verbs at Stage 1 is estimated to go from 50% to 75%. Using this feature as a vector in the machine learning algorithm, we expect to be able to add more precision to this estimation. A second type of improvement is the identification of new features or feature engineering. In text classification, feature vectors often contain up to 10,000 features (Joachims, 1997). It is probable that we have not yet identified all the relevant features to classify learner texts according to their stage of development. Since the Direkt Profil annotation is far richer than the 25 features identified manually, there is a potential for identifying more relevant features.

Raw scores for new features can be obtained by simply counting how many times a certain rule has been applied by the analyzer. Via simple processing, we can also obtain ratios which are often better measures, for example the ratio of inflected verbs to the total number of verbs.

5.2. Machine Learning Algorithms and Tools

The machine-learning module uses decision trees based on the ID3 algorithm (Quinlan, 1986) and Support Vector Machines (Boser et al., 1992). The training phase auto-

	NPs	VPs	PPs	Conj	MWE	Total
Reference structures	216	152	112	69	29	578
Detected structures	222	163	86	73	26	570
Correctly detected structures	208	137	85	69	26	525
Recall	96%	90%	76%	100%	90%	91%
Precision	94%	84%	99%	95%	100%	92%
F-measure	0.95	0.87	0.86	0.97	0.95	0.91

Table 3: Results on segments. We have excluded misspelled words from the reference annotation. Taking into account all the words would probably yield lower figures.

matically induces classifiers from texts in the CEFLE corpus that we manually classified and the features we extract with the analyzer. Once trained, the system uses the decision trees to automatically classify texts from the learners at runtime. We will present results for three types of classifiers: C4.5, SVM classifiers, and LMT (Landwehr et al., 2003). All our experiments have been done with the Weka collection of machine learning algorithms (http://www.cs.waikato.ac.nz/ml/weka).

5.3. The Profiler Optimization Sequence

In order to describe how we are working with profile optimization, consider first the following sample learner text from the CEFLE corpus:

> *Marie et Sofia est deux filles. Marie est grosse et a blonde cheveux. Sofia est mince et a marron cheveux. Elles aimaies travaillient. Sur une semaine elles sommes travaillient en Italie. Iatlie est dans le sud en Europe. Marie a une petite vert voiture. Dans la autoroute farie de la voiture sur Italie. Le temps est belle. Arrive l'hotel Marie et Sofia sortient sur votres etage dans l'hotel. La etage est petit et a une grosse venster. Prochein semain elles baigne dans la mer. Sur la soir Marie et Sofia avec deux hommes faire le disco. Il est amour dans le voyage! Un de voyage en Italie elles faire un a rote bus sur un sightseeing. Le finir en de voyage travaillent Marie,Sofia et de deux hommes "back to" Suede!*

This text was written by a learner at stage 1. The text contains a number of features typical for learner texts and it can be analyzed for developmental stage using the developmental sequences in Table 2. Here we will focus on those features that we have used in our first experiments to train the automatic classifier. These include some features from Table 2, e.g. percentages of finite forms of lexical verbs in obligatory contexts and subject-verb agreement (all grammatical persons collapsed) but also a number of other features. In addition to grammatical features, we have used lexical features, e.g. type-token ratio (TTR), a list of all the words in the text and word frequency information. For the last feature, token frequency in a large corpus of written French, we have extracted information from the *Lexique* database (New et al., 2004) . In total, 33 features were used in the training session. These are presented in Table 4 with their respective values for this particular learner text.

Figure 1 shows an example of a resulting decision tree for classifying learner texts according to their developmental stage. Without going into to details at this preliminary stage, it is particularly interesting that the decision tree presents the features in an hierarchical manner, following their classifying weight. This will help us in further developing the profiles and adding relevant features to them (feature engineering).

5.4. Evaluating Classifier Performance

We evaluated the performance of the three different classifiers used, C4.5, SVM and LTM. We carried out two separate evaluations. We first clustered the five stages into three larger stages, where stages 1 and 2, respectively 3 and 4, were collapsed into two stages. We then ran a second evaluation with the original five stages. Currently, the best classifier, SVM, obtains an average precision and recall in the vicinity of 70 % for the three-stage classification, and an average of 43 % precision and 36 % recall in the five-stage classification. As can be seen in Table 5, the C4.5 and LTM classifiers perform less well.

Tables 5 and 6 show that the difficulty is to automatically discriminate between texts from neighboring stages (i.e. 1 from 2, 3 from 4, etc.). We believe that one reason is due to the fact that Direkt Profil 1.5.2 only analyzes a subset of the phenomena described in (Bartning and Schlyter, 2004). Consequently, the classifying algorithm can currently not be trained with the full range of developmental sequences. We are therefore developing an enhanced, more flexible parser, which will make more features detectable, and hopefully improve classification accuracy significantly. The improved parser is near completion, and further results are expected in 2006.

6. Conclusion

In this paper, we have presented a new CLC in French. The CEFLE corpus (*Corpus Écrit de Français Langue Étrangère*) contains written texts in French produced by adolescent Swedish learners of French. It also contains a control group with texts written by French adolescents on the same topics. We have developed an analyzer called Direkt Profil on the basis of this CLC. The analyzer carries out a sentence analysis of learner texts based on developmental sequences. In this paper we have presented two new features of Direkt Profil and evaluated them. The first one is the introduction of a chunking layer to our annotation. In this layer the system identifies four syntactic groups. The evaluation of this annotation is presented in Table 3.

```
Finiteness - inflected and uninflected verbs <= 5
|   Inflected verbs <= 4: 1 (10.0/2.0)
|   Inflected verbs > 4: 2 (2.0)
Finiteness - inflected and uninflected verbs > 5
|   Average sentence length <= 10
|   |   TTR <= 47
|   |   |   Verbs in the conditional <= 0
|   |   |   |   Percentage lexical present tense verbs with agreement <= 60: 2 (10.0)
|   |   |   |   Percentage lexical present tense verbs  with agreement > 60
|   |   |   |   |   Lexical verbs in present tense <= 2: 2 (6.0)
|   |   |   |   |   Lexical verbs in present tense > 2
|   |   |   |   |   |   Occurrences of the 1,000 most frequent words <= 589: 2 (6.0/2.0)
|   |   |   |   |   |   Occurrences of the 1,000 most frequent words > 589: 3 (9.0)
|   |   |   Verbs in the conditional > 0: 3 (3.0)
|   |   TTR > 47: 1 (3.0/1.0)
|   Average sentence length > 10
|   |   Occurrences of the next 2,000 words <= 33
|   |   |   Word count <= 344: 3 (8.0/1.0)
|   |   |   Word count > 344
|   |   |   |   Percentage inflected verbs <= 91: 3 (2.0/1.0)
|   |   |   |   Percentage inflected verbs > 91: 4 (14.0/2.0)
|   |   Occurrences of the next 2,000 words > 33
|   |   |   Percentage participles with stem error <= 14
|   |   |   |   Lexical verbs in the present tense <= 0: 4 (3.0/1.0)
|   |   |   |   Lexical verbs in the present tense > 0
|   |   |   |   |   Sentences without verbs <= 1
|   |   |   |   |   |   Average sentence length <= 13
|   |   |   |   |   |   |   Occurrences of the 1,000 most frequent words <= 654: 3 (2.0)
|   |   |   |   |   |   |   Occurrences of the 1,000 most frequent words > 654: 6 (2.0)
|   |   |   |   |   |   Average sentence length > 13: 6 (15.0)
|   |   |   |   |   Sentences without verbs > 1
|   |   |   |   |   |   Finiteness - inflected and uninflected verbs <= 16
|   |   |   |   |   |   |   Occurrences of non-dictionary words <= 334: 2 (3.0/1.0)
|   |   |   |   |   |   |   Occurrences of non-dictionary words > 334: 3 (2.0)
|   |   |   |   |   |   Finiteness - inflected and uninflected verbs > 16: 6 (2.0)
|   |   |   Percentage participles with stem error > 14: 2 (3.0/1.0)
```

Figure 1: A excerpt of the decision tree.

	C4.5		SVM		LMT	
Stage	Precision	Recall	Precision	Recall	Precision	Recall
1–2	0.63	0.62	0.67	0.77	0.68	0.69
3–4	0.54	0.57	0.76	0.66	0.70	0.64
6	0.62	0.59	0.91	0.91	0.76	0.86

Table 5: Results of the classification of texts into 3 stages for the three classifiers. Each classifier used 33 attributes and was trained on the *Voyage en Italie* corpus.

The second new feature is the introduction of a machine-learning module to optimize profiles, carry out parameter tuning and identify new features for profiling linguistic development on the basis of learner texts. We presented some initial results on classification using five different features. For a three stages classification the average precision and recall reaches 70%. As Direkt Profil continues to develop we expect the performance of the classifier system to increase considerably within the next couple of months.

7. Acknowledgments

The research presented here is supported by a grant from the Swedish Research Council, grant number 2004-1674 to the first author and by grants from the Elisabeth Rausing foundation for research in the Humanities and from Erik Philip-Sörenssens foundation for research.

8. References

Malin Ågren. 2005. Le marquage morphologique du nombre dans la phrase nominale. une étude sur l'acquisition du français L2 écrit. Technical report, Institut d'études romanes de Lund. Lund University.

Nathalie Bailey, Carolyn Madden, and Stephen Krashen. 1974. Is there a 'natural sequence' in adult second language learning. *Language Learning*, 24(2):235–243.

Inge Bartning and Suzanne Schlyter. 2004. Stades et itinéraires acquisitionnels des apprenants suédophones en français l2. *Journal of French Language Studies*, 14(3):281–299.

Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh. ACM.

Heidi C. Dulay and Marina K. Burt. 1974. Natural sequences in child second language acquisition. *Language Learning*, 24:37–53.

Véronique Gendner, Anne Vilnat, Laura Monceaux, Patrick Paroubek, and Isabelle Robba. 2004. Les annotations syntaxiques de référence peas. Technical report, LIMSI, Orsay. http://www.limsi.fr/Recherche/CORVAL/easy/PEAS_reference_annotations_v1.6.html.

Jonas Granfeldt, Pierre Nugues, Emil Persson, Lisa Persson, Fabian Kostadinov, Malin Ågren, and Suzanne

	C4.5		SVM		LMT	
Stage	Precision	Recall	Precision	Recall	Precision	Recall
1	0.50	0.40	0.57	0.40	0.44	0.40
2	0.37	0.38	0.47	0.62	0.45	0.48
3	0.23	0.25	0.43	0.36	0.46	0.39
4	0.47	0.44	0.67	0.63	0.56	0.56
6	0.62	0.59	0.91	0.91	0.76	0.86

Table 6: Results of the classification of texts into 5 stages for the three classifiers. Each classifier used 33 attributes and was trained on the *Voyage en Italie* corpus.

Schlyter. 2005. Direkt profil: A system for evaluating texts of second language learners of French based on developmental sequences. In *Proceedings of The Second Workshop on Building Educational Applications Using Natural Language Processing, 43rd Annual Meeting of the Association of Computational Linguistics*, pages 53–60, Ann Arbor, June 29.

Sylviane Granger. 2004. Practical applications of learner corpora. In Barbara Lewandowska-Tomaszczyk, editor, *Practical Applications in Language and Computers (PALC 2003)*, pages 291–301. Peter Lang, Frankfurt.

Thorsten Joachims. 1997. Text categorization with support vector machines: Learning with many relevant features. Technical Report LS-8 Report 23, Universität Dortmund.

Niels Landwehr, Mark Hall, and Eibe Frank. 2003. Logistic model trees. In Nada Lavrac, Dragan Gamberger, Ljupco Todorovski, and Hendrik Blockeel, editors, *Proceedings of the 14th European Conference on Machine Learning (ECML)*, volume 2837 of *Lecture Notes in Computer Science*, pages 241–252. Springer.

David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical Diversity and Language Development*. Palgrave MacMillan, Basingstoke.

Boris New, Christophe Pallier, Marc Brysbalert, and Ludovic Ferrand. 2004. Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, and Computers*, 36(3):516–524.

Manfred Pienemann. 1998. *Language Processing and Second Language Development*. Benjamins, Amsterdam.

John Ross Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

Larry Selinker. 1972. Interlanguage. *International Review of Applied Linguistics*, 10(3):209–231.

Jacques Vergne. 1999. *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur. Analyse syntaxique automatique non combinatoire. Synthèse et Résultats.* Habilitation à diriger des recherches, Université de Caen, 29 septembre.

TTR	47
Occurrences of the 1,000 most frequent words	589
Occurrences of the next 2,000 words	13
Occurrences of less frequent words	0
Occurrences of non-dictionary words	397
Conjunctions	7
Word count	136
Number of sentences	16
Average sentence length	8
Sentences without verb	1
Finiteness: total of inflected and uninflected verbs	4
Inflected verbs	3
Lexical verbs in the present tense	1
Verbs in the passé composé	1
Modal auxiliaries + infinitives	0
Verbs in imparfait	0
Être/avoir in imparfait	0
Lexical verbs imparfait	0
Lexical verbs imparfait with agreement	0
Verbs in the simple future	0
Lexical verbs in the simple future	0
Verbs in the pluperfect	0
Verbs in the conditional	0
Percentage inflected verbs	75
Percentage inflected verbs with agreement	33.33
Percentage sentences without verb	6.25
Percentage lexical present verbs with agreement	0
Percentage verbs in passé composé with agreement	0
Percentage participles with stem error	0
Percentage simple future being être/avoir	0
Percentage lexical simple future verbs with agreement	0
Percentage lexical conditional with agreement	0
Stage	1

Table 4: An example of feature vector.