Evaluating Stages of Development in Second Language French: A Machine-Learning Approach

Jonas Granfeldt
Centre for languages and literature
Lund university
Box 201, S-221 00 Lund, Sweden
jonas.granfeldt@rom.lu.se

Pierre Nugues
Department of Computer science
Lund Institute of Technology
Box 118, S-221 00 Lund, Sweden
Pierre.Nugues@cs.lth.se

Abstract

This paper describes a system to define and evaluate development stages in second language French. The identification of such stages can be formulated as determining the frequency of some lexical and grammatical features in the learners' production and how they vary over time. The problems in this procedure are threefold: identify the relevant features, decide on cutoff points for the stages, and evaluate the degree of success of the model.

The system addresses these three problems. It consists of a morphosyntactic analyzer called Direkt Profil and a machine-learning module connected to it. We first describe the usefulness and rationale behind its development. We then present the corpus we used to develop the analyzer. Finally, we present new and substantially improved results on training machine-learning classifiers compared to previous experiments (Granfeldt et al., 2006). We also introduce a method to select attributes in order to identify the most relevant grammatical features.

1 Introduction

Since the beginning of systematic research in second language acquisition (SLA) in the 1970s, one line of investigation was to identify and analyze stages of development that learners pass through when acquiring a second or a foreign language. See Sharwood-Smith and Truscott (2005) for a recent discussion. Stage identification can be applied to data from all linguistic levels, but it is perhaps most interesting for the development of morphology and syntax. Within SLA, the learner's internal grammar is considered as its own system, an interlanguage grammar, that develops and restructures over time (Selinker, 1972). The objective of this research is to determine and model the growth of the learner's grammar, where the identification of relevant grammatical features, the definition of development stages, and their evaluation are complex tasks requiring a systematic methodology (Ellis and Barkhuizen, 2005), pp. 97–98.

In this paper, we describe and evaluate a system that has fully automated this process. As possible applications for it, we can think of diagnostic tools for assessing language development and we hope that both learners and teachers will find it useful in this respect. However, we focus here on how our system, and more generally the methodology we propose, can assist researchers when working with grammatical stages. In order to understand its relevance, we begin with a simplified description of how stage identification is commonly carried out in the field of SLA.

2 Background

2.1 Current methodology for identifying stages of development

The first step to identify stages of development is to determine and extract grammatical features in the production (oral or written) of a representative population of learners. The selection of features can

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 73–80

be theoretically or empirically motivated, the crucial point being that the selected features have a content validity, i.e. that they are features whose realizations can translate a qualitative change in the learner's grammar. A second step is to understand and model the development of these features over time. Some linguistic features show a straightforward linear development, i.e. the scores for adequate use of the feature increases steadily with time at some observable rate. Other features show a nonlinear, sometimes U-shaped, development where the scores initially are high and then decrease in a second phase, only to regain a high level of correctness in a third phase.

Once the developmental trajectories are known, a third step is to decide on cutoff points in the data where the learner has reached a new stage of development. Most researchers work on several grammatical features at the same time, a procedure sometimes referred to as grammatical profiling. This means that the establishment of a stage of development has to take into consideration the analysis of a large number of categories.

2.2 Some problems with the current methodology

A necessary component in the method described above is an in-depth morphosyntactic analysis of the language samples produced by the learners. In our case, these are written texts but they might also be transcriptions of oral productions. Most analysts working with first and second language acquisition have now access to relatively large amounts of machine-readable data (large in SLA terms). It is also common for widespread languages, like English and French, to use tools such as morphological parsers and part-of-speech taggers (MacWhinney, 2000). These tools can considerably reduce the otherwise very time-consuming analysis step.

But even so, a lot of manual analysis is left to be done. First there is currently no reliable automated tool to parse learner's data although there have been some attempts for English (Sagae et al., 2005). For French, some of the linguistic structures and features used in grammatical profiling can be captured using available part-of-speech taggers and morphological parsers. But other more complex structures such as the agreement between constituents cannot. Another problem in grammatical profiling is that current tools usually work on one single feature at the time in a pipeline architecture, while one needs to analyze a large number of phenomena at the same time. A third problem concerns the artificiality in identifying stages (Ellis and Barkhuizen, 2005), p.98.

The result of the morphosyntactic analysis is typically a frequency analysis of certain features. For a particular linguistic phenomenon, say 3rd person agreement in the present tense, a typical procedure is to identify the different realizations of the phenomena and count them. The compiled data for all the features are then often inspected intuitively in order to identify suitable stages of development. In the SLA domain, there are currently multiple ways of dealing with this step and there has not been any principled evaluation of them. A possible reason for this is that there is currently no framework that has connected any sophisticated statistical treatment to the first two steps: the morphosyntactic analysis and the frequency count. If a fully automated processing pipeline were available, all steps in this tricky process could be evaluated more thoroughly.

We report here the current status of our system that aims at overcoming the methodological problems discussed above. The rest of the paper is organized as follows. We begin by summarizing briefly the previous work on the morphosyntactic development of second language French. Then we describe the corpus we are using to develop the analyzer to extract the grammatical features and constructions. The analyzer, called Direkt Profil, is also presented briefly. In the last sections, we discuss our machine-learning approach to identify the stages of development and select attributes and we present our current results.

3 Morphosyntactic development of second language French

Studies on the morphosyntactic development of second language French have to a large extent been empirically driven. One of their specific aims was the identification of a large number of developmentally related grammatical features and constructions along with hypotheses about their sequence of acquisition. The study by Bartning and Schlyter

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

(2004) is an example of it for spoken French, where the authors identified some 25 different morphosyntactic features and proposed a definition of their development over time in adult Swedish learners. Taken together, these features delineate six stages of development in the shape of grammatical profiles – ranging from beginners to very advanced learners. Examples of features are shown in Table 1. As the language learner moves towards an increasing automation of the target language, the produced structures become more frequent, more complex, and more appropriate. Developmental sequences describe this process in linguistic terms.

4 The CEFLE Corpus

To develop our analyzer (see Sect. 5) and to test the machine-learning approach to stages of development, we used the Lund CEFLE Corpus (*Corpus Écrit de Français Langue Étrangère*) (Ågren, 2005). CEFLE consists of texts in French as a foreign language written by 85 Swedish students with different levels of proficiency. It contains approximately 400 texts and 100,000 words. It also features a control group of 22 French native speakers. CEFLE was compiled throughout the academic year 2003/2004. During this period, each student wrote four or five texts in French at two months intervals. The aim of this study was to analyze the morphosyntactic development in written production.

For the present study, we used a random selection of 317 texts from the CEFLE corpus, see Table 2. A member of the team annotated one text from each learner using the criteria in Bartning and Schlyter (2004) and classified it according to the developmental stage the text was reflecting. For our current experiments (see below), we subsequently assigned the same classification to the three or four other texts of the same learner in the CEFLE corpus. The assumption behind the decision to propagate the stage of development from one annotated text to all the texts of the same learner is that a learner generally does not move up to the next stage during the short period under which the collection of the texts took place.

5 Direkt Profil

Direkt Profil (Granfeldt et al., 2005; Granfeldt et al., 2006) is a morphosyntactic analyzer designed for French as a second language. The initial aim was to implement the grammatical features and constructions in Table 1. In the current version of the system, a few features are still lacking but there is also a great number of additional ones that were not present from the beginning. The system has been presented in some detail in previous papers and we only give a brief description of the main parts.

Verb groups and noun groups represent the essential grammatical support of the profile classification. The majority of syntactic annotation standards for French take such groups into account in one way or another. However, in their present shape, these standards are insufficient to mark up constructions

CEFLE corpus			**Selection of CEFLE used** (averages)		
Task name	Elicitation type	Words		Text length	Sent. length
Homme	Pictures	17,260	Stage 1 (N=23)	78	6.9
Souvenir	Pers. Narrative	14,365	Stage 2 (N=98)	161	8.4
Italie	Pics	30,840	Stage 3 (N=97)	212	9.8
Moi	Pers. Narrative	30,355	Stage 4 (N=58)	320	11.6
Total		92,820	Control (N=41)	308	15.2

Table 2: General description of the CEFLE corpus and the selection used in the experiments reported in this paper.

of Table 1, 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 Direkt Profil, v. 2.1, detects three types of syntactic groups, nonrecursive noun groups, verb groups, prepositional groups, and conjunctions, that it annotates using the XML format.

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. 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 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. The performance of Direkt Profil version 1.5.2 was evaluated in Granfeldt et al. (2005). The results showed an overall F-measure of 0.83 (precision and recall).

6 A machine-learning approach to evaluate stages of development

The frequency count of the grammatical constructions and features form a basis to establish general stages of development. In our system, the frequency analysis is obtained automatically as the output from Direkt Profil.

One core problem in this last step of the procedure is that the data from the frequency analysis show a gradual increase that looks more like a development through continua than a development in discrete stages. Any definition of a stage will be to some extent arbitrary. Currently, there are a variety of methods that are used in field, but there is no principled way of evaluating these procedures. In the work of Bartning and Schlyter (2004), six stages of development were defined, five of which were subsequently identified by a human annotator in the CEFLE corpus. In the following section, we evaluate the probability of the existence of five different stages using machine-learning techniques.

6.1 First experiment: Classification analysis using all features

As experimental setup, we used the texts from each of the 85 learners that were manually assigned with their stage of development. The classification was done using the criteria in Table 1. Then we reused the same classification for the learner's three or sometimes four other texts in the CEFLE corpus, resulting in 276 classified texts. An additional 41 texts came from the control group of native speakers, resulting in a total of 317 classified texts.

We then used three machine-learning algorithms: the ID3/C4.5 algorithm (Quinlan, 1986), support vector machines (Boser et al., 1992), and logistic model trees (Landwehr et al., 2003). The training phase automatically induces classifiers from the selection of texts in the CEFLE corpus and the features we extract with the analyzer. We did all our exper-

	C4.5			SVM			LMT		
Stage	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
1–2	0.66	0.70	0.68	0.70	0.71	0.71	0.76	0.75	0.75
3–4	0.70	0.68	0.69	0.71	0.72	0.71	0.76	0.79	0.77
Control	0.71	0.66	0.68	0.70	0.63	0.67	0.89	0.83	0.86

Table 3: Results of the classification of texts into three stages for the three classifiers. Each classifier used 142 attributes and was trained on 317 texts from the CEFLE corpus.

iments with the Weka collection[1] of machine learning algorithms (Witten and Frank, 2005) and we evaluated them using the embedded 10-fold cross-validation.

We first clustered the five stages into three larger stages, where stages 1 and 2 together with stages 3 and 4 were into two stages and we trained the classifiers on them. We then ran a second evaluation with the original five stages. The results for the 317 texts and a feature vector consisting of 142 features are shown in Tables 4 and 5.

These results can be compared to those we obtained with a previous version of Direkt Profil (1.5.4) using a smaller number of features (33) and a smaller training corpus (80 texts). Those results (Granfeldt et al., 2006) showed that the best classifier at that point, SVM, obtained 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. The current results with more than 100 more features and a nearly four times bigger training corpus show an improvement of nearly 10 percentage points. The currently best classifying algorithm, LMT, obtains an average precision and recall of 79% for the three-stage classification (Table 3). For the five-stage classification, the improvement is even greater. LMT obtains 62% precision and 59% recall. In comparing the two best performing algorithms, SVM and LMT, one observation is that LMT outperforms SVM on the intermediate and advanced stages of development – 3, 4, and the control group of native speakers – but not on the first two stages of development. We have currently no explanation for this fact.

In conclusion of this first experiment, we can say that the increased number of attributes and the larger training corpus resulted in better overall performance for all three classifiers. But the improvement was not as great as we expected. We suspected that with the introduction of more than 100 new features compared to our previous experiments, we also introduced some irrelevant features for the classification. We ran an attribute selection procedure in order to identify the best features at this point. The results of this second experiment are presented in the next section.

[1] Available from: http://www.cs.waikato.ac.nz/ml/weka/.

6.2 Second experiment: Classification analysis using attribute selection

To evaluate the 142 attributes, we measured the information gain for each attribute with respect to the class. This method is derived from ID3 and is part of the Weka software. We used the ranker search method that ranks individual attributes according to their evaluation. Tables 5 and 6 show the results for the top 10 and top 20 attributes according to the information gain evaluation method.

In the next step, we ran a new classification experiment using the same three algorithms as in the first experiment and the same selection of 317 texts from the CEFLE corpus. We first evaluated the performance of the classifiers using the top 10 attributes. The results for the five-stage classification are shown in Table 7.

This experiment produced mixed results. On an average, the radical reduction of the number of attributes from 142 to 10 does not seem to affect the results very much. The average precision and recall figures for LMT are respectively 66% and 58%. This would suggest that there is a lot of noise in the remaining 132 attributes. On the other hand, the results for the lowest stage of development deteriorate. The SVM algorithm does not identify one single text as being on stage 1 using the top 10 attributes. This would suggest that within the remain-

	C4.5			SVM			LMT		
Stage	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
1	0.37	0.42	0.39	0.54	0.58	0.56	0.44	0.33	0.38
2	0.50	0.52	0.51	0.60	0.60	0.60	0.59	0.61	0.60
3	0.42	0.46	0.44	0.45	0.46	0.45	0.51	0.54	0.53
4	0.48	0.38	0.42	0.52	0.50	0.51	0.64	0.66	0.65
Control	0.71	0.66	0.68	0.70	0.63	0.67	0.89	0.83	0.86

Table 4: Results of the classification of texts into five stages for the three classifiers. Each classifier used 142 attributes and was trained on 317 texts from the CEFLE corpus.

Avg. merit	Avg. rank	Attribute name
0.405	1.4	Percentage of Determiner-Noun sequences with agreement (number and gender)
0.354	2.2	Percentage unknown words
0.33	3.2	Percentage NPs with gender agreement
0.313	3.9	Percentage prepositions (out of all parts-of-speech)
0.311	4.3	Average sentence length
0.208	6.2	Percentage Noun-Adjective sequences with agreement (number and gender)
0.198	7.4	Percentage subject-verb agreement with modals + infinitive
0.187	8.3	Percentage subject-verb agreement in passé composé structures
0.177	9.3	Percentage subject-verb agreement with être/avoir in 3rd person plural
0.176	9.8	Percentage subject-verb agreement with modal verbs and pronominal subjects

Table 5: The top 10 attributes. Attributes 1–10

Avg. merit	Avg. rank	Attribute name
0.168	11.4	Percentage verbs in present tense (out of all tenses)
0.165	11.8	Percentage verbs in Passé composé (out of all tenses)
0.15	14	Percentage subject-verb agreement with modal verbs (all subjects)
0.142	15.7	Percentage subject-verb agreement with modal verbs in sg
0.14	16.2	Percentage subject-verb agreement with modal verbs in present tense and 3rd person pronominal subject
0.136	16.7	Percentage finite lexical verbs in finite contexts
0.133	17.3	Percentage subject-verb agreement with finite lexical verbs
0.131	18.1	Percentage subject-verb agreement with sg pronominal subjects and modal verbs
0.125	19.3	Percentage subject-verb with lexical verbs in 3rd person plural
0.116	21.4	Percentage subject-verb with pronominal subjects and être/avoir

Table 6: The 10 next attributes. Attributes 11–20

	C4.5			SVM			LMT		
Stage	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
1	0.46	0.46	0.46	0.00	0.00	0.00	0.78	0.29	0.42
2	0.50	0.49	0.49	0.53	0.72	0.61	0.57	0.70	0.63
3	0.43	0.42	0.43	0.50	0.43	0.46	0.55	0.49	0.52
4	0.50	0.57	0.53	0.62	0.71	0.66	0.63	0.64	0.63
Control	0.84	0.76	0.79	0.94	0.76	0.84	0.78	0.78	0.78

Table 7: Results of the classification of texts into five stages for the three classifiers. Each classifier used the top 10 attributes evaluated with the InfoGain method in Weka and was trained on 317 texts from the CEFLE corpus.

ing 132 attributes there are some attributes that are very important for identifying texts in stage 1. In our next evaluation, we therefore included the next 10 attributes in the attribute ranking (attributes 11–20) resulting in a feature vector of 20 attributes. The results for a five-stages classifications are shown in Table 8

Arguably, the overall results are better but the average for LMT actually shows a slight decrease compared to the previous experiment with only the top 10 features. All three classifiers identify texts on stage 1 using a feature vector with the top 20 features. We also note a difference in precision and recall figures for stage 1 using the ranked attributes. While these figures were relatively close in the first experiment using all 142 attributes (see Table 3 and Table 4), they are wide apart in the two following experiments (with recall figures being considerably lower than precision figures). This means that the recall quality depends on a much larger set of attributes for the lowest stage of development than for the other stages. Since the precision and recall figures for the other stages are close throughout, this could in turn mean that the stage 1 is the most heterogeneous stage.

7 Conclusion and future work

There is an ongoing discussion in the field of second language acquisition on the existence of discrete "stages" and how to define them, see for instance Sharwood-Smith and Truscott (2005). We believe that language development is systematic but always gradual if one looks close enough at the data. Our view is that developmental stages should reflect this property.

In this paper, we have presented and evaluated a system that can assist researchers in working with stages of development in second language French. The system consists of a morphosyntactic analyzer called Direkt Profil and a machine-learning module connected to it. A set of 317 texts from the CEFLE corpus was classified according to the stage of development they were reflecting. In classifying the texts, we built on previous research on morphosyntactic development in French second language. We extracted vectors of 142 features from the texts using the morphosyntactic analyzer we constructed. We then trained three different classifiers to evaluate the hypothesis that there were five stages of development represented in the material.

The results from a first classification experiment using a feature vector containing all the 142 features showed a substantial improvement of more than 10 percentage points compared to our previous results. For a three-stage classification, the average precision and recall figure for the system is now 79%. In trying to identify the most relevant features for classification, we used an attribute selection method based on the information gain and we identified two sets of top ranked attributes: the top ten attributes and the top twenty attributes. The results showed that while the overall performance was surprisingly not affected by the radical reduction of the number of attributes (from 142 to 10 and 20 respectively), the results for the lowest stage of development were affected very negatively. One conclusion at this point is that the stage 1 texts are very heterogeneous constructs to the point that it has to be questioned if they have an independent status.

From the results on the morphosyntactic analysis

	C4.5			SVM			LMT		
Stage	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
1	0.56	0.38	0.45	0.60	0.38	0.46	0.53	0.38	0.44
2	0.51	0.53	0.52	0.61	0.62	0.62	0.61	0.61	0.61
3	0.49	0.47	0.48	0.54	0.57	0.56	0.56	0.59	0.57
4	0.45	0.55	0.50	0.61	0.69	0.65	0.61	0.62	0.62
Control	0.78	0.68	0.73	0.83	0.73	0.78	0.86	0.88	0.87

Table 8: Results of the classification of texts into five stages for the three classifiers. Each classifier used the top 20 attributes evaluated with the InfoGain method in Weka and was trained on 317 texts from the CEFLE corpus.

it is clear that there is room for improvement. We are currently looking into the possibility of using a statistical POS tagger and a chunker trained on an annotated corpus of native French. The preliminary results are encouraging despite the very different kinds of data (native and nonnative French).

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.

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.

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.

Rod Ellis and Gary Barkhuizen. 2005. *Analysing learner language*. Oxford University Press, Oxford.

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

Jonas Granfeldt, Pierre Nugues, Malin Ågren, Jonas Thulin, Emil Persson, and Suzanne Schlyter. 2006. CEFLE and Direkt Profil: A new computer learner corpus in French L2 and a system for grammatical profiling. In *Proceedings of the 5th International Conference on Language Ressources and Evaluation*, pages 565–570, Genoa, Italy, 22-28 May.

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.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum, Mahwah, New Jersey.

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

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics 2005*, pages 197–2004, Ann Arbor, USA, June.

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

Michael Sharwood-Smith and John Truscott. 2005. Stages or continua in second language acquisition: A MOGUL solution. *Applied Linguistics*, 26(2):219–240.

Ian H. Witten and Eibe Frank. 2005. *Data Mining. Practical Machine Learning Tools and Techniques*. Elsevier, Amsterdam.