DOI 10.1007/s10958-016-2813-2
Journal of Mathematical Sciences, Vol. 214, No. 6, May, 2016

A STRUCTURAL PATTERN BASED METHOD FOR AUTOMATED MORPHOLOGICAL ANALYSIS OF WORD FORMS IN A NATURAL LANGUAGE

E. Egorova, A. Lavrentiev, and A. Chepovskiy UDC 519.76+81'32

Abstract. In this paper, a computerized model for morphological analysis of languages with word formation based on affixation processes is proposed. The main idea consists in defining structural patterns of words and corresponding lists of suffixes. First, a detailed description of a stemming algorithm, its modification, and the technique of determining grammatical characteristics of word forms are given. The next part of this work focuses on the application of the proposed algorithms for the French language. Finally, some results of execution of these algorithms are provided.

1. Introduction

One of the main problems in the field of information retrieval systems is to improve the quality and exhaustiveness of search results, while optimizing the size of the indexes of such systems. The basic algorithm that allows indexing and subsequent search for information in information retrieval systems is the one that isolates the stems of word forms. This brings about the problem of designing a software implementing models and algorithms for analyzing the words from natural languages capable of operating with limited or no morphological dictionaries at all.

There is a great number of works dedicated to the problem of automatic morphological analysis of word forms of natural language (see [1] for a detailed review). In this paper, we will focus on the construction of an algorithm for morphological analysis of word forms in languages with extensive affix derivation and inflection. In such languages, the word forms have a relatively complex morphological structure, since word inflection and derivation are expressed by adding a series of affixes to the root of the word. However, this structural complexity of words does not complicate the analysis by any means. On the contrary, it allows identifying more accurately the stem of a word form and determining its grammatical characteristics, as each affix carries a certain derivational or grammatical meaning.

At present, languages of this type are most commonly processed with algorithms based on a simple stripping of affixes. However, this approach has a certain number of drawbacks, one of which is that it does not take into account derivational and inflectional rules of the language. The first attempt to address this drawback was the morphological model described in [3]. In this paper, we formulate a generalized model of structural patterns and consider its application to the French language. The main idea of this model is to build structural patterns of word endings that describe all possible affix sequences in the word forms of a given language. This approach takes into account the grammatical features of the language and makes it possible to process new words or specific vocabulary and, as a consequence, gives more accurate results in the identification of pseudo-stems. In addition, the work addresses the issue of determining the grammatical characteristics. By identifying meaningful parts of an ending, the technique of structural patterns allows determining the grammatical characteristics of words. Finally, we use the example of French as a language with considerable affix derivation and inflection to illustrate the process of building the necessary patterns for the proposed approach to morphological analysis.

Translated from Fundamentalnaya i Prikladnaya Matematika, Vol. 19, No. 3, pp. 91–109, 2014.

2. Structural Pattern Method

Consider languages with extensive affix derivation and inflection. In such languages, word forms consist of roots (or stems) carrying the basic lexical meaning, and a number of affixes, each of which expresses one or several grammatical or derivational values. Hereafter we will use the terms "root" and "stem" as synonyms, since our aim is not a perfect morphemic analysis in terms of linguistic theory but identifying word forms with similar meaning. Affixes are divided into several types, depending on their position relative to the root of the word. Prefixes are placed before and suffixes after the root. In linguistics, derivational affixes are usually distinguished from inflectional ones. The former are used in order to modify the lexical meaning or the grammatical class (part of speech) of a word, while the latter just express grammatical values (such as a noun case or a verb tense). In this paper, however, this difference is not significant, since our main goal is to isolate the root of the word. For convenience, we will use the term "ending" to designate the entire set of derivational and inflectional suffixes.

Consider several types of word patterns. Hereinafter, the part of the word that is not an ending or a prefix will be referred to as the stem. The following three cases of word structure are possible: prefix + stem + ending, stem + ending, prefix + stem. Therefore, a word form consists of a stem, and a certain set of affixes, which in some cases may be empty. Now let us provide, in terms of the theory of formal languages, a more formal description of the words of natural language that have the aforementioned features, and the way these words are inflected. Assume that there is given an alphabet $\Sigma = \{\sigma_1, \sigma_2, \ldots\}$, where σ_i is a letter.

Definition 2.1. Let a *word* be a finite sequence $\sigma = \sigma_{i_1}\sigma_{i_2}\ldots\sigma_{i_n}$, where $\sigma_{i_j} \in \Sigma$. The empty word is also allowed, we denote it as ε.

Definition 2.2. Let Σ^* be the set of all possible words over the alphabet Σ.

Let us define the structure of a word form. For this purpose, we introduce the following notation. Let r be the stem as type and R be the set of words of this type. Also consider the set A of types of affixes that are used to form an ending. Let m be the number of different types; then $A = \{a_1, a_2, \ldots, a_m\}$. For $i = 1, \ldots, m$ we associate each type a_i with a set A_i of its representatives (words). Let p be the type of prefixes and P be the set of its representatives. It is important to note that there are words whose type cannot be determined unambiguously, i.e., $P \cap A_1 \cap A_2 \cap \cdots \cap A_m \neq \varnothing$. Let a_0 denote the type of words that do not belong to the set $P \cup A_1 \cup \cdots \cup A_m \cup R$.

Now let us consider the structure of the affix part of a word in more detail. As mentioned previously, it may consist either of a prefix and an ending, or of an ending only, or of a prefix only. In the context of this model, the prefix is expressed as a single affix or as the empty word, if it is absent. The structure of the ending is more complicated. Every ending is a word, which may be empty. A nonempty ending can be represented as a sequence of subwords, each of these subwords being an affix. Thus, for each word it is possible to obtain a sequence of affix types, which will describe its ending and be an ordered vector. In general, such a sequence looks as follows: $(a_{i_1}, a_{i_2}, \ldots, a_{i_s})$, $a_{i_j} \in A$. To describe the affix part of a word as a whole, it is necessary to account for its prefix part; therefore, a p should be added at the beginning of this sequence in the presence of a prefix. We obtain the following sequence: $(p, a_{i_1}, a_{i_2}, \ldots, a_{i_s})$, $a_{i_j} \in A$. Given the characteristics of the languages under consideration, it can be argued that the number of such sequences is finite. Let Sch denote the set of all sequences of affix types admissible for a given language that can describe the affix part of the word. Let l denote the maximum length of the sequence describing an ending. Subsequently, in the general case Sch can be written as follows:

$$\text{Sch} := \{(a_{i_1}, a_{i_2}, \ldots, a_{i_s}) \mid a_{i_j} \in A,\ s = 0, \ldots, l\} \cup \{(p, a_{i_1}, a_{i_2}, \ldots, a_{i_s}) \mid a_{i_j} \in A,\ s = 1, \ldots, l\} \cup \{(p)\}. \tag{2.1}$$

It can be seen that such a set contains sequences that describe all possible representations of the affix part of the word: a prefix, an ending or both a prefix and an ending. The elements of the set will be called the descriptions or structures of the affix part of the word.

Now we describe a function that will form the basis for a future algorithm, i.e., the typing function of subword sequences.

Definition 2.3. A *typing function* is a function F such that

$$F\colon (\gamma_1, \gamma_2, \ldots, \gamma_k) \mapsto \{(c_1, c_2, \ldots, c_k) \mid c_i \in A\} \cup \{p, a_0\},\ i = 1, \ldots, k\},$$

where for $j = 1, \ldots, k$, $\gamma_j \in \Sigma^*$, and c_j is one of its possible types.

In other words, the typing function of each subword sequence associates a set of possible sequences of types of these subwords. The multiplicity of these sequences is due to the fact that one and the same affix can be characterized by several types. Given these concepts, we now turn to a more formal description of the affix word part.

Assertion 2.1. A word $\sigma \in \Sigma^*$ is an ending if and only if it can be decomposed into subwords $\sigma = \sigma_1\sigma_2 \ldots \sigma_k$, $\sigma_i \in \Sigma^*$, $i = 1, \ldots, k$, in such a way that $F[(\sigma_1, \sigma_2, \ldots, \sigma_k)] \cap \text{Sch} \neq \varnothing$.

In other words, the fact that a word is an ending is equivalent to the fact that it can be decomposed into subwords in such a way that the corresponding sequence of types is acceptable, i.e., belongs to the set Sch (2.1). Similar statements can be formulated for the case of prefixes and prefixes + endings.

Assertion 2.2. A word $\sigma \in \Sigma^*$ is a prefix if and only if $F[(\sigma)] = \{p\}$.

Assertion 2.3. A word $\sigma \in \Sigma^*$ has the structure prefix + ending if and only if it can be decomposed into subwords $\sigma = \sigma_1\sigma_2 \ldots \sigma_k$, $\sigma_i \in \Sigma^*$, $i = 1, \ldots, k$, in such a way that $F[\sigma_1] = \{p\}$ and $F[(\sigma_1, \sigma_2, \ldots, \sigma_k)] \cap \text{Sch} \neq \varnothing$.

Note that because of the ambiguity of word decomposition and because of the fact that some words belong to several types, one and the same ending may correspond to multiple descriptions. Under these definitions, we shall now describe the concept of a word-form.

Definition 2.4. A word form is a word σ over the alphabet Σ that can be represented in the following form:

$$\sigma = \sigma^p\sigma^r\sigma^d,$$

where $\sigma^p \in P$, $\sigma^r \in R$, and σ^d is an ending.

Note again that the prefix and the ending may be empty symbols, so this definition describes all possible structures of words that are valid in the model. Hereinafter, the word form will be our main object of study.

As we mentioned earlier, each word form can be represented as a sequence of subwords, each of them belonging to a particular type. Since we are interested in the analysis of word forms without a dictionary, we do not have the set R, so the main task is to identify the prefix and the ending of a word form correctly, which will in turn allow determining the grammatically correct stem. The algorithm described below takes as input a word form, and returns the set of its admissible affix parts denoting different variants of word form structural patterns representation.

Algorithm 2.1.

Input: $\sigma = \sigma_{i_1} \ldots \sigma_{i_l}$, $\sigma_{i_j} \in \Sigma$, $|\sigma| = l$ is a word form.

Output: T, the set of admissible affix parts of the word form σ.

Step 1. $T := \{\varepsilon\}$.

Step 2. We obtain the set M consisting of all subword sequences by separating the prefix and decomposing the ending in such a way that every subword is contained in the set $A_1 \cup \cdots \cup A_m \cup P$.

Step 3. For each decomposition $(\sigma_1^d, \sigma_2^d, \ldots, \sigma_k^d) \in M$, $\sigma_i^d \in \Sigma^*$ apply the function F. Set $:= F[(\sigma_1^d, \sigma_2^d, \ldots, \sigma_k^d)]$.

Step 4. If $\text{Set} \cap \text{Sch} \neq \varnothing$, then $T := T \cup \{\sigma^d = \sigma_1^d\sigma_2^d \ldots \sigma_k^d\}$.

To put it simply, the first step is to get all the possible decompositions of a word form, i.e., decompositions such that each of their components is an affix permitted in the given language. After that we apply the typing function to them in order to understand what the sequence of decomposition subwords looks like. In the final step, it is determined whether this sequence is possible within the studied language. As a result, the algorithm for a given word form allows determining its possible decompositions into a prefix, a stem, and an ending, which are consistent with the grammar of the given language. It may happen that at the end of the algorithm the set T will consist only of the empty word ε. This would mean that the whole word form is a stem and contains no prefix or ending. For each language under consideration, a number of its specific features should be taken into consideration. These include possible length of endings, the length of affixes, the letters that occur at the beginning of affixes, etc.

3. Grammatical Features in the Method of Structural Patterns

The algorithm described above makes it possible for each individual word form to determine the set of possible representations of its affix part. Each element of this set is one of the sequences of subwords permitted in a given language, hereinafter called affixes. Each affix carries some information on grammatical, derivational or inflectional features of the word form. Thus, analyzing each affix individually as well as the affix part of the word as a whole, we can determine the grammatical features of a given word form. Consider an algorithm based on the results of Algorithm 2.1, which allows recovering the grammatical features of word forms. To begin with, one needs to define more clearly what grammatical features exist, how they relate to each other, and how they differ. We distinguish two levels of grammatical features.

Definition 3.1. A grammatical feature of the first level is a feature that concerns all affixes, and that uniquely identifies the set of other possible features of a given word form.

Example 3.1. Consider the Russian language where the "part of speech" feature belongs to the first level, as for each of the affixes one can understand what part of speech it characterizes. Also, each part of speech (verb, noun, adjective, etc.) unambiguously determines the set of other properties that make sense and that can be associated with a particular part of speech. For instance, for a verb (unlike for a noun) one can determine the tense, mood, etc. For a noun, one can determine the case, which, in turn, cannot be done for a verb.

We further assume that there can be several grammatical features of the first level in a given language, and each affix will have them all. Let us now define the grammatical features of the second level.

Definition 3.2. A grammatical feature of the second level is a feature that is proper to some affix, but not necessarily every affix of a word form. Each affix may have several different grammatical features of the second level.

Example 3.2. In Russian, if the first-level grammatical feature "part of speech" takes the value "noun," the following features of the second level can be determined: gender, number, case, etc.

Now we turn to a more formal description of the concepts introduced. In this model, each feature of the first or the second level has a few "representatives," that is, for example, in the Russian language the representatives of the first-level "part of speech" feature are: verb, noun, adjective, etc., and the representatives of the second-level "gender" feature are: masculine, feminine, neutral. Thus, there is a finite number of features, and for each of them a finite set of its representatives can be defined.

Suppose that for a given language there are N different features of the first level. They will be denoted as $C_1, C_2, \ldots, C_N$. Each of the features is associated with the set of its representatives, i.e., for instance, a C_i feature will be associated with a set F_i of its representatives. More formally, this is formulated as follows: $F_i := \{f_{i1}, f_{i2}, \ldots, f_{ik}\}$, where f_{ij} is the jth representative of the ith feature, and k indicates the number of representatives for a given feature. For each first-level feature, the number of representatives may vary, but we will not describe this explicitly, since it is not necessary. For each feature, we denote

the number of its representatives as the cardinality of the corresponding set, i.e., for a C_i feature, the number of representatives is $|F_i|$.

From the definition of the first-level feature, it follows that for each affix it is possible to determine which representative describes it in each feature. All possible combinations of representatives of the features can be described as the Cartesian product of the sets $F_1, \ldots, F_N$, i.e.,

$$F_1 \times F_2 \times \cdots \times F_N = \{(f_1, f_2, \ldots, f_N) \mid f_i \in F_i\}.$$

Thus, we can associate each affix with a vector in the set $F_1 \times F_2 \times \cdots \times F_N$. Therefore, let us make one more definition.

Definition 3.3. The characteristic profile (or profile) of an affix will be called a vector from the set $F_1 \times F_2 \times \cdots \times F_N$ such that the affix is characterized by all members of the vector (feature representatives).

Since one affix can be described by several representatives of the first level features, we find that each affix may correspond to several characteristic profiles.

The number of all possible characteristic profiles is

$$n = |F_1| \cdot |F_2| \cdot \ldots \cdot |F_N|.$$

All these profiles can be ordered, and a vector v of length n can be created in such a way that each of its coordinates will correspond to one of the profiles. Now one can map each affix to a v vector, where units will only be placed there, where they correspond to the profile that describes this affix. Simply put, we make a vector that indicates the profiles corresponding to a given affix. In general, this model is quite complicated, but for the languages under consideration the number of characteristics of the first level is limited, and the situation is considerably simplified.

The situation is less complicated for the second-level features. Suppose there are different features of the second level, we denote them by $D_1, \ldots, D_M$. We associate each of the features with a set of its representatives, i.e., for a D_i second-level feature, the set of its representatives will be denoted as $S_i = \{s_{i1}, \ldots, s_{ik}\}$, where s_{ij} is the jth representative of the ith feature. The number of representatives of each second level feature will also be denoted by the cardinality of the set, i.e., the number of representatives of the feature is equal to $|S_i|$.

At the next stage, we should distinguish between two cases, which will condition the further construction of the model.

(1) The value of all the second-level features is uniquely determined regardless of the profile.

(2) The value of some second-level features directly depends on the profile.

Consider the first case. Here, the value of the second-level features does not depend on the profile, so it is logical to assign to each affix a vector of zeros and ones, the same way as was done for the first-level features. The length of this vector is equal to $m = |S_1| + |S_2| + \cdots + |S_M|$, then the value of the vector coordinate is equal to one if the feature corresponding to this coordinate is determined by the given affix, and to zero otherwise. It should be noted that representatives of one feature are not mutually exclusive. Each profile uniquely defines a set of second-level features for a given word form. For each profile, one can specify a set of numbers of the coordinates of the second-level features, and these coordinates will only match the features that are relevant for the given profile (for example, in Russian, the conjugation class is irrelevant for nouns). Now we proceed directly to the algorithm for recovering grammatical characteristics of a word form by its ending.

After applying Algorithm 2.1, we can obtain for a given word form a set of its potential affix parts and their decompositions into individual affixes. Now let us describe step by step the algorithm to restore the grammatical features of a word form by analyzing these parts.

Algorithm 3.1.

Input: A word form $\sigma = \sigma_{i_1} \ldots \sigma_{i_l}$, $\sigma_{i_j} \in \Sigma$, $|\sigma| = l$, and a set T of its permissible affix parts obtained from Algorithm 2.1.

Output: A set of first- and second-level features for each of the possible affix parts.

Step 1. Calculation of the characteristic profile of each affix. In fact, for each affix of the nonroot part of a word form we know which representatives of the first-level features correspond to it, and, therefore, we also know what the corresponding profile vector looks like.

Step 2. Calculation of the characteristic profile of the whole word form. For this purpose, we need to do a componentwise logical multiplication of these vectors. Simply put, a word form is characterized by an i profile if all profile vectors display 1 at the ith place. It is quite possible that a word form can be characterized by several profiles. It is also possible that none of the profiles is selected, which indicates that the corresponding affix decomposition is not correct and should not be analyzed any further.

Step 3. Calculation of vectors for the second-level features of each affix. At this stage, we know the set of possible profiles for a word form. For each profile, we know the numbers of the second-level feature vector coordinates whose value we need to know. Also, for each affix we have a vector of zeros and ones, reflecting the values of its second-level features. Now we need to take from this vector only the coordinates that are necessary for the profile, and make new vectors out of them.

Step 4. Calculation of the second level features for the whole word form. Now we make a componentwise logical addition of these vectors. Representatives of the characteristics whose positions are filled with 1 in the resulting vector are considered to belong to the word form under consideration. In other words, if in at least one of the vectors the ith position is filled with 1, then we say that the representative of a second-level feature corresponding to this coordinate describes this word form

Step 5. The steps 1–4 should be applied for each possible affix part of the word form.

This algorithm has been described for the case where the value of all the second-level features is uniquely determined, regardless of the profile. Now consider the case where there are features that depend directly on the profile. It turns out that in this case it is incorrect to assign to each affix one (and the same for all profiles) vector of the second-level features. Each affix for each profile should be assigned its proper vector of the second-level features, and only those that are relevant for this profile. This saves us from having to store for each profile the set of numbers of coordinates of the relevant features. In general, the algorithm remains the same. At the first step, in a similar way we define the profile, and depending on it, we select the appropriate second-level feature vector for each affix, and perform a logical addition. As a result, a word form will have those features of the second level whose positions were filled with 1 in the total vector.

Modification of the Pseudo-Stem Identification Algorithm. Here we consider a modification of the basic algorithm (Algorithm 2.1) for the identification of pseudo-stems, using the method of structural patterns. We will consider languages that commonly use derivational suffixes to produce new words with a similar sense but *a different part of speech*. Accordingly, we will propose an algorithm for identifying the pseudo-stems of word forms with a certain part of speech known in advance. Note also that in this case the prefix part will not be considered, it will be (conventionally) included into the stem, as the primary task is to strip the ending correctly.

Now let us define the introduced conditions more formally. We will consider languages where only one feature of the first level is given, i.e., $C_1 =$ "part of speech". The set of representatives of this feature is finite, we denote it as follows: $P = \{p_1, p_2, \ldots, p_n\}$, where p_i is a part of speech. Also given is the set of types of suffixes, we denote it (as earlier) as $A = \{a_1, a_2, \ldots, a_m\}$, where a_i is a type of suffix. A_i denotes the set of representatives of each of the types a_i, $i = 1, \ldots, m$. In addition, we have the set of all possible structural patterns for the given language. Recall that the set Sch consists of ordered vectors, and the types of affixes are elements of these vectors. We proceed from the assumption that all of the structural patterns for such languages can be divided into groups depending on parts of speech. Then the set Sch can be represented as $\mathrm{Sch} = \mathrm{Sch}_{p_1} \cup \cdots \cup \mathrm{Sch}_{p_n}$, where Sch_{p_i} is the set of patterns responsible for the part of speech p_i. It is worth noting that $\mathrm{Sch}_{p_1} \cap \cdots \cap \mathrm{Sch}_{p_n} = \{(\varepsilon)\}$, i.e., the only pattern that is the same

for different parts of speech is the primitive one, i.e., the one that defines the words without expressed endings.

Using this notation, we will now describe a modification of the basic algorithm. The input, in addition to the word σ, now includes its part of speech p_i. Then it is necessary to perform a similar analysis as before, but replacing the set Sch by its subset Sch_{p_i}. In the output, we obtain a set of different possible endings and their decompositions according to the patterns, but in this case they satisfy the structure of words of a given part of speech. Such modification would reduce the number of possible variant endings and, therefore, improve the accuracy of analysis.

The algorithm for identification of the pseudo-stem also needs certain modifications. Given the constraints introduced, this algorithm can be implemented much easier. If we know in advance the part of speech of the word form, then the number of corresponding descriptions is reduced dramatically. We see that after we learn the part of speech of a word form, it remains to verify whether it contains derivational suffixes arranged in one of the possible orders. This can be done by analyzing the ending of the word form, i.e., we need a function that would identify at the end of the word one of the possible suffixes corresponding to the specified type. To do this, we introduce the function R, a word and one suffix types will be its input, and the output will be a set of subwords obtained after cutting all sorts of suffixes of the specified type off the word. Formally,

$$R(\langle \sigma, a_i \rangle) = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}, \quad a_i \in A, \quad \sigma \text{ is a word}, \quad \sigma_i \text{ is a subword},$$

if nothing can be cut off, the function returns the empty set.

Algorithm 3.2.

Given: a word σ, its part of speech p_l, all possible patterns for that part of speech Sch_{p_l}. All of the following steps should be used for each pattern s from Sch_{p_l}.

Step 1. Compute the set $M = R(\langle \sigma, a_{|s|} \rangle)$, where $a_{|s|}$ is the last element in the pattern s.

Step 2. Repeat Step 1 for all subwords of the set M, but instead of $a_{|s|}$ take $a_{|s|-1}$, i.e., the following (backward) affix type in the pattern s.

Step 3. Repeat Step 2, using the function R for all the words obtained by the previous step, but with a change of the affix type: every time take the following backward. Perform this until the whole vector s is processed or until R returns the empty set.

Result: all possible stems returned by the function R, stopping at the end of the vector s. If in the end we get the empty set, then return the original word σ as a stem.

In general, this kind of algorithm is suitable for all cases where the patterns can be classified according to one of the features.

4. Computer Analysis of Word Forms in the French Language

Now let us describe some of the basic principles of building structural patterns for word forms in the French language. Emphasis is made on morphological and derivational features of the French language, as they allow the correct morphological analysis of word forms.

Consider the suffix method of derivation. According to [5], nine different parts of speech are distinguished in French: noun (nom), adjective (adjectif), verb (verbe), adverb (adverbe), article (article), pronoun (pronom), preposition (préposition), conjunction (conjonction), and interjection (interjection). The suffix derivation method is most often used to form nouns, adjectives, verbs and adverbs. For the rest of this article we will focus on these four parts of speech.

Suffixes are divided into derivational and inflectional types. For adjectives, verbs, and certain nouns, inflectional suffixes are used to modify the word's gender (masculine/feminine) and number (singular/plural). We will not distinguish a separate type for such suffixes. In verb conjugation, person, number, tense, gender, mood, and voice are affected. The suffixes expressing these features are also inflectional. We use v to denote their type. Derivational suffixes are used to form new words. With these, words can change their part of speech or take a certain connotation (diminutive, multiplicity, etc.).

The process of building the structures of endings is based on the grammar of the French language [2, 4–6]. For the French language, the most convenient way to do this is to analyze the parts of speech one by one. That is why in order to create our model we analyzed and built structural patterns for the four basic parts of speech: verbs, nouns, adjectives, and adverbs.

Verb. In French conjugation, verb forms change endings depending on person, number, gender, tense, mood, and voice. There are three formal conjugation classes or groups. The verbs of the first and second groups are called "regular," because they have certain rules of conjugation, i.e., for them the ending is uniquely determined given the values of the grammatical features. The verbs of the third group are called "irregular," since their conjugation may affect the form of their stem and they do not have uniform rules for the formation of ending. All possible endings of infinitives and the forms that the verbs take in conjugation will be considered in the suffixes corresponding to the type v. This way we obtain the first possible pattern describing the verb endings: stem-v. For example, a first group verb "marcher" (walk) can be represented as "march-er". In conjugation, the same verb can take such forms as "march-ions" ([we] were walking), "march-erai" ([I] will walk), etc.

In addition to simple infinitive suffixes "-er," "-ir," etc., French has some more complex suffixes, such as "-ifier" and "-iser." They are used to make verbs from nouns (common and proper), adjectives, and abbreviations. We will treat these suffixes as complex, i.e., split them into two parts: "-ifi-er" and "-is-er." This is explained by the fact that in conjugation, only the final part ("-er") will vary and by the fact that the segments "-ifi-" and "-is-" are included in derived new words, whereas "-er" is not. Some French suffixes do not modify the part of speech, but just add a certain shade of meaning. These include the suffixes "-aill-," "-ass-," "-ill-," "-och-," "-onn-," "-ot-," "-ouill-," "-ard-," "-âtr-," "-et-," "-in-." They will also be analyzed as a complex; e.g., "chant-er"–"chant-onn-er" (to sing–to sing to oneself), "touss-er"–"touss-ot-er" (to cough–to cough slightly), etc. Thus, we can introduce another type of suffix, which will account for one of the parts in the complex verb endings. Now we can propose another pattern to describe the verb endings: stem-fc-v. In French, it happens that the word to which a verb suffix is added, is itself formed by suffixation. In other words, a derivative may become a producing word. It is worth distinguishing between two cases: the formation of verbs on the basis of nouns and on the basis of adjectives. In this regard, we propose two more patterns to describe verb endings: stem-n-v , where n denotes nominal derivational suffixes that produce nouns that become in turn the basis for verbs; stem-a-c-v, where a denotes adjectival derivational suffixes that are used in turn to produce verbs.

Adverb. Most of the French adverbs are derived from adjectives by means of adding a "-ment" suffix. For that reason, the ending of adverbs is usually composed of a suffix producing an adjective followed by the adverbial suffix. The "-ment" ending can take several variant forms depending on which kind of adjective it is added to: "-ement," "ément," "-amment," "-emment." If the ending of the adjective is not expressed explicitly, the adverb can be described by the following pattern: stem-d , where d is a type of suffix used to produce an adverb.

For cases where the ending of adjectives is expressed explicitly, as, for instance, in the word "alphabét-ique-ment" (alphabetically), where the suffix "-ique" transforms a noun into an adjective, and "-ment" produces an adverb out of an adjective, we propose the following pattern: stem-a-d. Some adjectives in French were formed from nouns that have their own suffixes. The corresponding scheme is stem-n-a-d. There are also some adverbs which were not produced by suffixation (or the trace of suffixation is no longer visible). These adverbs include such words as "droit" (right), "bien" (well), "mal" (badly), etc. For them, the following description can be proposed: stem-ε, where ε is the empty symbol, in other words, the pattern is composed of a stem only.

Noun. French nouns can inflect in number and sometimes in gender. The gender varies only in nouns denoting professions or other terms relating directly to a person or an animal. This variation is operated by replacing the final suffix or by adding a new one. The majority of nouns have a constant gender that can not be changed. The plural is formed by adding a special suffix to the word, which depends on what

letter the word ends with. All these peculiarities of noun inflection concern their grammatical features and will continue to be taken into account in the pattern models using the suffix type nt.

In French, there are many nouns that do not have any explicit ending mark, e.g., "lac" (lake), "mur" (wall), "sol" (ground), "tome" (tome), "foudre" (thunder). Such words may form the stem of derived words, but they cannot be divided into meaningful parts themselves. Hence we obtain the following possible description: stem-ε. Another large group is formed of nouns with a simple (indivisible) ending, e.g., "siffle-ment" (whistling) "malad-ie" (illness), etc.: stem-nt. In French, there are nouns that derive from verbs, in particular, those which derive from the verbs ending in "-ifi-er" and "-is-er," e.g., "humid-ifi-er"–"humid-ifi-cation" or "cristall-is-er"–"cristall-is-ation." The verbs ending in "-ir," form nouns with an intermediary suffix "-iss-" preceding the noun suffix itself, e.g., "fin-ir"–"fin-iss-age/-eur/-euse." The corresponding pattern is stem-c-nt. It is also possible that an adjective with an explicit ending produced a verb in "-is-er," and then, a noun was formed on the basis of the verb. Here are some examples: "milit-aire"–"milit-ar-is-er"–"milit-ar-is-ation," "individu-el"–"individu-al-is-er"–"individu-al-is-ation." The corresponding pattern is stem-av-c-nt. Quite common in French are nouns derived from adjectives, e.g., "actu-al-ité," "techn-ic-ité," "act-iv-ité," "opt-ic-ien," "natur-al-isme/-iste," and so on. The corresponding pattern is stem-av-nt.

Adjective. French adjectives are inflected in gender and number. As well as for nouns, the inflection is expressed by adding special suffixes at the end of a word or by modifying the last suffix. Such grammatical features with all possible exceptions will be taken into account as suffixes of the type a. However, it should be noted that for a large number of adjectives the gender and number are not formally expressed. For these words, these grammatical characteristics must be determined from the context.

Now we will describe the basic patterns of endings. Let a denote the type of suffixes that form adjectives. Then possible patterns are: stem-ε and stem-a. For adjectives the number of different patterns of endings is not very large. The first one describes the cases where the adjective was derived from a previously derived noun. For example, "constitu-tion"–"constitut-ionn-el/constitut-ionn-elle" corresponds to the pattern stem-n-a. The second pattern describes the endings of adjectives that have been derived from verbs in "-ifi-er" or "-is-er." For example, the verb "mod-ifi-er" can form adjectives "modifi-ant(-e)," "mod-ifi-able," "mod-ifi-cateur," "mod-ifi-catrice," "mod-ifi-catif." As for the "-is-er" ending, the following examples can be provided: "automat-is-er"–"automat-is-able." The corresponding pattern is stem-c-a.

To sum up the above, for the French language we obtain the following data: a set of types of affixes $A = \{\mathrm{c}, \mathrm{v}, \mathrm{n}, \mathrm{av}, \mathrm{d}, \mathrm{a}, \mathrm{nt}\}$, and a set Sch and a set of structural patterns composed of 16 elements:

$$\mathrm{Sch} = \{(\mathrm{v}), (\mathrm{c}, \mathrm{v}), (\mathrm{n}, \mathrm{v}), (\mathrm{av}, \mathrm{c}, \mathrm{v}), (\mathrm{d}), (\mathrm{a}, \mathrm{d}), (\mathrm{n}, \mathrm{a}, \mathrm{d}), (\mathrm{nt}),$$
$$(\mathrm{c}, \mathrm{nt}), (\mathrm{av}, \mathrm{c}, \mathrm{nt}), (\mathrm{av}, \mathrm{nt}), (\mathrm{n}, \mathrm{av}, \mathrm{nt}), (\mathrm{a}), (\mathrm{n}, \mathrm{a}), (\mathrm{c}, \mathrm{a}), (\varepsilon)\} \quad (4.1)$$

In order to determine the grammatical features of word forms, one needs to obtain some information from each affix that can be identified in its composition. In the case of the French language, there is only one first level feature, namely the "part of speech." The set of representatives of this feature is as follows: $F = \{\mathrm{noun}, \mathrm{adjective}, \mathrm{adverb}, \mathrm{verb}\}$. The second-level features include "gender," its representatives are "masculine"/"feminine"; "number" ("singular"/"plural"), "person" ("first"/"second"/"third"), "tense" ("simple past"/"imperfect"/"present"/"future"), "mode" ("indicative"/"subjunctive"/"conditional"), "verb form type" ("finite"/"participle"/"infinitive"). In the case of the French language, the last suffix of the ending is the most "informative," as it carries the information on the part of speech and other grammatical features. In the same way as in building the patterns, we will describe for each part of speech the characteristics that can be defined.

When analyzing verbs, complex and simple forms should be distinguished. A complex verb form is a combination of a main and auxiliary verb expressing a semantically single meaning of a certain tense or voice. The general pattern is $\mathrm{A} + \mathrm{V_{pp}}$, where A is an auxiliary verb ("avoir" or "être"), and $\mathrm{V_{pp}}$ is the main verb in the form of past participle (participe passé). In this case, analyzing the form of the

auxiliary verb, one can determine the following grammatical features: tense, mood, person, and number. A simple form, respectively, refers to such a form of the verb that does not require the use of an auxiliary verb. In this case, the last affix of the verb should be analyzed. Grammatical characteristics that can be determined are mood, tense, person, and number.

As for the word forms of nouns and adjectives, it is also the last affix of the ending that should be analyzed. The grammatical features to be determined are gender and number. Obviously, a problem of multiplicity arises here. Some affixes can characterize both a noun and an adjective. In addition, the gender and the number of some nouns and adjectives cannot be uniquely determined. Therefore, it is impossible to avoid such uncertainties without analyzing the context.

For adverbs, the last affix of the ending can only provide information on the part of speech.

5. Results of Experiments

To determine the performance and the accuracy of the method of structural patterns of endings, the algorithms presented above have been implemented in C++ and applied to the French language. In particular, the algorithm of pseudo-stem identification with the division of the ending into affixes, as well as the algorithm of determining grammatical features have been implemented. A series of computational experiments has been performed. In all cases, we used collections of word forms, obtained from processed (marked up) literary French texts, the total size being 42,000 words. This allowed grouping words according to their part of speech, as well as analyzing different word forms (modified in gender, in number; for verbs, in tense, etc.).

First of all, we analyzed the accuracy of recognition of structural patterns and determining the grammatical features of random sets of words. We assumed that the recognition is correct if the set of proposed patterns includes the one with the correct stem and correct affix partition from the point of view of derivational processes. We also assumed that the grammatical features are determined correctly if the list of possible grammatical "profiles" includes the right one. The final result is that 95 of the 100 words are correctly divided and that 82 out of 100 words have the correct grammatical features.

Secondly, a series of experiments was performed in order to identify the relationship between the average number of suitable structural patterns for word forms of different parts of speech and restrictions on the number of letters in the stem. Table 1 shows the results of this experiment: the numbers represent the average number of identified structural patterns for the corresponding part of speech and the number of letters in the stem.

Table 1. Correlation of the average number of patterns with the restriction on the number of letters in the stem.

	3	4	5
verb	3.8288	3.4701	2.9294
adverb	7.225	6.8605	6.2403
noun	4.5446	3.9664	3.209
adjective	4.9874	4.5148	3.86

It can be seen that the greater the number of letters given in the stem, the fewer the possible patterns are proposed. For example, when the number of letters in the stem passes from 3 to 5 the multiplicity decreases by 22% for adjectives and by 29% for nouns. It is seen that by defining the parameters it is possible to reduce the multiplicity of the results of morphological analysis.

Another experiment that we carried out was to count the most frequently occurring patterns for each part of speech. In this experiment, we analyzed the lists of words of one part of speech. In one case, the set consisted only of patterns consistent with the part of speech under consideration. Thus, we obtained

a distribution of patterns for a specific part of speech. The following histogram (Fig. 1) displays the results of this type of experiment for verbs.

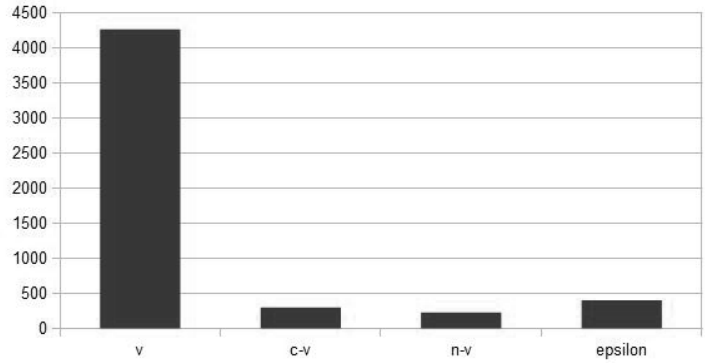

Fig. 1. Distribution of verb patterns.

In a second case, no restrictions were applied to the set Sch. As a result, we obtained a distribution over all possible patterns. For convenience in displaying the results, the patterns were combined according to their part of speech. The following histogram (Fig. 2) shows the results of this type of experiment for verbs.

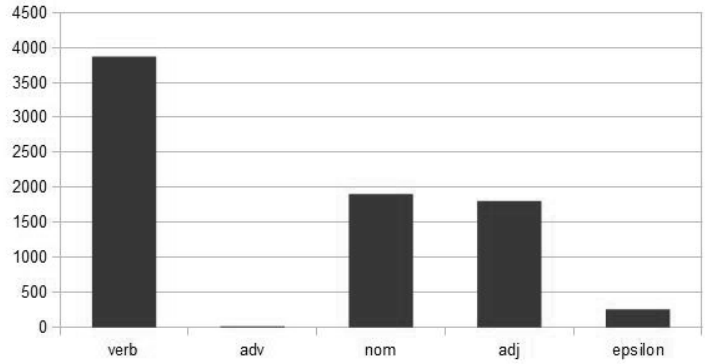

Fig. 2. Distribution of patterns of all parts of speech.

The histograms illustrate that in most cases we have correct patterns among the results. This fact once again proves the correctness of the proposed algorithms.

6. Conclusion

In this paper, a model for morphological analysis of word forms in languages with extensive affix inflection and derivation was proposed. The main task that the proposed model is designed to fulfill is to identify the pseudo-stems of word forms in a grammatically correct way. An accurate and powerful solution of this problem plays a key role in the optimization of the processes of information retrieval systems. In connection with this, an algorithm based on the method of structural patterns was proposed in order to identify the pseudo-stems of word forms and to divide endings into affixes. The basic idea of this method is to represent the affix part of the word as such a sequence of affixes that corresponds to one

of the (grammatically) available structures of the given language. The main advantage of this approach is that it does not require the use of morphological dictionary and is consistent with the grammatical features of the processed language.

We also addressed the determination of grammatical features of word forms. An algorithm based on the results of the previous one was proposed: the division of the affix part of words allows us to obtain information on the grammatical features of each affix, and as a result, of the entire word.

To demonstrate the possibility of using these algorithms, the process of building all the necessary structures has been described for the French language. In particular, the necessary types of affixes were identified, the patterns that describe the French word endings and the lists of grammatical features that may be determined have been provided. Also, using a computer program implementing all the proposed algorithms, a number of experiments for the numerical evaluation of the effectiveness of this method were performed. These results illustrate the possibility of applying the proposed approach for solving problems of information retrieval.

REFERENCES

1. A. Bolkovitianov and A. Chepovskiy, *The Morphological Analysis Algorithms of Computer Linguistics* [in Russian], I. Fedorov MSUPA, Moscow (2013).
2. J. DuBois and R. Lagane, *Livres de bord: Grammaire*, Larousse (2010).
3. E. Egorova and A. Chepovskiy, "Morphological approach to analyzing and indexing texts, the case of Indo-European languages," in: *Proc. CPT2013, 12–19 May 2013, Larnaca, Cyprus*, Izd. IFTI, Moscow; Protvino (2013), pp. 154–159.
4. M. Grevisse and A. Goosse, *Le bon usage*, Duculot, Paris (2011).
5. H. Huot, *La morphologie: forme et sens des mots du français*, Armand Colin (2006).
6. N. A. Katagoshchina, *How Words in French Are Formed* [in Russian], URSS; KomKniga, Moscow (2006).

E. Egorova
Higher School of Economics National Research University, Moscow, Russia

A. Lavrentiev
ICAR Research Lab — CNRS, Université de Lyon, ASLAN Labex, Lyon, France

A. Chepovskiy
Higher School of Economics National Research University, Moscow, Russia
E-mail: achepovskiy@hse.ru