Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora

Hervé Déjean
GREYC - CNRS - UPRESA 6072
Université de Caen - Basse Normandie
Herve.Dejean@info.unicaen.fr

Abstract

This paper describes an overview of a method which allows discovery of syntactic structures from untagged corpora. It is composed of three main steps: the discovery of the grammatical morphemes of the language. Then the construction of the chunks which are a multilingual conceptual level allowing the bypass of the limping notion of words. And Finally the discovery of the relations between chunks. We give an overview of the different procedures realized and we especially describe the discovery of morphemes. This operation is divided into three steps: the discovery of the most frequent morphemes of the language. Then the discovery of the other morphemes, and finally the segmentation of the words of the corpus. We concluded with the procedure of correction which required the chunk level. The concepts and algorithms were tested on a twenty natural languages like English, German, Turkish, Vietnamese, Swahili, Finnish, Latin, Indonesian.

1 Introduction

The method presented in this paper is inspired by the distributional approach developed by American structuralists between 1940 and 1950 (Harris, 1951). This approach is characterized by two facts: (a) the use of corpora and (b) the use of the notion of distribution instead of the sense of elements. The distribution of an element is the set of the environments in which the element occurs. Other works describe systems that induce structures from corpora, but they use tagged corpora (Brill, 1993), or grammatical informations (Brent, 1993), or work with artificial samples (Elman, 1990). Our originality lies in the fact that *we only use untagged and non artificial corpora without specific knowledge about the studied language.* We try to discover the structures of a natural language from raw texts of this language (on 100,000 words). We show that this kind of discovery is possible if we have some expectations of the structure of Natural Languages and if we use some formal properties.

The method relies on structural linguistic concepts: the morpheme, the chunk and the linearity of the language, i.e. the corpus is composed of a unidimensional sequence of elements. We first give here an overview of the concepts and general principles, from morphemes to syntactic structures discovery. Then we explain in detail how the segmentation is carried out.

2 The General Structure of Sentences

Natural Languages are a linear object. It means that sentences are sequences of sounds. In the case of written sentences, we consider them as sequences of letters (or characters). We also consider that languages are not only sequences of sound but are structured in several structural levels. We claim that these different levels are formally indicated in the sentences. How? Since sentences are unidimensional object, a simple way is the use of boundaries indicators between the elements which composed the sentences. Applying this principle on several Languages, we find out three multilingual and hierarchical levels: the morpheme level, the chunk level and the clause level. One useful formal criterium in the discovery of these structures is the position of words and morphemes relatively to beginnings and ends of sentences.

The morpheme level is already well known in Linguistics. The morphemes are the basic elements of the structure. In this paper, we call morphemes the affixes of the language. These elements are discovered during the operation of words segmentation. The morphemes contain as much structural informations as grammatical words and are essential to the discovery of the syntactic structures. Section 3 explains how we list them.

The higher level is the chunk one. We note that

Hervé Déjean (1998) **Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98 Workshop on Paradigms and Grounding in Language Learning,* ACL, pp 295-298.

some elements have a specific behaviour: they never occur at the beginning or at the end of the sentences. For example, the English word *the* never ends the sentences. There exists in all the studied languages similar elements (words or morphemes) that we can consider as indicating the beginning or the end of structures. *the grammatical words as well as the morphemes are consider as boundaries indicators.* We systematically consider grammatical words either as beginning indicator or as ending indicator. In practice, we tie them to their nearest lexical element (the following lexical for beginnings and preceding lexical for endings) In the same way, prefixes are considered as beginning and suffixes as ending. For example, both postpositions and inflexional suffixes are consider as ending indicators. The structures generated by these elements correspond to a lexical element (the nucleus of the chunk) surrounded by grammatical elements (words or morphemes, generally a combination of both).

The chunks may be viewed as non recursive phrases. Though each chunk of the corpus has not systematically boundaries indicators, there generally exists enough chunks which are delimited in order to allow the discovery of these boundaries. The discovery of such indicators is automatically realized for a large part.

The last level is the clause level. By working on boundaries indicators, we have noted that some indicators have a more specific behaviour. They mainly occur at the beginning or at the end of sentences. Furthermore, since some chunks have the same behaviour, clause boundaries are indicated either by morpheme, sole grammatical words or chunks. These elements always characterize elements of clauses: conjunctions or verbal phrases. For instance, English conjunction *but* begins sentences 672 times out of 760 occurrences. This behaviour is specific to clause boundaries indicators. German clause is, most of the times, closed either by grammatical words such as *her*, *zurück*, verbal particles, or by verbal phrases. In Turkish, the conjunction *ama* (but) occurs 763 times and begins 743 times. The Turkish clause is closed by verbal chunks, which implies that all the verbal morphemes (*-tir*, *yor*) are well marked as absolute endings. All the languages which have a *SOV* or *OSV* structure offer obvious end boundaries for clauses, and languages which have *VSO* or *VOS* structure offer beginnings boundaries for clauses. These formal informations do not cover all the formal characteristics of the languages, but they offer enough informations in order to discover the different syntactic relations between chunks, and offer a good starting point in order to find specific structures of a language, the position of the finite verb in German for instance.

In practice, we note that some languages privilege beginning indicators (prepositional languages as many European ones), others privilege ending indicators (postpositional languages as Turkish or Japanese) either at chunk level as at clause level, but they generally use the two methods. Some languages (Asian tonal languages) have a low number of boundaries indicators that complicates the chunks and clauses discovery. For the moment, we have stopped this study at clause level (or sequences of clause), but there perhaps exists higher levels.

3 The Morphemes Discovery

We now explain in details how the morphemes of a particular language are found. We refer the readers to other works dealing with this problem (Brent, Murthy, and Lunsberg, 1995), (de Marcken, 1995). Our aim is not the realization of a morphological analysis of each word of the corpus, but the production of the list of the morphemes for a given language. We do not try to discover all the morphemes contained in the corpus, since only the hundred most frequent ones are necessary in order to climb to chunks level. The method is inspired by the works of Zellig Harris. His algorithm is based on the number of different letters which follow a given sequence of letters. The increase of this number indicates a morpheme boundary. For instance, after the English sequence *direc*, we only find, in our corpus, one letter *t*. After *direct*, we find four letters: *i*, *l*, *o*, and *e* (*directly*, *director*, *directed*, *direction*). This increase indicates a boundary between the root (*direct* and the suffixes (*-ion*, *-ly*, *-or* and *-ed*). The algorithm works well when the corpus contains enough occurrences of a stem family. But, it may generate wrong segmentations. For example from the list *started, startled, startling*, the algorithm outputs this segmentation: *start-ed*, *start-led*, *start-ling*. The errors occur when two kinds of stem families are used for the segmentation. (Harris, 1955) exposes several variations more or less complex. Their implementation does not furnish great improvements.

Our idea for improving the segmentation is to divide into three steps this operation. The first step computes the list of the most frequent morphemes. The second steps extends the list by segmenting words with the help of the morphemes already generated. The third step consists in the segmentation of all the words with the morphemes obtained at the second step. The algorithm is illustrated with the suffixes segmentation, but the discovery of prefixes is totally symmetric: we just reverse the letters of the words.

3.1 The discovery of the most frequent morphemes

The discovery of the most frequent morphemes is based on Harris algorithm. We try to find beginnings or endings of words which have the following property: after a given sequence of letters, we count the number of different letters. If this number is higher than a threshold (half of the letters of the alphabet), we arrive at a morpheme boundary, except in the case we are in the sequence which corresponds to a longer morpheme, a case we can detect. For example, before the sequence *on*, we found 18 different letters, thus *on* may be a morpheme. But 292 of these words in the corpus end with *ion* out of 367 which end with *on*. Since the longest sequence *ion* represents more than 50% of the word ended by *on*, we consider that *on* is a part of the morpheme -*ion*[1]. We only keep on the sequences which have a frequency higher than 100.

Table 1: The most frequent morphemes.

English	-e -s -ed -ing -al -ation -ly -ic -ent
French	-s -e -es -ent -er -és -re -ation -ique
German	-en -e -te -ten -er -es -lich -el
Turkish	-ın -in -lar -ler -dan -den -ini -ını
Swahili	-wa -ia -u -eni -o -isha -ana -we
Swahili	wa- m- ku- ali- ni- aka- ki- vi-
Vietnamese	NONE

3.2 The discovery of other morphemes

Once these morphemes are found,, we use them in order to segment words and to find out other morphemes thanks to the following rule: For a given sequence of letters (*light* in Table 3.2), we check on if the next sequences of letters correspond to morphemes already found. If half of them belongs to the morphemes found (like *-s -ed -ing -ly -er* , then the others (*-ness -en -est*) are also considered as morphemes.

This algorithm also generates wrong morphemes, but the frequency of them is very low (1 or 2). Thus, we only keep on new morphemes which have a frequency higher than a given threshold (5 in practice). The morphemes with a frequency lower than this threshold are not found. The morphemes list may greatly depend on the type of corpus used. The number of morphemes depends on the morphology of the language. In Vietnamese, no morpheme is found.

Table 2: Second step of the morphemes discovery.

Morphemes found	words	New Morphemes
	light	
-s	lights	
-ed	lighted	
-ing	lighting	
-ly	lightly	
-er	lighter	
	lightness	-ness
	lightest	-est
	lighten	-en

In English, a list of fifty morphemes is generated (Table 3). The Turkish list contains more than 500 morphemes. We note that morphemes have a similar behaviour as words: a small number of them possesses a high frequency and corresponds to the major occurrences of the corpus. We do not try to generate all the morphemes of the corpus, since *the hundred most frequent morphemes are sufficient for the construction of the higher level* (the chunk level). Some morphemes of the list given in Table 3 are composed of a sequence of morphemes (*ful-ly, ence-s*). In highly morphological languages, most of the morphemes correspond to sequence of elementary morphemes. We do not try to resegment these elements now. Because of the presence of one letter morphemes, the resegmentation inevitably lead to the segmentation of the morphemes in letters. We wait the chunk level in order to refine these morphemes (Section 4).

Table 3: Final English Morphemes

suffixes:-y -ward -ure -s -ry -ously -ous -ors -or -ness -ments -ment -ly -less -ively -ive -ity -ious -ions -ion -ings -ingly -ing -in -ily -ies -ic -ible -fully -ful -est -es -ers -er -ence -ences -en -ement -ements -ely -ed -e -ations -ation -ance -ances -ally -al -age -ably -able -'s

prefixes: dis- in- pro- re- un-

3.3 The segmentation of the words

Once the list of the morphemes is found, we use it for segmenting all the words of the corpus. We segment the words by applying the longest match algorithm: we segment each word with the longest morpheme which matches beginning or ending of the word. In order to allow the chunks discovery, there are some words which are not segmented: the most frequent ones (5% of the words). They generally cor-

[1] Form the sequence *on*, we generate the morpheme -*ation*.

respond to grammatical words, and we do not segment them in order to make easier the chunks discovery. The following section explains how the lexical words which appear in this list are segmented. We check on the segmentation of 500 words randomly selected and we obtain 8 segmentations we consider as wrong (as *compla-in*, *forse-en* or in German word *antwortest*[2] segmented in *antwor-test* with the morpheme *-test* (correct in *lern-test*[3] , preterit 2 pers.).

Harris' algorithm realizes the segmentation of words during the discovery of morphemes. The dissociation of the two phase allows a more correct segmentation. With Harris algorithm, the words *startling*, *startled* and *started* generate the following segmentation: *start-ed*, *start-led*, *start-ling* (Section 3). With our method, the segmentation is *startl-ing*, *startl-ed* and *start-ed* since *-ling* and *-led* are not morphemes. It may be generated some errors (as *antwortest*) but only for few words.

4 The correction of words segmentation

We now explain how the frequent lexical words and the morphemes composed of a sequence of other morphemes are segmented. The method use the contextual informations discovered in the chunk level. During the construction of chunks, we generate bigrams of morphemes (Table 4). We use these bigrams in order to refine the segmentation. Each word or morpheme occurring in a context corresponding to chunk structure will be segmented. For example, the German word *Hauses* (house) occurring in *des Hauses* is segmented in *des Haus-es* thanks to the context *des S-es*[4]. The algorithm is the same for sequences of morphemes. The French sequence *antes* is segmented is *ante-s* thanks to the contexts *les S-s*.

Table 4: Segmentation correction.

bigrams		correct segmentation
German		
des S-es	des Hauses	des Haus-es
ich S-te	ich machte	ich mach-te
French		
les S-s	les S-es	les S-e-s
les S-s	les S-antes	les S-ante-s

5 The necessity of morphemes in a procedure of discovery

The morphemes level allows the emergence of structures which hardly appear at word level: structures which are marked by morphemes like the concordance structures. For example, the French structure (*les-S-s S-s*) or German one (*des-S-en S-es*) are easily found thanks to their frequencies. Other structures are also easily found like *adverb-verb* structure in English, characterized by the high frequency of the bigrams (*S-ly S-ed*). Another useful morphemes are inflectional ones which mark relations between chunks at clause level. The relations between chunks are discover since bigrams composed of grammatical words and morphemes belonging to contiguous chunks. Frequent bigrams generally correspond to relations between two chunks (like *S-ed S-ly*). A positional criterium allows the elimination of bad frequent bigrams like (*of-S S-ed*) (Noun Complement - Verb sequence): since this bigram never begins a sentence, we consider that the structure is not complete and requires another chunk in order to complete the relational structure (*the-S of-S S-ed*).

We conclude by claiming that morphemic level is essential and unavoidable in a procedure of syntactic structures discovery.

References

Brent, Mickael. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19:243–262.

Brent, Mickael, Sreerama K. Murthy, and Andrew Lunsberg. 1995. Discovering morphemic suffixes : A case study in mdl induction. In *Fifth International Workshop on AI and Statistics,Ft. Lauderdale, florida.*

Brill, Eric. 1993. Automatic grammar induction and parsing free text : a transformation-based approach. In *ACL93.*

de Marcken, Carl. 1995. The unsupervised acquisition of a lexicon from continous spreech. Technical report, MIT Artificial Intelligence Lab. Memo 1558.

Elman, J.L. 1990. Finding struture in time. *Cognitive Science*, 14:179–211.

Harris, Zellig. 1951. *Structural Linguistics.* The University of Chicago Press.

Harris, Zellig. 1955. From phonemes to morphemes. *Language*, 31(2):190–222.

[2] (you) answer. Antwort-en: to answer

[3] (you) learned. Lern-en: to learn.

[4] S for Stem