MORPHONETTE: A PARADIGM-BASED MORPHOLOGICAL NETWORK

NABIL HATHOUT[*]

ABSTRACT: Morphonette is a derivational morphological network of French in which words are characterized by the paradigms they belong to: their derivational families and series. This resource was built from a large lexicon of written forms. Its creation is based on a measure of morphological proximity and on formal analogy. Three additional criteria are used to separate the relation between members of the same family from the ones between members of the same series and to eliminate most of the morphologically invalid connections. The paradigmatic organization of the lexicon is described by means of filaments which account for the different morphological properties of individual words.

KEYWORDS: morphological network, derivational morphology, morphological paradigms, morphological similarity, formal analog,.

1. THE NEED FOR DERIVATIONAL RESOURCES

The relationship between morphology and computational linguistics is an old one, in part because sentences cannot be properly analyzed without a thorough treatment of inflectional morphology. Nowadays, this treatment has become standard for many European languages such as English, French or Italian. But the same is not true of derivational morphology. The difference is clearly visible when we consider morphological derivational resources. Apart from CELEX database (Baayen et al., 1995) which describes the derivational morphology of fragments of Dutch, English and German lexicons, no other large-coverage derivational resource is available. Several efforts to fill this gap are underway, such as the creation of the DADI dictionary of Italian derivational affixes (Grandi & Montermini,

[*] CLLE-ERSS, CNRS & Université de Toulouse

I thank the anonymous reviewer for his useful comments on this paper.

2010). For this dictionary, the process is primarily a manual one. The research I present in this paper has a similar objective, namely the creation of a morphological derivational resource, but this one for French. A second difference is that the construction is fully automatic. Of course, this resource will have to be revised manually in order to eliminate any errors induced by the irregularities accumulated in the lexicon over time.

Specifically, this paper presents a new morphological derivational resource for French: the Morphonette network. Morphonette is a relational lexicon which describes derivational relations between words. It is characterized by a paradigmatic structure based on derivational families and series. This structure is described by means of morphological filaments which combine information from both types of paradigms: families and series. This research is situated in a word based approach (Anderson, 1992, Aronoff, 1994, Stump, 2001). The creation of Morphonette does not involve any word formation rules or any statistical modelling like most of the systems competing in the Morpho-Challenge contests (Kurimo et al., 2010). It is based on a new measure of morphological proximity and on formal analogy. The paper presents these two techniques and how they are used for the identification and characterization of the morphological relations. These techniques are supplemented by a set of criteria which eliminate the less reliable relations and separate the relations between members of the same families from the ones between members of the same series.

The remainder of the paper is structured as follows. The next section presents the general theoretical framework within which the creation of Morphonette is formulated. Section 3 outlines the methods used to create the network, addresses the problem of computational complexity this creation raises and describes how it is reduced. In section 4, I present four criteria intended to eliminate some of the erroneous relations from the initial network and to separate the families from the series. Section 5 describes the filaments which compose the structure of Morphonette. Section 6 compares some related works and finally, section 7 offers a short conclusion and some directions for further research.

2. THE MORPHOLOGICAL STRUCTURE OF THE LEXICON

The research reported in this paper seeks to discover the morphological structure of the lexicon. The proposed method does not aim at individually parsing each of the words in the lexicon but at performing one global analysis of the entire. The creation of Morphonette is indeed situated in a theoretical framework radically word-based (Aronoff 1994) and also paradigmatic (Becker, 1993; Booij, 1997, 2008). In this view of morphology, words are not made up of morphemes. Rather, they are minimal morphological units. Thereby, they do not have any morphological structure. The morphological structure is conceived as a level of organization of the lexicon. This structure is composed of morphological relations between the words memorized in the lexicon (Bybee, 1985, 1995).

2.1 Paradigms

The morphological relations are organized into paradigms. Morphological paradigms can be divided into at least eight different types on the basis of three oppositions: inflexion versus derivation; families versus series; morphology versus lexicon. Table 1 presents two words belonging to an instance of each of these eight types. For example, *lavons* ('wash' 2nd plural present indicative) and *lavera* ('wash', 3rd singular future indicative) belong to the same morphological inflectional family while *furieux* 'furious' and *curieux* 'curious' belong to the same lexical derivational series (see Hathout, 2009a, b for more details). The other examples in table 1 are *allons* ('go' 2nd plural present indicative), *ira* ('go' 3rd singular future indicative), *coupons* ('cut' 2nd plural present indicative), *sommes* ('be' 2nd plural present indicative), *dériver* 'derive', *dérivable* 'derivable', *variable* 'variable', *prison* 'jail' and *carcéral* 'penitentiary'.

	inflectional		derivational	
family	*lavons, lavera*	*allons, ira*	*dériver, dérivable*	*prison, carcéral*
series	*lavons, coupons*	*lavons, sommes*	*dérivable, variable*	*furieux, curieux*
	morphological	lexical	morphological	lexical

TABLE 1. EIGHT TYPES OF MORPHOLOGICAL PARADIGMS

If the notion of morphological families is also well-established, that of morphological series is less well-known. Families can be defined as sets of words that are very close to each other in the sense that they share as much semantic and formal[1] properties as possible and that these properties are as much specific as possible. Note that this definition is different from the usual one where families are defined as sets of words that share a common root, or in other words, that are derived from each other (Schreuder & Baayen 1997). For instance, the family of *produire* 'produce' includes

[1] The formal properties of a word are both its phonemic and graphemic features.

produit ‘product’, *production* ‘production’, *reproduire* ‘reproduce’, *productif* ‘productive’, *improductif* ‘unproductive’, *productivité* ‘productivity’, *productiviste* ‘productivist’, etc. in addition to *produire* itself. The notion of family can be extended to the inflectional level: an inflectional family is the set of the inflected forms of one lexeme. For instance, the inflectional family of the adjective *vert* ‘green’ contains the forms *vert* (masculine singular), *verte* (feminine singular), *verts* (masculine plural) and *vertes* (masculine plural).

The second important notion is that of series. A series is a set of words as large as possible such that these words (1) share very general semantic and formal properties and (2) participate in the greatest possible number of analogies that involve other words of the series. For instance, the derivational series of the noun *lavage* ‘wash’ contains all the *-age* suffixed derivatives of the French lexicon: *façonnage* ‘shaping’, *rabotage* ‘planning’, *étiquetage* ‘labeling’, *maquillage* ‘making-up’, etc. Similarly, inflectional series can be defined as sets of equivalent forms from lexemes of the same category. For instance, the inflectional series of *lavera* contains all the 3rd singular future indicative verb forms: *marchera* ‘walk’, *pensera* ‘think’, *écrira* ‘write’, *aimera* ‘like’, etc.

In the remainder of this paper, I am concerned only with two types of paradigms: morphological derivational families and morphological derivational series. These paradigms form a lexical grid which could be illustrated as in figure 1 where the families are represented horizontally and the series vertically. The figure has been simplified for sake of readability since, for instance, *modifiable* ‘modifiable’ is also connected with *modifier* ‘modify’ and *modification* ‘modification’ within its family and with *rectifiable* ‘rectifiable’ and *sanctifiable* ‘worthy of being sanctified’ within its series. Families and series are interconnected through the words they words they share. Series are connected to each other via the families and vice versa. The other examples in figure 1 are *modificateur* ‘modifier’, *fructifiable* ‘which can bear fruits’, *fructificateur* ‘which brings moral benefits’, *fructifier* ‘bear fruits’, *fructification* ‘fructification’, *rectificateur* ‘rectifier’, *rectifier* ‘rectify’, *rectification* ‘rectification’, *sanctificateur* ‘sanctifier’, *sanctifier* ‘sanctify’ and *sanctification* ‘sanctification’.

	↕		↕		↕		↕	
↔	modifiable	↔	modificateur	↔	modifier	↔	modification	↔
	↕		↕		↕		↕	
↔	fructufiable	↔	fructificateur	↔	fructifier	↔	fructification	↔
	↕		↕		↕		↕	
↔	rectifiable	↔	rectificateur	↔	rectifier	↔	rectification	↔
	↕		↕		↕		↕	
↔	sanctifiable	↔	sanctificateur	↔	sanctifier	↔	sanctification	↔

↕ ↕ ↕ ↕

FIGURE 1. FAMILIES AND SERIES FORM A LEXICAL GRID

2.2 Analogy

The families and series which make up the grid are interconnected by analogies such as (1) where the notation a:b = c:d denotes the existence of an analogy between a, b, c and d, classically expressed as "a is to b as c is to d". Words in the lexicon participate in a large numbers of analogies. This number is of the order of $n \times m$ for a word with a family of n members and a series of m members.

Analogy plays an essential role in the characterization of the paradigms and their members. For instance, in order to find out the properties that characterize the series of *fructificateur*, we can use the analogies in (1) to determine a set of properties such as (2) where *Ppt(x)* represents the set of the properties of x.

(1) fructificateur:rectificateur = fructifier:rectifier
fructificateur:rectificateur = fructifiable:rectifiable
fructificateur:fructification = rectificateur:rectification
fructificateur:fructification = modificateur:modification

The first line in (2) describes the properties that are common to *fructificateur* and *rectificateur* but are not possessed by neither *fructifier* nor *rectifier*.[2] The sets of properties in (2) should in theory all be the same. In practice, the properties of the series of *fructificateur* could be defined as the union of all these sets.[3]

(2) (Ppt(fructificateur) \ Ppt(fructifier)) ∩ (Ppt(rectificateur) \ Ppt(rectifier))
(Ppt(fructificateur) \ Ppt(fructifiable)) ∩ (Ppt(rectificateur) \ Ppt(rectifiable))
(Ppt(fructificateur) \ Ppt(fructification)) ∩ (Ppt(rectificateur) \ Ppt(rectification))
(Ppt(fructificateur) \ Ppt(fructification)) ∩ (Ppt(modificateur) \ Ppt(modification))

2.3 Morphological analysis

This conception of morphology allows us to redefine morphological analysis as a global description of the paradigmatic structure of the lexicon. The aim is no longer to cut the individual words into morphemes and describe their composition. The analysis of a given word then consists in identifying its

[2] In this formula, backslash (\) denotes set subtraction.

[3] These properties only characterize the most regular parts of the paradigm. Possible irregularities can be described by comparing the properties computed at the level of the words they concern with the global properties of the paradigm.

position in the lexical grid and hence their relations with the other words in the lexicon. This position is defined by two morphological coordinates: its family and its series. For instance, in morpheme-based theories, a word such as *fructification* was classically analyzed as in (3) where the second line lists the chunks of the written form which correspond to each morpheme.[4]

(3) [[[fruit]$_N$ [-ifier]]$_V$ [-ion]]$_N$

fruct -ificat -ion

I propose instead to replace this decomposition by an identification of a large enough subset of its derivational family (e.g. *fruit* ‘fruit’, *fructifier*, *fructifable*, *fructification*, etc.) and a large enough subset of its derivational series (e.g. *modificateur*, *compilateur* ‘compiler’, *pollinisateur* ‘pollinator’, etc.). With this type of analysis, the morphological properties of the word[5] are fetched from the ones of these two paradigms and from the analogies they induce. Moreover, this analysis does not resort neither to the notion of morphological rule, nor that of morpheme, affix or morphological exponent. These notions are in fact useless when one seeks to identify morphological relations between words.

2.4 The task

The task I present in this paper is a first attempt to create a French morphological network which describes the derivational families and series of a significant fragment of the lexicon. The network is constructed from the word list of a large machine readable dictionary: the *Trésor de la Langue Française informatisé* (TLFi). I have only used the written forms and categories (i.e. parts of speech). Therefore, I did not resort to the semantic information, which distinguishes this work from previous ones (Hathout, 2009a, b) where morphological relatedness was computed both from formal features and semantic one extracted from the words definitions. However, the general method is still the same: it is mainly based on a measure of morphological similarity and on formal analogy. The novelty in the research presented here is the use of four criteria intended to remove a large part of the morphologically invalid relations from the network and to separate the families from the series. The primary objective of the present work is indeed to create a reliable resource which contains as little errors as possible. The downside is of course that the coverage of the network is small with respect

[4] The first two chunks are suppletive forms of their respective morphemes. Allomorphy is described in just the same way.

[5] These are the properties that are associated with the location of the word in the morphological grid.

to the size of the initial lexicon. Another significant contribution of this research is a finer characterization of the paradigmatic structure of the lexicon which leads to the definition of a novel data structure: the morphological filaments.

3. REDUCTION OF THE COMPUTATIONAL COMPLEXITY

The creation of Morphonette is based on two simple observations. First, formal analogies shows a better precision than isolated binary relations yielded by affixation schemata (Jacquemin, 1997; Gaussier, 1999; Hathout, 2005) because they must hold between four words and thereby involve four binary relations (*a*:*b*, *a*:*c*, *b*:*d* and *c*:*d*). The second advantage of analogies over binary relations is their redundancy. A binary relation included in one analogy is also normally involved in a large number of other analogies. An analogy made up of relations that do not occur in other analogies is often erroneous as in (4) where neither of the *allier*:*allouer* relation ('ally', 'allocate') nor the *dévier*:*dévouer* one ('deviate', 'devote') can be extended into larger morphological families. For instance, the other members of the family of *dévier* (*déviation* 'deviation', *déviant* 'deviant', *déviateur* 'deviator') do not form analogies with *dévouer* and conversely, the members of the family of *dévouer* (*dévoué* 'devoted', *dévouement* 'devotion') do not occur in analogies with *allier*. The redundancy thus enhances the reliability of the binary relations.

(4) *allier*:*allouer* = *dévier*:*dévouer*

The morphological structure of the lexicon can be obtained from the binary relations occurring in the analogies provided that one can separate the families from the series. In other terms, the overall task could be divided into two subtasks: (*i*) collect all the analogies that hold between the words in the lexicon; (*ii*) type the binary relations contained in these analogies. The entire process is automatic (see section 4).

3.1 Morphological neighbourhood

Analogies being quaternary relations, ideally, all the word quadruples that can be formed from the lexicon should be checked one by one. In practice, testing all the quadruples is impossible because they are numerous and the complexity of their verifications is too high. For instance, the 100,000 headwords of the TLFi yield 10^{20} quadruples. Moreover, checking whether one quadruple (*a*, *b*, *c*, *d*) is a formal analogy or not has a computational complexity on the order of $o(n^4)$ where n is the length of the longest of the four words, knowing that the mean length of the TLFi headwords is 15 characters. An exhaustive checking of all the quadruples is therefore out of reach for current computers.

This observation does not put into question the use of analogy. It is indeed possible to reduce the search space by 10 orders of magnitude by limiting the checking to the quadruples made up of words that are the most likely to be morphologically related, that is, words which belong to the same derivational family or the same derivational series. Actually, if a:b = c:d is an analogy, then a and d are each morphologically related to both b and c. The selection of the words that are likely to be morphologically related is achieved by means of a measure of morphological similarity able to bring closer the words that belong to the same derivational family or the same derivational series. The measure relies on two facts:

- words are all the more strongly morphologically related as they share a large number of phonemic properties;
- words are all the more strongly morphologically related as the properties they share are specific, that is, infrequent.

As said before, Morphonette is constructed with formal and categorical properties only. In this experiment, the formal features were extracted from phonetic transcriptions computed by means of the LIA_PHON phonetizer (Béchet, 2001). These transcriptions are written in a format where each phoneme is represented by two characters. For instance, the transcription of the adjective *atomique* `atomic' is `aattoommiikk` which corresponds to the IPA /atɔmik/. One additional # character is added at the beginning and at the end of the transcription to mark its limits.

The formal properties used to estimate the morphological proximity between words all the n-grams of phonemes occurring in their transcriptions, that is, all the subsequences of n phonemes for $1 \leq n \leq l$ where l is the length of the transcription. These features are extremely redundant since almost all the subsequences are part of many features. This redundancy can be seen in (5) which lists the formal properties of the word *atomique*, grouped by length for the sake of readability.

(5) #atɔmik#
#atɔmik atɔmik#
#atɔmi atɔmik tɔmik#
#atɔm atɔmi tɔmik ɔmik#
#atɔ atɔm tɔmi ɔmik mik#

#at atɔ tɔm ɔmi mik ik#
#a at tɔ ɔm mi ik k#
a t ɔ m i k

These properties are able to capture a large part of the formal regularities in a language with a concatenative morphology such as French because they favour longer sub-sequences of phonemes over short ones. For instance, *atomique* ‘atomic’ and *atomiser* ‘atomize’ will be brought closer because they share the *n*-gram #atɔmi as well as 20 smaller ones. Furthermore, #atɔmi only appears in the transcriptions of 9 other words in the lexicon (*atomicien* ‘nuclear physicist’, *atomicité* ‘valence’, *atomiquement* ‘atomically’, *atomisation* ‘atomisation’, *atomiseur* ‘atomiser’, *atomisme* ‘atomism’, *atomiste* ‘atomist’, *atomistique* ‘atomistical’) which makes it a highly specific feature (i.e. *n*-gram).

Practically, the morphological similarity between words is calculated by simulating a spread of activation through a bipartite graph. The graph contains on one side the words and on the other their properties (see figure 2). The words are connected to all their features and the edges are weighted so that the activation is uniformly spread from each word to all its features and from each feature to all the words which have it. More precisely, the neighbors of a word *w* are identified by initiating activation at the vertex which represents *w* and then spreading it evenly to all the features of *w* (i.e. every feature of *w* receives the same fraction of the initial activation). In the next step, the activations located at the feature vertices are uniformly spread back to the vertices of the words which possess them. The strength of the activation obtained by these words is taken as an estimate of their level of relatedness to *w* (Hathout, 2008).

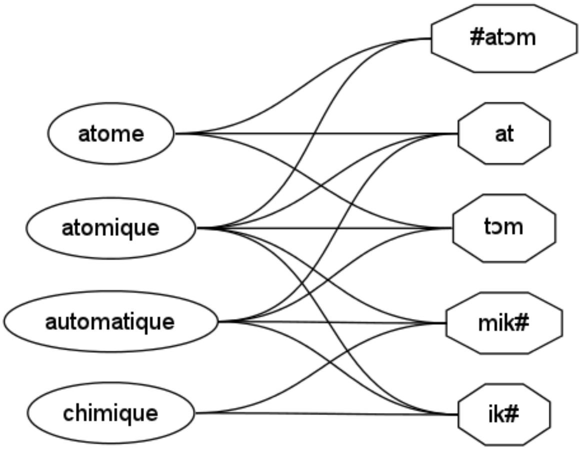

FIGURE 2. EXCERPT OF THE BIPARTITE GRAPH SET UP TO COMPUTE THE MORPHOLOGICAL PROXIMITY BETWEEN WORDS

This method enables us to compute morphological neighbourhoods for all the words in the lexicon. These neighbourhoods are illustrated in figure 3 which displays the 50 nearest neighbours of the verb *fructifier* 'bear fruits'. The words which belong to the family of *fructifier* are in boldface, those which belong to its series are in italic and those which are not morphologically related to *fructifier* are underlined.

fructifier fructifiant fructificateur fructification fructifère *sanctifier rectifier présanctifier* **fructivore** fructidorien fructidorienne fructidoriser fructidor **fructueusement fructueux fructuosité fructose** obstructif constructif instructif désobstructif destructif autodestructif **usufructuaire infructueusement infructueux infructuosité** sanctifiant sanctifiable rectifieuse rectifieur rectifiant rectifiable *transsubstantifier substantifier stratifier cimentifier certifier savantifier refortifier ratifier présentifier pontifier plastifier notifier nettifier mortifier mythifier mystifier quantifier*

FIGURE 3. THE 50 NEAREST NEIGHBOURS TO *FRUCTIFIER* 'BEAR FRUITS'[6]

The figure shows that the members of the family appear globally at the beginning of the list and that the members of the series come next. This trend is a direct consequence of the fact that the properties associated with stems are larger in number and more specific than those associated with affixes or compound elements. The morphological neighbourhoods are key to the reduction of the search space of analogies because they enable us to check only those quadruples (a, b, c, d) where b is a neighbour of a, c is a

[6] Translation of the neighbours: *fructifier* 'bear fruits' *fructifiant* 'bearing fruits' *fructificateur* 'which brings moral benefits' *fructification* 'fructification' *fructifère* 'which bears fruits' *sanctifier* 'sanctify' *rectifier* 'rectify' *présanctifier* 'pre-sanctify' *fructivore* 'frugivorous' *fructidorien* 'participant in the coup of 18 Fructidor' *fructidorienne* 'female participant in the coup of 18 Fructidor' *fructidoriser* 'send into exile following the coup of 18 Fructidor' *fructidor* 'twelfth month in the French Republican calendar' *fructueusement* 'fruitfully' *fructueux* 'fruitful' *fructuosité* 'fruitfulness' *fructose* 'fructose' *obstructif* 'obstructive' *constructif* 'constructive' *instructif* 'instructive' *désobstructif* 'unblocking' *destructif* 'destructive' *autodestructif* 'auto-destructive' *usufructuaire* 'usufructuary' *infructueusement* 'unfruitfully' *infructueux* 'unfruitful' *infructosité* 'unfruitfulness' *sanctifiant* 'sanctifying' *sanctifiable* 'sanctifiable' *rectifieuse* 'female rectifier' *rectifieur* 'rectifier' *rectifiant* 'rectifying' *rectifiable* 'rectifiable' *transsubstantifier* 'transform' *substantifier* 'transform into substance' *stratifier* 'stratify' *cimentifier* 'make something acquire the properties of cement' *certifier* 'certify' *savantifier* 'give an erudite aspect' *refortifier* 're-strengthen' *ratifier* 'ratify' *présentifier* 'make present to the consciousness' *pontifier* 'pontificate' *plastifier* 'plasticise' *notifier* 'notify' *nettifier* 'clarify' *mortifier* 'mortify' *mythifier* 'mythologize' *mystifier* 'mystify' *quantifier* 'quantify'

neighbour of *a* and *d* is a neighbour of both *b* and *c*.[7] For instance, the number of quadruples to be checked drops from 10^{20} to 10^{10} for a lexicon of 100,000 entries, that is, the number of headwords of the TLFi. The task then becomes quite within the reach of current computers.

The adequacy of the measure of morphological similarity can be assessed by calculating the number of analogies that can be formed when the size of the neighbourhoods increases. Figure 4 shows that this number increases logarithmically. In other words, it shows that globally the words which are morphologically related to the headwords are located at the top of their neighbourhoods and that nearly two thirds of the analogies can be collected if one only considers the first 100 neighbours of each word.

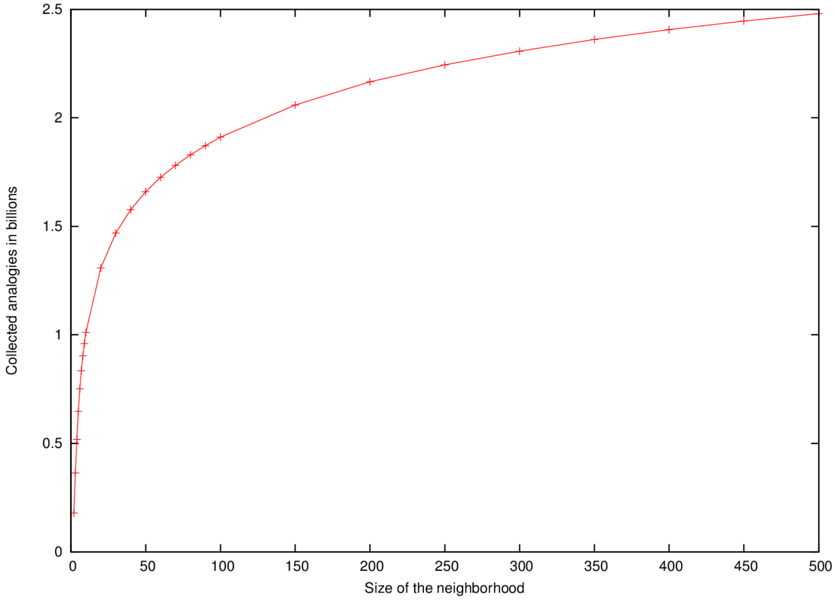

FIGURE 4. NUMBER OF FORMAL ANALOGIES THAT CAN BE FORMED AS A FUNCTION OF THE SIZE OF THE NEIGHBOURHOODS.

3.2 Formal Analogy

The discovery of the relations between the words which compose the Morphonette network is based on the morphological neighbourhoods and on formal analogy (Lepage, 2003; Stroppa, 2005; Stroppa & Yvon, 2005). A formal analogy is an analogy which holds between formal representations, such as phonemic transcriptions, written forms, feature structures, trees, etc.

[7] These conditions are based on the assumption that the neighbourhood of a word includes its entire family and its entire series.

Four representations (a, b, c, d) form a formal analogy if the differences between a and b are identical to those between c and d. This is the case for example for the written forms in (6) which can be represented as in figure 5.

(6) *fructueux:infructueusement = soucieux:insoucieusement*
‘successful’, ‘unsuccessfully’, ‘anxious’, ‘carelessly’

The differences are first, the insertion of the string *in* at the beginning of the written form and second, the substitution of the string *sement* for *x* at the end. *ε* represents the empty string.

ε	fructueu	x	ε	soucieu	x
in	fructueu	sement	in	soucieu	sement

FIGURE 5. THE DIFFERENCES BETWEEN *FRUCTUEUX* AND *INFRUCTUEUSEMENT* ARE THE SAME AS THE ONES BETWEEN *SOUCIEUX* AND *INSOUCIEUSEMENT*

Formally, given an alphabet L, four strings $(a, b, c, d) \in L^{*4}$ form an analogy $a{:}b = c{:}d$ if there exist four factorizations of length n of the four strings $(f(a), f(b), f(c), f(d)) \in (L^{*4})^n$ such that:

$$\forall i,\ 1 \leq i \leq n,\ (f_i(a), f_i(d)) \in \{(f_i(b), f_i(c)),\ (f_i(c), f_i(b))\}$$

In the worst case, n is the length of the longest string and the computational complexity of the verification of the analogy is in $o(n^4)$. The accurate verification of analogies is therefore extremely expensive and not suitable for the number of quadruples that can be formed among the morphological neighbours. So, I adopted a less expensive method, the trade-off of which is that full completeness is no longer guaranteed. The method associates each couple of words with a string edit signature which describes the edit operations to be performed on the representation of the first word in order to transform it into the representation of the second one. These signatures were calculated by means of the python library pyLevenshtein. For example, the signature of the pair (*fructueux* ‘successful’, *infructueusement* ‘unsuccessfully’) is presented in (7) where *ε* represents the empty string and @ an identical substring in both words.

(7) (insert, *ε*, *in*), @, (replace, *x*, *sement*)

Notice that (7) is also the signature of the pair (*soucieux* ‘anxious’, *insoucieusement* ‘carelessly’). The primary advantage of this method is its computational complexity in $o(n^2)$. Another one is that signatures are calculated only once for each pair which further reduces the cost of the verification.

4. MORPHOLOGICAL NETWORK

The Morphonette lexicon has been constructed from an initial graph created from formal analogies. More specifically, a neighbourhood of 100 words is calculated for each entry of the input lexicon (the TLF word list). Then all the quadruples (a, b, c, d) such that b is a neighbour of a, c is another neighbour of a which follows b in the neighbourhood of a and d belongs to the intersection of the neighbourhoods of b and c are formed. Next, for each quadruple, we check whether $s(a, b) = s(c, d)$ and $s(a, c) = s(b, d)$ where $s(x, y)$ is the edit signature of the couple (x, y). The set A of all these analogies is then used for the construction of the initial graph $G = (V, E)$ where V is the set of the words that occur in A, $E \subset V \times V$ is a set of edges (x, y) such that there exists an analogy $x{:}y = z{:}t \in A$. The initial graph contains 75,832 vertices and 3,159,981 edges. The next step is to type the edges as relations between members of the same family or between members of the same series. For instance, in the analogy *fructueux*:*infructueusement* = *soucieux*:*insoucieusement*, the edges (*fructueux*, *infructueusement*) and (*soucieux*, *insoucieusement*) connect members of the same family and dually the edges (*fructueux*, *soucieux*) and (*infructueusement*, *insoucieusement*) join members of the same series. This step also involves a cleaning up of the graph to remove some spurious relations such as in (8) where the second lines give the phonetic transcriptions.

(8) a. *destructeur*:*structural* = *descripteur*:*scriptural*
ddaissttrruukkttoer:ssttrruukkttuurraall =
ddaisskkrriippttoerr:sskkrriippttuurraall
‘destructor’, ‘structural’, ‘descriptor’, ‘scriptural’
b. *foyère*:*cloyère* = *foisonner*:*cloisonner*
ffwwaayyairr:kkllwwaayyairr = kkllwwaazzoonnei:ffwwaazzoonnei
‘hearth marble’, ‘oysters basket’, ‘abound’, ‘partition’
c. *paissant*:*abaissant* = *paye*:*abeille*
ppaissan:aabbaissan = ppaiyy:aabbaiyy
‘pasturing’, ‘dropping’, ‘pay’, ‘bee’

4.1 Categorical criterion

Binary relations occurring in the formal analogies are primarily typed on the basis of a categorical criterion that can be described by the following two propositions. Notice however that the second proposition is valid only for analogies where familial relations do not involve prefixation.

- Two words belonging to the same series have the same category.

- In a morphological analogy, two words with different categories belong to the same morphological family.

This criterion assigns the analogies to three types: F (for *family*) when (*a*, *b*) and (*c*, *d*) are words of the same family, S (for *series*) when they belong to the same series and U (for *undetermined*) when all four words have the same category. It can help us type analogies such as (9a) and (9b) but not (9c) and (9d). Edge types can also be determined indirectly as for (9c).

(9) a. *fructificateur*$_{\text{N}}$:*fructifier*$_{\text{V}}$ = *rectificateur*$_{\text{N}}$:*rectifier*$_{\text{V}}$ (type F)
'which brings benefits', 'bring benefits', 'rectificator', 'rectify'
b. *maçonnage*$_{\text{N}}$:*poinçonnage*$_{\text{N}}$ = *maçonner*$_{\text{V}}$:*poinçonner*$_{\text{V}}$ (type S)
'walling', 'stamping', 'build', 'stamp'
c. *développeur*$_{\text{N}}$:*développement*$_{\text{N}}$ = *envelopppeur*$_{\text{N}}$:*enveloppement*$_{\text{N}}$
(type U)
'developer', 'development', 'wrapper', 'envelopment'
d. *adapté*$_{\text{A}}$:*adaptable*$_{\text{A}}$ = *inadapté*$_{\text{A}}$:*inadaptable*$_{\text{A}}$ (type U)
'adapted', 'adaptable', 'inadequate', 'which cannot be adapted'

Indeed, other analogies such as the ones presented in (10) allow us to assign the F-type to *développeur*:*développer* and to *developpement*:*développer*, and therefore place *développeur* and *developpement* in the same family. This edge is thus typed transitively.

(10) a. *développeur*$_{\text{N}}$:*développer*$_{\text{V}}$ = *enveloppeur*$_{\text{N}}$:*envelopper*$_{\text{V}}$ (type F)
b. *développement*$_{\text{N}}$:*développer*$_{\text{V}}$ = *enveloppement*$_{\text{N}}$:*envelopper*$_{\text{V}}$
(type F)

Its F-type can then be extended to the second couple in (9c) *enveloppeur*:*enveloppement*. The analogy (9d) presents an additional challenge because all its binary relations have both types.

For analogies where the familial relations involve prefixation as in (11), the categorical criterion is inoperative or may predict a wrong type.

(11) a. *trouver*$_{\text{V}}$:*troussage*$_{\text{N}}$ = *retrouver*$_{\text{V}}$:*retroussage*$_{\text{N}}$ (type F)
'find', 'trussing', 'recover', 'sleeves rolling up'[8]

For (11), the predicted F-type is wrong: *trouver* and *troussage* do not belong neither to the same family nor to the same series. Indeed, being bases in the same prefixation is not enough to make two words belong the same family. Notice however that most of the errors induced by the categorical criterion

[8] *Troussage* and *retroussage* are highly polysemous nouns. I just picked up one of their meanings.

are eliminated by the frequency criterion presented below; prefixation relations being rarer than suffixation ones because of the lower number of prefixed derivatives in the lexicon.

4.2 Filtering out the errors

Morphonette is destined to become a resource for French NLP and as such it must contain the least possible errors. Three criteria were used for this purpose.

4.2.1 Criterion 1

The first is to remove the analogies of the type (8c) repeated in (12) in order to keep only the quadruples for which there exists an analogy for both their written forms and their phonetic transcriptions.

(12) *paissant*:*abaissant* = *paye*:*abeille*
ppaissan:aabbaissan = ppaiyy:aabbaiyy
'pasturing', 'dropping', 'pay', 'bee'

4.2.2 Criterion 2

The second criterion concerns the size of the series which normally contain many members, as illustrated by the example (13) which presents an excerpt of the series of the name *siffleur* 'whistler'.[9] As a result, if *x* belongs to the family of *w*, then the binary relation *w*:*x* must occur in a large number of analogies (normally, one for each member of the series of *w*).

(13) *hâbleur*, *haleur*, *engueuleur*, *enrouleur*, *entauleur*, *entôleur*, *branleur*, *chialeur*, *dégonfleur*, *dribbleur*, *effileur*, *épileur*, *fileur*, *flirteur*, *footballeur*,

[9] Translation of the examples: *hâbleur* 'boaster', *haleur* 'hauler', *engueuleur* 'someone who likes to give hell', *enrouleur* 'recruiter', *entauleur* 'bad paying client', *entôleur* 'crook', *branleur* 'wanker', *chialeur* 'sniveler', *dégonfleur* 'deflator', *dribbleur* 'dribbler', *effileur* 'threder', *épileur* 'epilator', *fileur* 'spinner', *flirteur* 'player', *footballeur* 'football player', *frôleur* 'someone who brushes by', *gifleur* 'slapper', *jongleur* 'juggler', *caleur* 'lazy worker', *camoufleur* 'backstage dresser', *colleur* 'gluer', *crawleur* 'crawl swimmer', *cribleur* 'sifter', *miauleur* 'meower', *ourleur* 'hemmer', *parfileur* 'someone who unweaves a fabrics in order to recover the gold or silver threads', *parleur* 'speaker', *persifleur* 'mocker', *pileur* 'looter', *rafleur* 'someone who snaps up something', *recéleur* 'receiver of stolen goods', *ronfleur* 'buzzer', *rouleur* 'worker who moves loads', *siffloteur* 'whistler', *ciseleur* 'engraver', *souffleur* 'blower', *trembleur* 'shaker', *trimbaleur* 'someone who trails somebody around', *trimballeur* 'someone who trails somebody around', *troubleur* 'disturber', *hurleur* 'howler', *vitrioleur* 'someone who splashes a person with sulfuric acid to deface or kill her', *voleur* 'thief', *vielleur* 'watcher', *violeur* 'rapist'.

frôleur, *gifleur*, *jongleur*, *caleur*, *camoufleur*, *colleur*, *crawleur*, *cribleur*, *miauleur*, *ourleur*, *parfileur*, *parleur*, *persifleur*, *pileur*, *rafleur*, *recéleur*, *ronfleur*, *rouleur*, *siffloteur*, *ciseleur*, *souffleur*, *trembleur*, *trimbaleur*, *trimballeur*, *troubleur*, *hurleur*, *vitrioleur*, *voleur*, *vielleur*, *violeur*

For example, if (8b) was correct, *foyère* would belong to the family of *foisonner* and the relation *foisonner*:*foyère* would be part of large number of analogies. But this relation does not occur in any other analogy, allowing for the elimination of (8b). Specifically, the frequency criterion is used as follows: if a binary relation occurs in at least 10 analogies, it is likely to be correct and to connect two words of the same family. Several values of the threshold have been tested and 10 proved to be a good compromise between precision (i.e. it eliminates the maximum number of erroneous relations) and recall (i.e. it removes the minimum number of correct relations). Precision and recall were estimated manually.

This test has a dual function. It allows to type some of the relations for which the categorical criterion cannot decide. It also eliminates some of the wrong analogies: if an F-typed relation does not meet the minimum frequency requirement, all the analogies where it occurs are removed from the set *A*. The operation is iterated until all the F-typed relations pass the test.

The dual property is that usually families are small sets as illustrated by the example (14). However, a criterion on family size would be useless because series can be small too. Such a criterion would lead us to consider these series as families, for example in the case of a compound such as *thyroïdite* ‘thyroiditis’ which belong to a series of seven words only (15).

(14) *Sifflable* ‘whistlable’, *sifflade* ‘catcalls’, *sifflement* ‘whistling’, *sifflerie* ‘whistling’, *siffler* ‘whistle’, *siffleuse* ‘female whistler’

(15) *Adénoïdite* ‘adenoiditis’, *arachnoïdite* ‘arachnoiditis’, *choroïdite* ‘choroiditis’, *mastoïdite* ‘mastoiditis’, *parotidite* ‘parotitis’, *péricardite* ‘pericarditis’, *thyroïdite* ‘thyroiditis’

4.2.3 Criterion 3

The third criterion is structural in nature. It relies on the fact that series are dense regions in the morphological graph. Indeed, if two words y_1 and y_2 belong to the series of a word *w*, one can expect that y_1 also belongs to the series of y_2 and y_2 to that of y_1. In other words, morphological graphs are small worlds in the sense of Watts & Strogatz (1998).

This criterion is used to homogenize the subseries of compounds such as *zoomorphie* ‘zoomorphy’ which contain both words formed with the first compound element as *anthropomorphie* ‘anthropomorphy’ and others with the second as *zoologie* 'zoology'. Therefore, this subseries mixes up two types of morphological relations: composition in the first case and *-ie* suffixation in the second. The effect of this third criterion is to eliminate from these subseries the words that correspond to the minority relation. In the example above, it eliminates the *-ie* derivatives such as *zoologie* or *zoophagie* 'zoophagy'.

5. FILAMENTS

The techniques and criteria I have just presented have been used to create a first version of the French morphological network that includes 29,310 entries, 96,107 relations between members of the same family and 1,160,098 relations between members of the same series.[10]

The construction of the Morphonette revealed a paradigmatic organization of the lexicon slightly more complex than that presented in Figure 1. In this organization, families can be seen as sets of words expressing the same ideas or concepts and series as sets of words showing the same morphological regularities and sharing the same properties. Since words normally have multiple properties, they are likely to be part of several paradigms, or rather several subseries.

For example, a form as *gazouillard* 'babbler' is both a noun that has a corresponding feminine *gazouillarde* 'female babbler' and a derivative of the verb *gazouiller* 'babble'. The first property includes *gazouillard* in a subseries of nouns that have corresponding feminines in /aʀd/ as in (16)[11] and the second in a subseries of *-ard* deverbal as in (17).

(16) *babillard*, *becquillard*, *béquillard*, *braillard*, *douillard*, *égrillard*, *fripouillard*, *gaillard*, *grenouillard*, *grognard*, *justiciard*, *montagnard*, *prétentiard*, *savoyard*, *citrouillard*, *trouillard*, *vadrouillard*, *vasouillard*

(17) *bafouillard* 'mumbler', *douillard*, *grenouillard*, *citrouillard*, *vadrouillard*, *vasouillard*, *ventrouillard* 'pot-bellied'

The two subseries are different. The first contains all the words of the

[10] This resource is licensed under the Creative Commons and is available at the following address: http://redac.univ-tlse2.fr/lexiques/morphonette.html.

[11] Translation of the examples: *babillard* 'talkative', *becquillard* 'someone on crutches', *béquillard* 'someone on crutches', *braillard* 'yelling', *douillard* 'rich person', *égrillard* 'dirty-minded', *fripouillard* 'crook', *gaillard* 'strapping lad', *grenouillard* 'someone who looks like a frog', *grognard* 'grunter', *justiciard* 'magistrate', *montagnard* 'highlander', *prétentiard* 'pretentious', *savoyard* 'man from Savoy', *citrouillard* 'someone with a head like a pumpkin', *trouillard* 'coward', *vadrouillard* 'rolling stone', *vasouillard* 'clumsy'.

second. But it also includes nouns for which there exists a feminine in /aʀd/ but no corresponding verb as *gaillard* ‘stapping lad’, *montagnard* ‘highlander’, *trouillard* ‘coward’ or *savoyard* ‘man from Savoy’. Also note that each of these subseries can be identified by a member of the family of *gazouillard*: *gazouillarde* for the first and *gazouiller* for the second.

These observations led me to propose a new data structure that I call filament. A filament is a triple consisting of an entry w, an element x of his family and a subseries $s(w, x)$ such that, for all $y \in s(w, x)$, there exists z such that $w{:}x = y{:}z$.

If the subseries of some of the filaments of a same word may differ, others are identical. For example, *gazouiller* should have the same subseries with respect to *gazouillard* and *gazouillarde*. The filament structure will therefore have to be generalized in a future version of the Morphonette network so that it associates with each entry a subfamily and a subseries. The filament structure is thus an intermediary level of organization which reflects the fact that families and series are themselves composite structures.

6. RELATED WORKS

From a theoretical point of view, this work is situated in a framework related to the Network Morphology of Bybee (1995), to the Surface-to-Surface Morphology of Burzio (2002), and to emergentist approaches of Aronoff (1994), Albright (2002) or Goldsmith (2006).

The construction of Morphonette uses a technique similar to that of Goldsmith (2006) or Bernhard (2006). The main differences with these ones is that it is fully lexeme-based and does not make use of morpheme nor contain any representation of them. Morphological regularities emerge directly from a very large set of analogies. Collecting them is one of the contributions of the work presented here. It was made possible through the use of the measure of morphological similarity proposed in Hathout (2008). This measure, inspired by work on small worlds done by Gaume *et al*, (2002), avoids word decomposition. In this respect, it could be compared to the experiments of Yarowsky & Wicentowski (2000) and Baroni *et al*, (2002) where the Levenshtein string edit distance is used to identify formal similarity between words. The present research is also close to the ones by Langlais *et al*, (2009) and Lavallée & Langlais (2009) who use formal analogies to analyze words morphologically and to translate them.

The Morphonette network could also be compared to the morphological families constructed by Xu & Croft (1998), Gaussier (1999) or Bernhard (2009) among others. With respect to these methods, the main contribution of Morphonette is the generation of a huge collection of formal analogies and the exploitation of the structural properties of the morphological graph in order to set apart the familial and the serial relations.

7. CONCLUSION

This research program is still in a preliminary stage. Many improvements are needed and many developments are to be made. In the near future, I plan to extend this research in two directions. The first is to improve the coverage of the French Morphonette which currently contains about 30% of TLFi headwords. A bootstrapping method will be used to achieve this goal. Moreover, a larger coverage will not be achieved without adding to the method some semantic knowledge and using the dictionary macro-structure. Both will be help us improve the measure of proximity morphological presented in section 2.1, and better characterize the derivational families and series.

I also plan to extend this research to other Romance languages which like French do not yet have large coverage morphological databases, but also to build a morphological network for English with the aim of evaluating the construction method. This network will be compared to the English part of the CELEX database, which provides detailed morphological descriptions for a significant fragment of the English lexicon.

REFERENCES

Albright, A. (2002). *The identification of bases in morphological paradigms*. PhD thesis, University of California, Los Angeles.

Anderson, S. R. (1992). *A-Morphous Morphology*. Cambridge, UK: Cambridge University Press.

Aronoff, M. (1994). *Morphology by Itself. Stem and Inflectional Classes*. Linguistic Inquiry Monographs. Cambridge, MA: MIT Press.

Baayen R. H., Piepenbrock R. & Gulikers L. (1995). *The CELEX lexical database (release 2)*. Philadelphia, PA: Linguistic Data Consortium.

Baroni, M., Matiasek, J. & Trost, H. (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-2002* (pp. 48–57). Philadelphia, PA: ACL.

Béchet, F. (2001). LIA_PHON : un système complet de phonétisation de textes. *Traitement automatique des langues*, 42 (1), 47–67.

Becker, T. (1993). Back-formation, cross-formation, and 'bracketing paradoxes' in paradigmatic morphology. *Yearbook of Morphology* 1992, 1–27.

Bernhard, D. (2006). Automatic acquisition of semantic relationships from morphological relatedness. In *Advances in Natural Language Processing, Proceedings of the 5th International Conference on NLP, FinTAL 2006*, volume 4139 of Lecture Notes in Computer Science (pp. 121–132). Berlin / Heidelberg: Springer Verlag.

Bernhard, D. (2009). Morphonet: Exploring the use of community structure for unsupervised morpheme analysis. In *Working Notes for the MorphoChallenge at CLEF 2009*. Corfu.

Booij, G. (1997). Autonomous morphology and paradigmatic relations *Yearbook of Morphology* 1996, 35–53.

Booij, G. (2008). Paradigmatic Morphology. In B. Fradin (Ed.), *La raison morphologique. Hommage à la mémoire de Danielle Corbin* (pp. 29–38). Amsterdam / Philadelphia: John Benjamins.

Burzio, L. (2002). Surface-to-surface morphology: when your representations turn into constraints. In P. Boucher (Ed.), *Many Morphologies* (pp. 142–177). Somerville, MA: Cascadilla Press.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, J. L. (1995). Regular morphology and the lexicon. *Language and cognitive processes*, 10 (5), 425–455.

Gaume, B., Duvigneau, K., Gasquet, O. & Gineste, M.-D. (2002). Forms of meaning, meaning of forms. *Journal of Experimental and Theoretical Artificial Intelligence*, 14 (1), 61–74.

Gaussier, E. (1999). Unsupervised Learning of Derivational Morphology from Inflectional Lexicons. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing* (pp. 24–30). College Park, MD: ACL.

Goldsmith, J. (2006). An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12 (4), 353–371.

Grandi, N. & Montermini, F. (2010). *DADI: Dizionario degli affissi derivazionali italiani.* Talk given at the "Morphology meets computational linguistics" workshop. Bologna.

Hathout, N. (2005). Exploiter la structure analogique du lexique construit : une approche computationnelle. In *Cahiers de Lexicologie*. 87 (2), 5–28.

Hathout, N. (2008). Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *Proceedings of the 3^rd^ Textgraphs Workshop on Graph-Based Algorithms for Natural Language* (pp. 1–8). Manchester: ACL.

Hathout, N. (2009a). Acquisition of morphological families and derivational series from a machine readable dictionary. In F. Montermini, G. Boyé, and J. Tseng (Eds.) *Selected Proceedings of the 6^th^ Décembrettes: Morphology in Bordeaux* (pp. 166-180), Cambridge, MA: Cascadilla Proceedings Project.

Hathout, N. (2009b). *Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie*. HDR thesis. Université de

Toulouse.

Jacquemin, C. (1997). Guessing Morphology from Terms and Corpora. *Proceedings of 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)* (pp. 156–167). Philadelphia, PA: ACM.

Kurimo, M., Virpioja, S., Turunen, V. & Lagus K. (2010). Morpho Challenge 2005-2010: Evaluations and Results, *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology* (pp. 87–95). Uppsala: ACL.

Langlais, P., Yvon, F. & Zweigenbaum, P. (2009). Improvements in analogical learning: Application to translating multi-terms of the medical domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)* (pp. 487–495). Athens.

Lavallée, J.-F. & Langlais, P. (2009). Morphological acquisition by formal analogy. In *Working Notes for the MorphoChallenge at CLEF 2009*. Corfu.

Lepage Y. (2003). *De l'analogie rendant compte de la commutation en linguistique.* HDR thesis, Université Joseph Fourier, Grenoble.

Schreuder, R. & Baayen, R. (1997). How complex simplex words can be. *Journal of Memory and Language*, 37, 118–139.

Stroppa, N. (2005). *Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles*. PhD thesis, École Nationale Supérieure des Télécommunications, Paris.

Stroppa, N. & Yvon, F. (2005). An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)* (pp. 120–127). Ann Arbor, MI: ACL.

Stump, G. (2001). *Inflectional Morphology*. Cambridge, UK: Cambridge University Press.

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393, 440–442.

Xu, J. & Croft, W. B. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transaction on Information Systems*, 16 (1), 61–81.

Yarowsky, D. & Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the Association of Computational Linguistics (ACL-2000)* (pp. 207–216). Hong Kong: ACL.

Chomsky, N. (1970). Remarks on nominalization. In R. A. Jacobs & P. S. Rosenbaum (Eds.), *Readings in English Transformational Grammar* (pp. 232-286). Waltham: Ginn and Co.

Carlson, G. (1977). *Reference to Kinds in English*. Ph.D. dissertation. Amherst, MA: University of Massachusetts.

Comrie, B. (1989). *Language Universals and Linguistic Typology*, 2nd edition. Oxford: Blackwell.

Jackendoff, R. (1990). *Semantic Structures*. Cambridge, MA: MIT Press.

Hopper, P. J. & Thompson, S. A. (1984). The discourse basis for lexical categories in universal grammar. *Language* 60 (4), 703–752.

Name Surname
Affiliation
Full address
Country
e-mail: xxx@xxx.xx