Morphological Pairing based on the Network Model*

Nabil Hathout
CNRS - ERSS
Nabil.Hathout@univ-tlse2.fr

Abstract

The paper presents a method of semi-automatic construction of a lexicon enhanced with constructional informations. The construction is carried out in two steps: (*i*) automatic pairing of the constructed words with their base words, then (*ii*) manual validation of these base words by human operators. The method is based on the Network Model proposed by J. Bybee, which gives a central role to analogy and lexical memory (i.e. the mental lexicon): the lexicon is viewed as a network of fully inflected forms connected to each other by relations set up according to shared semantic and phonological features. Connections are organized into paradigms corresponding to analogical schemata. The automatic pairing of the constructed words with their bases takes place in three steps: (*i*) unsupervised learning of analogical schemata, (*ii*) association of a set of candidate bases with each constructed word, and (*iii*) filtering of these candidates by means of a measure (ex. type frequency). The method can be used for any language with concatenative morphology since no linguistic knowledge is used. Moreover, the only resource needed is a lexicon of inflected forms. An evaluation of the method is presented in § 4.

1. Network Model

Analogy plays a central role in natural language morphology from the inflexional point of view as well as the constructional one:[1] it is the basic mechanism of the paradigmatic organization of the lexicon. On the other hand, word formation primarily depends on the existing lexicon, that is on the forms it contains and the connections that hold between them. For a given speaker, the analysis of a constructed (i.e. derived) word is performed in relation to the units of its mental lexicon and the semantic and phonological relations that connect them. This view of the lexicon corresponds to the "Network Model" proposed by J. Bybee [4, 5, 6]. As shown by L. Burzio [2, 3], this model is fully consistent with the Optimality Theory framework. The lexicon is viewed as a network of fully inflected forms connected to each other by relations set up according to shared semantic and phonological features. Connections between lexical units are of different strength: the degree of relatedness of two units depends on the type and the number of the semantic features they share (phonological similarities do not intervene directly even if it usually parallel the degree of semantic relatedness).

Bybee's model is purely representational. Connections are organized into paradigms (or schemas). An affix such as *-able* is regarded as a set of connections between the units of a proportional series. This set gathers on the one hand links between the adjectives ending in *-able* (eg. activ**able**:agit**able** 'activable:agitable'; accept**able**:admir**able** 'acceptable:admirable') and that share this segmental material and the semantic properties that can be paraphrased as "capable of, fit for or worthy being..." and on the other links between these adjectives and verbs or nouns (eg. **activ**able:**acti**ver 'to activate'; **corvé**able:**corvé**e 'liable to fatigue:fatigue') that share the segmental material corresponding to the radical (eg. *activ-*; *corvé-*) and the semantic properties of the verb or noun (eg. "make active"; "fatigue"). The productivity of a schema (its capacity to include new items) depends on its consistency with general phonological constraints such as the Obligatory Contour Principle (OCP), on the number of links it gather (i.e. its type frequency) and on its "validity cue" (i.e. the ratio between type frequency and the number of items compatible with the schema).

2. Lexicon enhanced with constructional informations

The paper is devoted to the application of the Network Model to morphological pairing in order to build semi-automatically a lexicon for NLP and IR (hereafter LECI, lexicon enhanced with constructional informations). This work is part of the research project "MorTAL" (Morphology for NLP) [9, 10] which gathers G. Dal, Ch. Jacquemin, F. Namer and the author.

These informations could be used in several NLP and IR tasks such as knowledge acquisition from corpora or document retrieval [8]. For instance, they may be

*The present work is funded by the Ministère de l'Éducation Nationale, de la Recherche et de la Technologie (*French Ministry of Education, Research and Technology*) as part of the program Actions Concertées Incitatives 1999 (*Concerted Incitement Action*).

[1] We follow D. Corbin by prefering the term "constructional" to "derivational" because it is more explicit than the latter and rather neutral from a theoretical point of view ("derivation" implicitly implies the existence of rules, representation levels...).

entry	base words
acceptable/Afpms	accepter/Vmn----
activation/Ncfs	activer/Vmn----
activer/Vmn----	actif/Afpms
affable/Afpms	
imperturbable/Afpms	perturber/Vmn----
inacceptable/Afpms	acceptable/Afpms
inactivable/Afpms	inactiver/Vmn---- activable/Afpms
recouvrable/Afpms	recouvrer/Vmn---- recouvrir/Vmn----

Table 1: Some entries of the LECI.

used to identify morphological variations such as *activer un processus*:*un processus activable* 'to activate a process:an activable process' or *un processus actif*:*l'activité du processus* 'an active process:the process activity on the basis of the relations *activer* → *activable* and *actif* → *activité*. The LECI entries are the lemmata of a French lexicon; constructed word entries give their base words (table 1); non-constructed words and words with foreign or infralexical bases (eg. *affable* 'affable') have empty entries. Some constructed words may have several bases that correspond to different meaning of these words. This homonymy may be due either to different schemata connecting the constructed word with different bases (for instance a prefixation and a suffixation as in the case of *inactivable* 'inactivable': [in- activable] "not activable"; [inactiver -able] "capable of being inactivated") or to the same schema connecting the constructed word with different bases as for *recouvrable* which means either 'recoverable' or 'coverable'.

3. Morphological pairing

In this paper, we are interested in the identification of individual schemata, seen as bipartite subgraphs of the entire lexicon, composed of links connecting on the one hand words formed with a particular affix (eg. *-able*, *-ité*, *-iser*...) and on the other their bases.

3.1. Learning of analogical schemata

In the Network Model, lexical units are attested words which could be approximated by the set of inflected words of a lexicon of reference. In this study, we have used TLFnome96, a lexicon built from the *Trésor de la Langue Française* (TLF) word list, extended with TLFindex99 built from the TLF index. This reference lexicon includes 735 000 entries for 97 000 lemmata. The LECI is a subgraph of the lexicon that only contains links corresponding to morphological constructions, that is to a parallel sharing of semantic and phonological features. This sharing is identified on the basis of the written forms only by means of an unsupervised learning of analogical schemata. For instance, when exposed to couples as *activer*/Vmn----:*activable*/Afpms, *agiter*/Vmn----:*agitable*/Afpms..., the learning program, `trouvaffix`, induces a suffixation schema *er*/Vmn----:*able*/Afpms that could be used to link *achetable*/Afpms 'buyable' and *acheter*/Vmn---- 'to buy'. Similarly, from couples like *inaccordable*/Afpms:*accordable*/Afpms 'grantable:ungrantable', *inaliénable*/Afpms:*aliénable*/Afpms 'alienable:unalienable'..., it induces a prefixation schema *in*/Afpms:/Afpms that enables *inacceptable*/Afpms 'unacceptable' and *acceptable*/Afpms to be paired. Prefixation and suffixation schemata are learned separately with an algorithm similar to the one of `findaffix` (script of the `ispell` checker). Schemata learning relies on two assumptions:

1. The longer the form, the stronger the correspondence between written form and semantic is. As consequence, if two forms share a sufficiently long common prefix (or suffix), they are very likely to be semantically connected.

2. The lexical frequency of a morphological relation is an index of its regularity, the latter being a gage of the rule validity.

Schemata corresponding to a given suffix (eg. *-able*) are learned on $C \times B$ where C is the set of words supposed to be constructed (eg. the adjectives ending in *-able*) and B is a set of words of the same categories as their supposed bases (eg. verbs and nouns). Members of B may be either inflected forms or lemmata. For a prefix (eg. *in-*), learning takes place on $C \times B'$ where B' is the set of words of the same category as the ones in C (only homocategorical prefixes are considered). Notice that some constructed words are both prefixed and suffixed (eg. *imperturbable* 'unperturbable'). We then have to learn prefixation schemata even when we are dealing with a set C of words supposed to be suffixed. In this case, learning takes place on $B' \times B'$ since the prefixations that apply to the words in C are not specific to C but concern the whole set of words of the same category. For instance, the prefixations found in adjectives ending in *-able*, may apply to all adjectives. $B' \times B'$ is preferred to $C \times B'$ because it has a much greater size and yields more accurate frequency values.

Our learning program differs from `findaffix` on two points: (*i*) it takes into account the morphosyntatic categories of the words, which enhances significantly the pairing precision; (*ii*) it provides for each schema some additional informations intended to evaluate its validity (validity cues, distribution according to the size of the bases...). Moreover, the objective being to bring out the morphological paradigmatic structure of the lexicon, `trouvaffix` computes the set of schemata once for all and not on demand, as in other researches on analogy based computational morphology [11, 13].

3.2. Schema application

Learned refixation and suffixation schemata are utilized to connect the words in C to words in B or B'. For a suffix as *-able*, C is the set of adjectives ending in *-able* and B is the set of inflected forms of verbs and nouns. The resort to inflected forms instead of the supposed bases lemmata is justified from a theoretical point of view by the fact that lexical units are attested words and that these only occur in inflected form. It is also justified from a practical point of view because it gives access to the verb long radicals. Adjectives ending in *-able* are precisely constructed on these long radicals which can, for instance, be found in present participles (*finissant*:*finissable* 'finishing:finishable', *faisant*:*faisable* 'doing:feasible...). We can also see in table 1, that these adjectives can be only suffixed (*accepter*:*acceptable*), or only prefixed (*acceptable*:*inacceptable*) or both prefixed and suffixed (*perturber*:*imperturbable*, 'to perturb:unperturbable') because °*perturbable* 'perturbable' is not attested although it is morphologically well-formed. Adjectives ending in *-able* may then have other adjectives ending in *-able* as base words (second case); in the third case, the base being a verb or a noun, the categorical constraints of the prefix enforce an interpretation where the suffix is inside the prefix scope, for instance [in-[pertuber -able]].

The application of the learned schemata yields the connections needed to pair these three types of constructed words with their bases. It is carried out by the `applicaffix` program which links every word c in C to the subset of words b from B and from B' such that there exists a prefixation and/or a suffixation schema kept by `trouvaffix` and of which $b : c$ is an instance. The candidate bases associated with each $c \in C$ are then sorted in order to only retain the most likely one.

3.3. Selection of the best candidate base

Four measures have been used to compare the candidate bases b associated with a word $c \in C$: (*i*) freq_σ, the type frequency of the schema σ that connects b and c; (*ii*) pointwise mutual information $MI(\sigma)$ [7]; (*iii*) its variant $MI^3(\sigma)$:

$$MI(\sigma) = \frac{\text{freq}_\sigma}{\text{freq}_\text{S} \times \text{freq}_\text{a}} \qquad MI^3(\sigma) = \frac{\text{freq}_\sigma^3}{\text{freq}_\text{S} \times \text{freq}_\text{a}}$$

where freq_S is the type frequency of the affix stripped from the base word and freq_a is the type frequency of the affix added to the radical in order to get the constructed word.[2] The fourth measure is relative to schema orientation. The schemata learned by `trouvaffix` are not oriented in the sense that it keeps schemata that (*i*) link words from C to their bases in B (eg. *respecter*:*respectable* 'to respect:respectable'), (*ii*) link constructed words from B to their bases in C (eg. *rentabiliser*:*rentable* 'to make profitable:profitable') and (*iii*) link words that only share a common radical (eg. *respiration*:*respirable* 'breathing:breathable'). A quite simple heuristics that predicts schema orientation ($b \rightarrow c$ in the first case and $b \leftarrow c$ in the second one) consists in comparing the size of the affixes stripped from b and added to get c. We also observed that in both cases, the orientation can be predicted by comparing the validity cues for the base words, $V_\text{S}(\sigma)$ and for the constructed words $V_\text{a}(\sigma)$; $V_\text{S}(\sigma) < V_\text{a}(\sigma)$ if the orientation is $b \rightarrow c$ and $V_\text{S}(\sigma) > V_\text{a}(\sigma)$ if it is $b \leftarrow c$, where:

$$V_\text{S}(\sigma) = \frac{\text{freq}_\sigma}{\text{freq}_\text{S}} \qquad V_\text{a}(\sigma) = \frac{\text{freq}_\sigma}{\text{freq}_\text{a}}$$

Both functions take their valued in $]0, 1]$. We also observe that the closer $V_\text{S}(\sigma)$ to 0 and $V_\text{a}(\sigma)$ to 1, the more valid $b \rightarrow c$ schemata are. We then defined a measure $SO(\sigma)$ of the strength of the schema orientation:

$$SO(\sigma) = \frac{\log(V_\text{S})}{\log(V_\text{a})}$$

$SO(\sigma)$ expresses that the less the schema applies in the bases set and the more it applies in the constructed words set, the more strong the schema orientation is.

4. Evaluation

The morphological pairing method has been evaluated against a fragment of the LECSI (where S stands for semantic) that G. Dal and F. Namer [12] are building as part of the MorTAL project, by means of a finely hand tuned morphological analyzer (that uses exception lists). Their analysis include a complete constructional decomposition of the constructed words, an exhaustive handling of foreign and infralexical bases and a gloss of the constructed word meaning. However, with a very few exceptions, when a constructed word has more than one base, only one of them is provided. This fragment consists of 3 suffixes: *-able* that constructs adjectives from verbs and nouns, *-ité* (cognate to English *-ity*) that constructs nouns from adjectives and nouns, and *-iser* (cognate to English *-ise*) that constructs verbs from adjectives and nouns.

The evaluation takes into account the fact that [12] use a lexicon of reference different from ours and that some constructed words have infralexical bases. It then only bears on the entries of the fragment that belong to $C \cap C_F$ which have their base words in $B \cap B_F$ where C_F is the set of the fragment entries and B_F is the set of their bases. Moreover, constructed words with infralexical bases are regarded as non-constructed words. The evaluation concerns first the morphological pairing with unsupervised learned schemata and second the usefulness of some of the theoretical hypothesis of Bybee's model as the role of type frequency or validity cues.

[2] All the schemata being learned on the same C and B word corpora, we ignore the number of words of these corpora when computing the MI values, since these numbers are constant factors that do not affect the candidate order.

Schema orientation		−size −V.C.	+size −V.C.	−size +V.C.	+size +V.C.
freq_σ	recall	91.3%	94.5%	93.7%	94.4%
	precis.	90.6%	94.0%	93.2%	94.1%
$MI(\sigma)$	recall	21.6%	46.9%	40.4%	61.4%
	precis.	21.5%	46.6%	40.2%	61.7%
$MI^3(\sigma)$	recall	83.9%	94.2%	91.8%	94.4%
	precis.	83.3%	93.6%	91.3%	94.0%
$SO(\sigma)$	recall	95.5%	95.4%	95.5%	95.4%
	precis.	95.3%	95.6%	95.7%	95.8%

Table 2: Recall and precision of the pairing of 1182 adjectives ending in *-able*; $E = 0,78$. The schemata are filtered according to their orientation defined in terms of affix sizes and validity cues: [−size] = no filtering according to the size; [+size] = filtering according to the size; [−V.C.] = no filtering according to validity cues; [+V.C.] = filtering according to validity cues.

The paring quality depends mainly on the homogeneity of the set where the bases are looked for. The paring system achieves very satisfactory results for the *-able* suffix (see table 2) since its bases are mainly verbs and because French inflected verbs and nouns are well differentiated. The results are not as good for *-ité* (precision is around 91% for a 89% recall with freq_σ) because, even if the bases are mainly adjectives, their forms are not easily distinguishable from nouns forms. For *-iser*, the pairing is insufficient (precision is around 54% for 57% recall with $SO(\sigma)$) because of the more even distribution of its bases between nouns and adjectives; in this case, deciding between them has to rely on semantic informations. Table 2 also shows that orientation based on validity cues is almost always outperformed by an orientation based on the size of the schema affixes and that the less efficient the measures, the more useful this orientation is (as for MI).

5. Conclusion

This study has proposed a method to build semi-automatically a constructional lexicon. This method presents several strong points: it uses an unsupervised learning; it only takes as input a lexicon of inflected forms; it does not make use of any linguistic knowledge which makes it independent from individual languages; production and validation of the pairings are separated which allow the latter be carried out by persons with no particular computational abilities. This study also confirmed that type frequency and validity cues are strongly correlated to schema validity.

References

[1] P. Boucher, editor. *Morphology book*. Cascadilla, Cambridge, Mass., to appear.

[2] L. Burzio. Multiple correspondence. *Lingua*, 1998 (104):79–109, 1998.

[3] L. Burzio. Surface-to-surface morphology: when your representations turn into constraints. Talk given at the "1999 Maryland Mayfest", University of Maryland, College Park, Aug. 1999.

[4] J. L. Bybee. *Morphology. A Study of the Relation between Meaning and Form*, volume 9 of *Typological Studies in Language*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 1985.

[5] J. L. Bybee. Morphology as lexical organization. In M. Hammond and M. Noonan, editors, *Theoretical Morphology. Approaches in Modern Linguistics*, chapter 7, pages 119–141. Academic Press, San Diego, CA, 1988.

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

[7] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, may 1990.

[8] B. Daille, C. Fabre, and P. Sébillot. Applications of computational morphology. In Boucher [1].

[9] G. Dal, N. Hathout, and F. Namer. Construire un lexique dérivationnel : théorie et réalisation. In P. Amsili, editor, *Actes de la VI^e conférence sur le Traitement Automatique des Langues Naturelle (TALN'99)*, pages 115–124, Cargèse, Corse, jul 1999. ATALA.

[10] N. Hathout, F. Namer, and G. Dal. An experimental constructional database : The mortal project. In Boucher [1].

[11] Y. Lepage. Solving analogies on words: an algorithm. In *Proceedings of the of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, volume 2, pages 728–735, Montréal, Canada, Aug. 1998.

[12] F. Namer and G. Dal. Gédérif: Automatic generation and analysis of morphologically constructed lexical resources. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, may 2000. ELRA.

[13] V. Pirrelli and F. Yvon. The hidden dimension: a paraadigmatic view of data-driven nlp. *Journal of Experimental & Theoretical Artificial Intelligence*, 1999(11):391–408, 1999.