From WordNet to CELEX: acquiring morphological links from dictionaries of synonyms

Nabil Hathout

ERSS – CNRS & Université de Toulouse Le Mirail – France
5, allées Antonio Machado. F-31058 Toulouse CEDEX 1
Nabil.Hathout@univ-tlse.fr

Abstract

Morphological resources such as CELEX do not exist for many languages. NLP and RI systems that operate on texts and documents written in these languages have then to rely on morphological resources acquired from lexica or corpora. These resources usually suffer from a problem of precision because no *a priori* semantic knowledge is used for their acquisition. The paper proposes a robust and language independent technique to acquire morphological constructional relations from dictionaries of synonyms. The idea is to explore simultaneously synonymy and morphological relations in order to make more accurate prediction. The paper presents an evaluation of the technique and a comparison of the acquired morphological links with the CELEX database.

1. Constructional morphology for NLP and IR

In the last decade, the interest for morphology and especially constructional[1] morphology has been growing in theoretical linguistics and in computational linguistics, related domains as information retrieval (IR). Some recent experiments in IR (Xu and Croft, 1998; Jacquemin and Tzoukermann, 1999) have shown that constructional morphology can contribute to improve the efficiency of IR systems. For highly inflected languages as French, a proper treatment of inflexional morphology is imperative (Namer, 2000). This is less the case for for poorly inflected languages (Krovetz, 1993).

Word formation is commonly regarded as lexical. For instance, (Bybee, 1988; Bybee, 1995) develops a theory where the lexicon is viewed as a network of lexical items (eg. fully inflected forms) connected to each other by relations set up according to shared semantic and phonological features.

From a computational point of view, word formation can be dealt with in two ways:

- by means of a morphological analyzer such as Englex (Antworth, 1990) for English (based on the two level model) or DeriF (Dal et al., 1999; Namer and Dal, 2000) for French (based on the SILEX model). This solution has many limitations: it is expensive; lot of linguistic knowledge has to be implemented into the morphological analyzers which implies a long and tight collaboration between linguists and programmers; morphological analyzers cannot be easily adapted to other languages; they strongly depend on their underlying linguistic models...
- by means of morphological databases such as CELEX (Baayen et al., 1995). This solution can only be used for a very small number of languages: Dutch, English and German. For instance, as far as we know, no such database is available for romance languages. One reason of this lack of morphological databases is that their creation of is quite expensive.

However, several methods of supervised and unsupervised acquisition of constructional morphology have been proposed by authors. All of them involves some amount of symbolic or statistical learning. The input may be lexical data: inflected forms as in (Gaussier, 1999) and (Hathout, 2000) or medical nomenclature as in (Grabar and Zweigenbaum, 1999). More often, morphological knowledge is acquired from text corpora as do (Jacquemin, 1997), (Goldsmith, 2001), (Schone and Jurafsky, 2000; Schone and Jurafsky, 2001) or (Déjean, 1998). The learning of constructional morphology relies on a double approximation:

1. Word forms are good approximation of the phonological features.
2. Word forms can be used as approximation of words meaning: word forms that share a long enough substring are associated to lexemes that have good chances to be semantically related.

Corpus based methods can be very helpfull for specific NLP and RI tasks. In particular, they can adapt to the vocabulary of the texts. However their results cannot be easily accumulated into databases repositories because most of them do not have a sufficient precision. The problem of precision is common to all methods and tools that do not use a priori linguistic knowledge. While the first above approximation is quite satisfactory, the second one is very coarse and cannot be improved without integrating a minimal amount of semantic knowledge in the process. Semantics can be either included in the tools or described in an external resource. The latter option is superior to the former because it preserves the generality of the method and guarantees its independence from individual languages.

[1] We adopt the terminology proposed by Danièle Corbin and her team (Corbin, 2001); we prefer the term "constructional" to "derivational" which does not always imply a single notion.

2. Combining word formation and synonymy into analogies

Almost all unsupervised methods that acquire constructional morphology from corpora or lexicons proceed in two steps:

1. they connect word forms by stripping and adding graphemic affixes;
2. the connections are then filtered on the base on various parameters such as the frequency of the stripping/adding patterns, the number of characters stripped/added, the co-occurrence of the words in some segments of text (eg. fixed size windows), the similarity of the words contexts (measured by means of TF∗IDF weighting), etc.

The weakness of the methods (especially regarding precision) comes from the nature of the information used to decide whether the words can be connected or not: either it is statistical or it relies on *a priori* approximation. This problem may be solved by the use of ressources that contain some semantic knowledge and which have been build or checked by humans.

These ressources include lexical databases like WordNet and machine readable dictionaries (MDRs). Among these, dictionaries of synonyms are perfectly suited for semantic filtering for at least three reasons:

1. they have a uniform and standard format;
2. most of their information is encoded explicitly (it is made up of binary synonymy relations between entries);
3. synonymy relations is almost exactly the kind of semantic knowledge we are looking for (they precisely hold between words that share semantic features).

Dictionaries of synonyms have additional desirable features:[2]

- they exists (at least in printed form) for a many languages;
- their format does not depend on individual languages;
- they often have a quite small size (they can be made machine readable at a reasonable cost).

However, synonymy relations usually hold between words that belong to different constructional families while word formation connect members of the same family. Synonymy relations can be viewed as orthogonal to the constructional ones. Never the less, they can be easily exploited because the relation "share semantic features with" is transitive. More specifically, we aim at filtering the morphological links predicted on the base of the sharing of a common graphemic substring. For instance, in figure 1, abandon/V and abandonment/N are connected because they share the graphemic prefix abandon and because they match a suffixation pattern (/V:ment/N) which can connect 221 couples of the lexicon lemmata (see §4.). Similarly, the connection of desert/V and desertion/N is predicted on the base of their common graphemic prefix desert and of the 115 frequency of the pattern they match (/V:ion/N).[3]

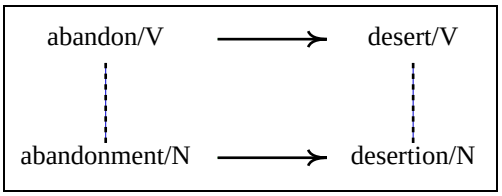

Figure 1: Example of a morpho-synonymy analogy (segment extracted from a proportional series). Nodes are lemmata tagged with categorial labels. Arrows correspond to synonymy relations given by a dictionary. Dashed lines correspond to predicted morphological connections.

Since words are morphologically connected on the base of their shared phonological and semantic features, the prediction of morphological links between abandon/V and abandonment/N and between desert/V and desertion/N involves a prediction that these pairs of lexemes share an large part of their semantic features. On the other hand, the dictionary of synonyms indicates that synonymy relations hold between abandon/V and desert/V and between abandonment/N and desertion/N and therefore, that the members of each of these pairs share a sizable part of their semantic features.

By using the transitivity of the relation of sharing of semantic features, we can assume that if a word X is a synonym of abandon/V and a word Y is a synonym of abandonment/N, then X and Y share a part of their semantic features. If, in addition, X and Y can be connected morphologically (on the base of X and Y word forms), then this link has a greater chance to be correct since it has been predicted "independently". The prediction that are based on transitivity uses only one predicted link. The other two are safer. To summarize, the proposed method combines constructional links and synonymy relations in order to make convergent (and consequently more accurate) predictions.

We call *morpho-synonymy analogies* the quadruplets as (abandon/V, abandonment/N, desert/V, desertion/N). From the linguistic point of this structure can be seen as a proportional series (Cruse, 1986). In particular, it can be read as "abandonment/N is to abandon/V as desertion/N to desert/V."

3. Extraction of dictionary of synonyms from WordNet

WordNet (Miller et al., 1990; Fellbaum, 1999) is well known lexical database which describes basic semantic relations between word senses such as synonymy, hypernymy, meronymy, antonymy, etc.[4] Word senses are represented as synsets that is "sets of synonyms that are inter-

[2] By an abuse of language, we will use the term *synonymy* for relations that might be better termed *semantic proximity*.

[3] Both frequencies correspond to patterns have been learned from the *S-dict* dictionary made up of WordNet synsets (see § 3.).

[4] `http://www.cogsci.princeton.edu/~wn/`

changeable in some context." WordNet is divided in four separate parts, one for each major category (noun; verb; adjective; adverb).

Here, we are only concerned with synonymy relations. WordNet describes three relations of semantic proximity. They can be displayed by `wn`,[5] the command line interface to WordNet when called with the `-syns(v|n|a|r)` and `-coor(v|n)` options.

In our experiment, each relation has been listed in a separate dictionary in order to compare the robustness of the proposed method. In all three dictionaries, words are tagged with their category (in other words, the categorial informations are distributed on the WordNet entries).

Synsets. The most strong proximity relation is the synonymy relation described by the synsets. For instance, abandon/V belongs to five synsets {abandon#1}; {abandon#2, give up#10}; {abandon#3, give up#2}; {vacate#2, empty#3, abandon#4}; {abandon#5, forsake#1, desolate#1, desert#1}. The words that compose the synsets have sense numbers attached to them. We do not use sense numbers in our experiment because only direct "synonymy" links are involved in the search for analogies. Besides, only suffixation relations between "simple" word forms are considered. For example, all synonymy relations that involve give up will not be ignored since this verb form is composed of two distinct words.

The word pairs connected by the strict synonymy relation (defined by the membership of the same synset (eg. abandon/V:desert/V)) have been gathered in a dictionary we called *S-dict*. Notice that these relation are not directed.

Synonymy/immediate hypernymy. The second relation is less tight. It corresponds to the synonymy/immediate hypernymy relation between distinct synsets (eg. the verb synset {abandon, forsake, desolate, desert} has the verb synset {leave} as synonym/immediate hypernym). For adverbs, the synonymy/immediate hypernymy relation is only given for their root adjectives. Synonymy and immediate hypernymy are not distinguished in WordNet. Unlike the previous strict relation, synonymy/immediate hypernymy is directed (eg. the verb synset {leave} is not given as having {abandon, forsake, desolate, desert} among its synonym/immediate hypernym). This orientation will be preserved in the *M-dict* dictionary which describes this semantic proximity relation. More specifically, each synonymy/immediate hypernymy link between synsets is decomposed into a set of links between word. *S-dict* has also been added to *M-dict*.

Coordinate terms. The third relation is the weakest one. It corresponds to the coordinate terms (sisters) relation and is only available noun and verb synsets (eg. the verb synset {abandon, forsake, desolate, desert} has the verb synset {jilt} as a coordinate term). We dealt with this relation as if it was directed: we did not symetricalize it. The coordinate terms (sisters) relation yielded an *L-dict* similar to *M-dict*. *M-dict* has been added to *L-dict*.

Table 1 presents some statistics on these three dictionaries. The differences in the number of entries come from the fact that only connected words are taken into account.

[5]We used WordNet 1.7.

	entries(#)	links(#)	ratio
S-dict	42 918	125 754	2.9
M-dict	68 176	306 887	4.5
L-dict	70 170	2 283 679	28.8

Table 1: Size and number of connections of the extracted dictionaries.

4. Computing the morpho-synonymy analogies

The search for morpho-synonymy analogies relies on the one hand on the synonymy relations described in the dictionaries presented in §3., and on the other on predicted constructional links.[6] The computation is independent of the individual dictionaries. Let us suppose we selected *X-dict*, one of the three previous dictionaries.

4.1. Creating a morphological graph

Morphological links are computed from *X-dict* word forms by means of a two steps procedure:

In the first step, a set of suffixation patterns is learned from *X-dict* word forms by means of a technique presented in (Dal et al., 1999). Patterns learning relies on two assumptions:

1. The longer the word, the stronger the correspondence between written form and semantic is. As consequence, if two word 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.

More specifically, a pattern is a triplet (s_1, s_2, f) where s_1, s_2 are graphemic suffixes with categorial tags and f is the pattern frequency. For instance, the learning program have kept the pattern (er/N, ion/N ,24) which connects 24 word pairs: abstracter/N:abstraction/N, asserter/N:assertion/N, ..., suppresser/N:suppression/N. More generally, a pattern (s_1, s_2, f) can connect a pair of word forms (w_1, w_2) if w_1 and w_2 share a common graphemic prefix p such that $w_1 = p \cdot s_1$ and $w_2 = p \cdot s_2$.[7] The pair (s_1, s_2) will be called constructional signature of the pair of word forms (w_1, w_2) (Jacquemin, 1997).[8] For the experiment, the main constraints that have been imposed on the learning process are that p must be at least 3 characters long and f must be greater that 2.

Other authors as have proposed similar techniques (Gaussier, 1999). The technique is also used in the `findaffix` script of the *Ispell* package. For languages with non concatenative morphology (eg. semitic languages), morphological links could be discovered by means

[6]Recall that "synonymy relation" is an abuse of language.

[7]For a formal presentation of these constructional patterns, see (Pirrelli and Yvon, 1999).

[8]The notation $(s_1 : s_2)$ and $(w_1 : w_2)$ will be used as a variate of (s_1, s_2) and (w_1, w_2).

of more sophisticated techniques like those proposed by (Lepage, 1998).

In the second step, the learned patterns are applied on *X-dict* word forms in order to generate a morphological graph. The graph contains correct links attraction/N:attractiveness/N and attraction/N:attractively/R, but also some wrong links as attraction/N:attention/N, attraction/N:attestation/N, attraction/N:attest/V, attraction/N:attain/V or attraction/N:attended/A.[9] Morpho-synonymy analogies are used in order to discriminate between the two sets of links.

4.2. Combining the morphological graph and the synonymy graph

X-dict can be seen as a graph of synonymy relations with exactly the same nodes as the corresponding morphological graph that is *X-dict* word forms. Combining the two graphs is quite easy. We just need to explore the morphological graph in order to find all the quadruplets of word forms $(x_1 : x_2, y_1 : y_2)$ such that (x_1, x_2) and (y_1, y_2) belong to the morphological graph and (x_1, y_1) and (x_2, y_2) belong to the synonymy graph. Figure 2 presents some examples of analogies that occur in *S-dict*. We define the constructional signature of an analogy $(x_1 : x_2, y_1 : y_2)$ as the quadruplet $(s_1 : s_2, t_1 : t_2)$ formed by the constructional signatures $(s_1 : s_2)$ and $(t_1 : t_2)$ of the word pair forms $(x_1 : x_2)$ and $(y_1 : y_2)$.

Filters are applied during the exploration of the graphs in order to eliminate four classes of the morpho-synonymy analogies that tend to be often incorrect. These are:

1. Analogies where the word forms of x_1, x_2, y_1, y_2 are not all distinct. This constraint eliminates the analogies with words that are converted from each other as (abstract/A:abstraction/N, abstractionist/A:abstract/N). These analogies are usually incorrect because they correspond to distinct senses of polysemous words.

2. Analogies where one pair of words is formed on the other by prefixation as in (engraved/A:engraft/V, graved/A:graft/V).[10] This constraints has been devised for a similar experiment on a French dictionary of synonyms (Hathout, 2001) and has been kept for English, even if it did apply for the dictionaries extracted from WordNet.

3. Analogies where x_1, x_2, y_1, y_2 belong to the same constructional family as (abstinence/N:abstinent/N, abstention/N:abstainer/N). More specifically, the constraint filters analogies where the graphemic stem of $(x_1 : x_2)$ is a prefix of the graphemic stem of $(y_1 : y_2)$ or vice versa. In most such cases, the semantic relations that hold between $(x_1 : x_2)$ and between $(y_1 : y_2)$ are different.

4. Analogies where $(x_1 : x_2)$ and $(y_1 : y_2)$ are also synonyms as (advancement/N:advance/N, movement/N:move/N). These analogies present a problem similar to the previous one: the semantic relations that hold between $(x_1 : x_2)$ and between $(y_1 : y_2)$ often are different.

Among the results shown in figure 2, some are correct analogies as (unceasing/A:unceasingly/R, unending/A:unendingly/R) in the sens that unceasing/A is to unceasingly/R as unending/A to unendingly/R. Other as (bluing/N:bluish/A, blue/N:blueish/A) are not since bluish/A and bluish/A are only graphemic variants while bluing/N is the process of becoming blue and blue/N is a name of color. But both analogies are made up of correct constructional links: bluing/N:bluish/A, are morphologically related and so are blue/N:blueish/A.

	analogies	morphological links	
dictionary	(#)	(#)	precision(%)
S-dict	35 044	22 878	**98**
M-dict	89 504	58 448	87
L-dict	376 390	118 838	66

Table 2: The first column indicates the number of acquired analogies. The second column gives the number of distinct morphological links. The third column presents an estimation of the precision of the morphological links acquisition.

The acquisition of analogies have been performed in the dictionaries *S-dict*, *M-dict* and *L-dict*. Table 2 presents the number of analogies and of morphological links acquired for each of these dictionaries. Our aim been the creation of a constructional database, we only evaluated the morphological links. More specifically, we have checked manually one sample of 100 morphological links for each dictionary in order to estimate the precision of the process. As expected, the precision decreases with the increase of the number of the links in the dictionaries and with the loosening of the semantic relation.

5. Typing the analogies

The results presented in table 2 are quite good, especially for *S-dict*. However, they can be improved significantly by typing the analogies.

We can first divide the analogies into two groups. The first group which we will term *strong* (S) includes analogies as (transformation/N:transformable/A, transmutation/N:transmutable/A) where both pairs of word forms have the same constructional signature (tion/N:ble/A). The second group which we will term *weak* (W) includes analogies as (protester/N:protest/N, objector/N:objection/N) where the constructional signature of the first word form pair is (er/N:/N) and the one of the second word form pair is (er/N:ation/N). For now, this typing does not take into account allomorphy. For instance, (ed/A:ation/N) and (ed/A:ion/N) are seen as distinct signatures and an analogy as (deformed/A:deformation/N, distorted/A:distortion/N) is regarded as weak. The typing of the analogies can also be regarded as a typing of their constructional signatures.

[9] These links belong to the morphological graph of *S-dict*. Since *S-dict* is a part of the other two dictionary (*M-dict* and *L-dict*), the links also belong to their morphological graph.

[10] This example has been coined. WordNet 1.7 does not contain the adjective graved.

break/V:breakup/N	separate/V:separation/N
slight/A:slightly/R	slim/A:slimly/R
embezzler/N:embezzle/V	defalcator/N:defalcate/V
clench/V:clenched/A	clinch/V:clinched/A
bluing/N:bluish/A	blue/N:blueish/A
trickily/R:trick/V	foxily/R:fox/V
lumpy/A:lump/V	chunky/A:chunk/V
thieving/N:thief/N	stealing/N:stealer/N
unceasing/A:unceasingly/R	unending/A:unendingly/R
whiskered/A:whiskers/N	bearded/A:beard/N

Figure 2: Examples of morpho-synonymy analogies.

Two types of morphological links can then be defined relatively to the two types of analogies. A morphological link will be termed *strong* (S) if its constructional signature belongs to the signature of at least one strong analogy. Morphological links with signatures that only occur in weak analogies will be said to be *weak* (W). The typing of the morphological links can also be regarded as a typing of their signatures. For instance, in *S-dict*, (tion/N:ble/A) is a strong signatures while (ry/N:/A) is a weak one. Morphological links such as rivalry/N:rival/A which have that weak signature are also weak.

The two types of morphological links signatures induce a sub-typing of the weak analogies. More specifically, weak analogies can be composed of two strong links (eg. (adaptation/N:adapt/V, adjustment/N:adjust/V)) , or two weak links (eg. (acidulate/V:acidulousness/N, acidify/V:acidity/N)) or one strong link and one weak link (eg. (settlement/N:settle/V, resolution/N:resolve/V) with a strong signature (ment/N:/V) and a weak one (ution/N:ve/V)).[11] The sub-typing of the weak analogies can be refined further. One may distinguish the weak link signatures than occur in {S,W} analogies from the ones that only occur in {W,W} analogies...[12]

The typing of the analogies reveals that all four types are not of the same importance. More precisely, strong analogy signatures have 5.2 to 6.9 more instances than weak ones (see tables 3, 4 and 5). The three sub-types of the weak analogy signatures also show sharp differences in their numbers of instances.

The types of analogies can also be used as filters in order to improve the precision of the morphological links. Table 6 show that the precision varies with the analogy types and with the dictionary.[13] Some types concentrate the main part of the incorrect links. Moreover the contrast between the types is more sharp as the dictionaries grow in size and the corresponding semantic proximity relations loosen. In other words, the efficiency of the filtering based on analogy typing increases with the weakening of the semantic relations. However, the growth does not change the relative order of the precision and for all three dictionaries, and {S,S} weak analogies are the most secure ones. This is not an expected result and the fact that strong analogies are consistently less good as {S,S} weak ones has still to be explained.

[11] All three examples are *S-dict* analogies.

[12] {S,W} analogies include the (S,W) and (W,S) ones.

[13] The precision have been estimated by checking manually samples of 100 morphological links.

type	analogies(#)	signatures(#)	mean
strong	13 652	1 370	9.9
weak	21 392	11 168	1.9
weak {S,S}	12 008	3 492	3.4
weak {S,W}	7 516	5 880	1.2
weak {W,W}	1 868	1 796	1.0

Table 3: Analogies acquired from *S-dict*. Strong analogy signatures have 5.2 times more instances than weak ones.

type	analogies(#)	signatures(#)	mean
strong	22 438	2 008	11.1
weak	67 066	34 870	1.9
weak {S,S}	33 482	8 248	4.0
weak {S,W}	25 322	18 618	1.3
weak {W,W}	8 262	8 004	1.0

Table 4: Analogies acquired from *M-dict*. Strong analogy signatures have 5.8 times more instances than weak ones.

type	analogies(#)	signatures(#)	mean
strong	44 666	3 372	13.2
weak	331 724	215 276	1.5
weak {S,S}	101 336	22 800	4.4
weak {S,W}	128 844	92 108	1.3
weak {W,W}	101 544	100 368	1.0

Table 5: Analogies acquired from *L-dict*. Strong analogy signatures have 6.9 times more instances than weak ones.

6. Comparison of with CELEX

The comparison of the acquired morphological links with CELEX has two aims. First it give an estimation of the acquisition recall. Second, it addresses the question of the actual usefulness of the acquired links.

The CELEX English database gives a constructional analysis. The morphological descriptions are bracketed structure such as (((equal)[A],(ize)[V|A.])[V],(ation)[N|V.])[N]. These structures are transformed into a morphological links corresponding to their most peripheral constructional step (eg. (((equal)[A],(ize)[V|A.])[V],(ation)[N|V.])[N] gives equalization/N:equalize/V, ((equal)[A],(ize)[V|A.])[V] gives equalize/V:equal/A... Only suffixation is considered, and as for

	S-dict		*M-dict*		*L-dict*	
type	(#)	precision(%)	(#)	precision(%)	(#)	precision(%)
strong	11 838	**97**	20 864	**93**	30 116	87
weak {S,S}	8 598	**99**	25 232	**97**	39 964	**92**
weak {S,W}	6 130	**96**	23 964	87	60 260	62
weak{W,W}	1 612	76	9 728	62	44 072	23

Table 6: The precision of the morphological links varies with the analogy types and with the dictionaries.

	word forms intersection		links in the restriction		common	recall
	(#)	(%)	(#)	(%)	(#)	(%)
S-dict	20 728	72.7	10 296	**54.9**	9 748	**94.6**
M-dict	23 234	81.5	28 042	**52.0**	24 132	86.0
L-dict	23 262	81.6	54 928	**53.7**	34 008	61.9

Table 7: Comparison of the morphological links acquired from the dictionaries extracted from WordNet with *STD*.

	S-dict		*M-dict*		*L-dict*	
type	common links(#)	recall(%)	common links(#)	recall(%)	common links(#)	recall(%)
strong	4 048	**97.4**	7 908	**97.6**	10 858	**95.4**
weak {S,S}	3 438	**98.6**	10 616	**98.1**	16 482	**93.2**
weak {S,W}	2 768	**92.4**	10 006	85.1	25 212	58.3
weak{W,W}	744	71.3	4 368	44.2	19 336	17.1

Table 8: Recall with respect to *STD* for each type of analogies.

WordNet, only entries with simple word forms have been used, that is 39 302 entries on a total number of 52 447. The extraction have produced 20 063 suffixation links connecting 28 501 distinct entries.

However, these links only describe single constructional steps while the morphological links acquired from WordNet can be complex. The comparison must therefore be made with the symmetric and transitive closure of the set of links extracted from CELEX. The closure resulted in a morphological base of reference composed of 99 358 links. Let us call it *STD*.

Table 7 presents a comparison of the morphological links acquired from the dictionaries extracted from WordNet with *STD*. The first two columns indicate the size of the intersection of the word lists of the acquired morphological links with that of *STD*; the percentages are relative to *STD* word list. The third and fourth columns give the number of acquired morphological links with both ends in the intersection of the word lists; the percentages are relative to the total number of links acquired from the respective dictionaries. The fifth column indicates the number of links in the restriction that also belong to *STD*. The sixth column presents the recall with respect to *STD*. The recall of the acquisition of morphological links is quite good, especially when it uses a strong synonymy relations as the membership to the same synset. The contribution of the different types of analogies to the overall recall is not uniform as table 8 shows.

For all three extracted dictionaries, approximately half the morphological links connect words that do not belong to CELEX suffixed word list. Once checked by humans, these morphological links could represent a significant complement to CELEX.

7. Conclusion

We have presented a method that exploit synonyms or semantic proximity relations in order to extract pairs of constructional links that form proportional series, namely morpho-synonymy analogies. The method is very general because it is independent of specific languages, all the linguistic knowledge involved in the acquisition is external to the system as it is encoded in the dictionaries of synonyms. A similar experiment carried out on a French dictionary of synonyms is described in (Hathout, 2001). The French results are similar in number and precision to the ones from *M-dict*.

The method is also very robust and does not require the dictionaries to have specific features. This point is established by repeating the acquisition of morphological links from three dictionaries describing a strict synonymy relation *S-dict*, a loose one *L-dict* and an intermediate one *M-dict*. The only incidence of the loosening of the semantic relation have been a degradation in precision and recall. However, a typing of the analogies signatures has been proposed in order to further filter the acquired morphological links.

8. References

Evan L. Antworth. 1990. *PC-KIMMO: a two-level processor for morphological analysis*. Number 16 in Occasional Publications in Academic Computing. Summer Institute of Linguistics, Dallas, TX.

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX lexical database (release 2). CD-ROM. Linguistic Data Consortium, University of Pennsylvania, Pennsylvania, USA.

Joan L. Bybee. 1988. Morphology as lexical organization. In Micheal Hammond and Michael Noonan, editors, *Theoretical Morphology. Approaches in Modern Linguistics*, chapter 7, pages 119–141. Academic Press, San Diego, CA.

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

Danielle Corbin. 2001. Préfixe et suffixes : du sens aux catégories. *International Journal of French Language Studies*, 11:41–69.

D. Alan Cruse. 1986. *Lexical Semantics*. Cambridge University Press, Cambridge, UK.

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

Hervé Déjean. 1998. Morphemes as necessary concept for structures discovery from untagged corpora. In *Proceedings of the Workshop on Paradigms and Grounding in Natural Language Learning*, pages 295–299, Adelaide, Australia.

Christiane Fellbaum, editor. 1999. *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Éric Gaussier. 1999. Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the Workshop on Unsupervised Learning in Natural Language Processing*, University of Mariland, USA. Association for Computational Linguistics, ACL'99.

John Goldsmith. 2001. Unsupervised learning of the morphology of natural language. *Computational Linguistics*, 27(2):153–198.

Natalia Grabar and Pierre Zweigenbaum. 1999. Acquisition automatique de connaissances morphologiques sur le vocabulaire médical. In *Actes de la 6^e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-99)*, pages 175–184, Cargèse, July.

Nabil Hathout. 2000. Morphological pairing based on the network model. In *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)*, pages 35–38, Pyrgos, Grèce.

Nabil Hathout. 2001. Analogies morpho-synonymiques. une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes. In Denis Maurel, editor, *Actes de la 8^{e}Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2001)*, Tours. ATALA.

Christian Jacquemin and Evelyne Tzoukermann. 1999. Nlp for term variant extraction: synergy between morphology, lexicon, ans syntax. In Tomek Strzalkowski, editor, *Natural Language Information Retrieval*, pages 25–74. Kluwer Academic Publishers, Dordrecht.

Christian Jacquemin. 1997. Guessing morphology from terms and corpora. In *Proceedings of 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'97)*, pages 156–167, Philadelphia, PA. ACM.

Robert Krovetz. 1993. Viewing morphology as an inference process. In Robert Korfhage, Edie Rasmussen, and Peter Willett, editors, *Proceedinds of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 191–202, Pittsburg. ACM.

Yves Lepage. 1998. 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, August.

Georges A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4):335–391.

Fiammetta Namer and Georgette Dal. 2000. 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. ELRA.

Fiammetta Namer. 2000. FLEMM : un analyseur flexionnel du français à base de règles. *Traitement automatique des langues*, 42(2):523–548.

Vito Pirrelli and François Yvon. 1999. The hidden dimension: a paradigmatic view of data-driven nlp. *Journal of Experimental & Theoretical Artificial Intelligence*, 1999(11):391–408.

Patrick Schone and Daniel S. Jurafsky. 2000. Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the Conference on Natural Language Learning 2000 (CoNLL-2000)*, pages 67–72, Lisbon, Portugal.

Patrick Schone and Daniel S. Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL-2001)*, Pittsburgh, PA.

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