Phonotactics in morphological similarity metrics

Nabil Hathout

CLLE/ERSS. CNRS & Université de Toulouse

Abstract

This study explores the contribution of phonotactic regularities to the way we measure morphological similarity between words. I focus on how morphological similarity fits with conventional morphemic analysis and present a method to exploit the latter to create two reference resources from the English part of the CELEX database. The resources are used for the comparison and evaluation of morphological similarity metrics. I also compare four measures that estimate morphological similarity: the standard edition distance of Levenshtein; the measure of De Pauw and Wagacha; Proxinette; and PHACTS. The study provides an idea of the contribution of phonotactic regularities to morphological similarity.

1. Introduction

In this paper, I investigate a basic concept that has not been extensively studied: morphological similarity between words. This concept is fundamental in morphology because morphological relations connect words that share semantic and phonological properties, or that are semantically and phonologically similar. Because the shared properties may vary from one pair of words to another, the morphological relations between them also vary. In this paper, I show how it is possible to measure the variation.

Morphological similarity is a gradable property. I will assume that it is measurable and that it takes values ranging from 0 to 1: 1 in case of identity and 0 if the two words have no relation. For example, the pairs in (1) all have some degree of morphological similarity. In contrast, the pairs in (3) have no morphological similarity because they share only semantic properties (as in (3a)), only phonological properties (as in (3b)), or no semantic or phonological properties (as in (3c)).

(1) a. law:unlawful; admirable:admirably

b. projectionist:percussionist

c. photograph:phonograph

(2) indistinguishably:readability; collectivization:decisiveness

(3) a. legality:lawfulness; rich:wealthy

b. piece:peacefully; rights:writer

c. moon:fishery

The similarity between a base and one of its derivatives, such as *admirable*:*admirably* in (1), is clearly perceivable. The second example in (1) shows that words with no direct derivational relationship may also have some morphological similarity, such as *law*:*unlawful*,where two derivations are involved (*law* → *lawful* → *unlawful*). The examples in (1b) illustrate a morphological similarity between two words constructed by the same two derivations. *Projectionist* and *percussionist* result from a suffixation in *-ist* of a *-ion* derivative. Morphological similarity also exists between words that share a compounding element, as in *photograph*:*phonograph* in (1c), or an affixe, as in (2), where *indistinguish**ably***:*read**abili**ty* share a suffixation in *-able* and *collec**tiv**ization*:*decis**iv**eness* share a suffixation in *-ive*.

The examples in (1) also show that not all words have the same morphological similarity. Some are more similar than others. For instance, the similarity between *admirable* and *admirably* is clearly stronger than that between *collectivization* and *decisiveness*. This similarity allows a speaker who knows *admirable* and hears *admirably* for the first time to have a good idea of the meaning of the latter. In contrast, knowledge of the word *collectivization* does not allow the same speaker to know what *decisiveness* means because the semantic and phonological properties shared by these words are not sufficiently informative. More generally, the pairs in (1a) are more similar than the ones in (1b) and (1c), which are themselves more similar than the ones in (2).

The notion of similarity between words is commonly used in various fields of linguistics. It plays a central role in distributional semantics (Harris, 1954, 1979; Firth, 1957; Rubenstein and Goodenough, 1965) based on the assumption that words with similar distributions are semantically close. Distributional similarity has been used to identify morphological relationship, such as in (Schone and Jurafsky, 2000, 2001) and (Baroni et al., 2002).

This work is a continuation of Hathout (2011b). It focuses on the way morphological similarity can be estimated. The paper makes two main contributions.

1. Section 3.2 shows how morphological similarity fits with conventional morphemic analysis and presents a method to exploit the latter to create reference resources for the comparison and evaluation of morphological similarity metrics.
2. Section 4 compares four measures that estimate morphological similarity: the edition distance of Levenshtein (1966) used, for instance, in (Baroni et al., 2002); the measure of De Pauw and Wagacha (2007); Proxinette of Hathout (2009); and PHACTS of Calderone and Celata (2011, 2012).

2. Morphological similarity

Under what conditions can we consider two words morphologically similar? The simple answer follows from the conception of a morphological relationship as a relation between words that simultaneously share semantic and phonological properties. This relatively loose definition does not require the existence of a regular association between the shared semantic and phonological properties. In fact, almost all the morphological relations are regular. They correspond to the configurations shown in (1): relations between members of the same derivational family (1a, 1c) or the same derivational series (1b). In contrast, non-paradigmatic similarity, as in (2), is less regular.[1] In what follows, I only consider these three configurations.

Any definition of morphological similarity must properly reflect the speaker's and the linguist's intuition. The first is that the similarity between a base and a derivative decreases with the number of steps of derivation, as in (4), where *fiction* is more similar to *fictional* than to *fictionalization*.

(4) fiction:fictional; fiction:fictionalize; fiction:fictionalization

We can generalize this correlation to words that are not derived from each other by considering derivational families graphs whose edges represent derivational relations. The similarity between two words of the same family is then estimated by the length of the shortest path that connects them. For example, (5) shows that *gracelessness*:*gracelessly* are more similar than are *indiscernible*:*discernment*. *Gracelessness* and *gracelessly* are separated by a path of length 2, whereas *indiscernible* and *discernment* are separated by a path of length 3.

(5) a. gracelessness ↔ graceless ↔ gracelessly

b. indiscernible ↔ discernible ↔ discern ↔ discernment

Speakers are also likely to have some intuition for pairs of words that belong to the same derivational series, as in (6), where the similarity in (6a) is greater that in (6b) and (6c).

(6) a. constitutionally:operationally

b. constitutionally:architecturally

c. constitutionally:gloriously

The similarity between two words from the same series increases with the specificity of the latter: *constitutionally* and *operationally* belong to a series of adverbs in *-ionally*, *constitutionally* and *architecturally* to a series of adverbs in *-ally*, and *constitutionally* and *gloriously* to a series of adverbs in *-ly*. The specificity of a series can be estimated by the number of steps of derivation common to all the words it contains: three steps in (6a) with suffixations in *-ion*, *-al* and *-ly*; two in (6b) with suffixations in *-al* and *-ly*; and one in (6c) with a suffixation in *-ly*. It then becomes possible to compare the similarity of pairs of words from different series. For instance, *constitutionally*:*operationally* are more similar than *safety*:*certainty*, which share only one suffixation in *-ty*.

For NP-similar pairs, illustrated in (2), the intuition is unclear. Presumably, the similarity increases with the number of shared affixations, but such an estimation does not take into account the possible differences in the semantic contribution of the affixes.

[1] In the following, the pairs that share one or more affixes but do not belong to the same derivational family or the same derivational series will be called "NP similar", NP standing for non-paradigmatic (see Section 3.2.1 for a formal definition).

Another difficult question remains unanswered: how does the similarity of a pair of words from the same family compare with the similarity of a pair of words from the same series? For instance, how do we know whether *constitutionally*:*operationally* are more similar than *indiscernible*:*discernment* or vice versa? What about the similarity of *operationally*:*constitutionally* and *operationally*:*cooperation*? The next section provides a tentative answer to these questions.

3. Data and methods

The first part of the answer is the definition of two reference resources created from CELEX[2] (Baayen et al., 1995). The second part is the comparison of four measures and two baselines. All experiments were performed in English.

3.1. Data

The most important quality of a reference resource is reliability. Ideally, a reference resource of morphological similarity should be created by lexicographers using clear and detailed guidelines. In practice, the creation of such a resource faces two major challenges: cost that exceeds the financial capacity of most of the research teams and the difficulty of defining the concept, one consequence being that speakers have no insight into the similarity of many pairs of words. To avoid these problems, I used the English part of the CELEX database to create two reference resources. CELEX is a well-known lexical database that is regularly used in psycholinguistics and natural language processing. It provides detailed phonological, morphological, syntactic, and distributional information on a significant fragment of the lexicons of English, German, and Dutch. The morphological descriptions in CELEX include a representation of the morphological structure of the words as shown in Figure 1.

```
governable      ((govern)[V],(able)[A|V.])[A]
traditionally   (((tradition)[N],(al)[A|N.])[A],(ly)[B|A.])[B]
```

Figure 1: CELEX morphological structures. On the first line, `(able)[A|V.]` states that *-able* is an affixation that derives adjectives from verbs. The position of the affix with respect to the base is indicated by the dot.

I also extracted two lexicons from CELEX. The first is a large lexicon that contains nouns, verbs, adjectives, and adverbs with typographically simple canonical forms (i.e. lemma exclusively composed of lowercase letters). This lexicon has 38 670 entries. The second lexicon is a subset of the large one. It contains entries with a lemma of at least 3 characters and a frequency greater than or equal to 20 for a total of 17 887 entries. The need for a reduced lexicon is due to limitations of the MaxEnt library[3].

[2] celex.mpi.nl.

[3] opennlp.apache.org.

3.2. Similarities of reference

3.2.1. Structural matching similarity

CELEX morphological structures were first used to build a “structural matching” reference resource or SMS. The SMS contains three parts corresponding to the three types of morphological relationships discussed in Section 1: relationships between members of derivational families, between members of derivational series, and between NP-similar words.

Derivational families are sets of words that share a stem or one or more compounding elements. Families are extracted from a derivational graph created from CELEX. The family of a word includes its derivational ancestors, their descendants, and the words it shares with one or more compounding elements. The length of the shortest path from one family member to another is used as an estimate of their similarity in SMS. For example, (7) shows that *governable* is connected to *government* by a path of length 2.

(7) governable ↔ govern ↔ government

Series are sets of words that are built by the same sequence of morphological processes. The morphological structures of these words share a common “outer shell” that can be described by a derivational schema. The series to which an entry belongs can be deduced from its morphological structure by considering the abstractions of the components that contain a stem or a compounding element. For instance, from the decomposition of *navigability* in (8), we can deduce that it belongs to a series of nouns in *-ability* described by the schema (9a) and a series of nouns in *-ity* matching the schema (9b). The schema in (9a) results from the abstraction of the representation of the verb *navigate*, and the one in (9b) results from the abstraction of the representation to the adjective *navigable*.

(8) `(((navigate)[V],(able)[A|V.])[A],(ity)[N|A.])[N]`

(9) a. `((@,(able)[A|V.])[A],(ity)[N|A.])[N]`

b. `(@,(ity)[N|A.])[N]`

The similarity between the members of a series can be estimated by the complexity of its schema. The complexity can be measured by the number of variable positions (represented by the symbol `@`) and nodes that carry categorical information. For instance, the complexity of the series of nouns in *-ability* is 3 (`@, [A], [N]`) , and the complexity of the nouns in *-ity* is 2 (`@, [N]`).

NP-similar words can easily identified from their structures. Two words are NP similar if they do not belong to the same family or the same series and if their structures contain at least one common affix. This is the case for the pairs in (10). The similarity between these words can be estimated by counting the number of common affixes in their structures. For instance, the SMS similarity of the pair (10) is 2, the common affixes being *-al* and *-ion*.

(10) occupat**ional**ly:industri**al**izat**ion**

```
((((occupy)[V],(ation)[N|V.])[N],(al)[A|N.])[A],(ly)[B|A.])[B]
((((industry)[N],(al)[A|N.])[A],(ize)[V|A.])[V],(ation)[N|V.])[N]
```

In SMS, each type of morphological relationships has its own scale. The estimates are completely different in nature and are not directly comparable. They define relative orders within each subset of similar words. However, this order can be extended by ordering these three subsets with respect to the intuition that family members are closer to each other than series members, which are themselves closer than NP-similar words (see Section 2). The resulting resource, SMS, is illustrated in Figure 2,which presents the 20 words that are most similar to the adjective *governable* in the Large corpus.

govern_V **ungovernable**_A **governance**_N **government**_N **governor**_N **misgovern**_V **governess**_N **governmental**_A **governorship**_N **guv**_N **misgovernment**_N **misgovernment**_N predictable_A comfortable_A playable_A fortifiable_A attainable_A approachable_A certifiable_A navigable_A

Figure 2: The words most similar to *governable* with respect to SMS. Words in boldface belong to the derivational family of *governable*. Words in light type belong to its derivational series. The letters following the underscore indicate the grammatical category.

Table 1 shows the average number of similar words in the Small and Large corpora for the three types of morphological relationships. It highlights that families are small sets and that most of the similar words belong to the series (Hathout, 2011b). The ratio between the sizes of the series and the NP-similar words shows that the contexts in which affixes may occur are subject to strong constraints.

Type	Large	Small
F	9	3
S	1386	584
NP	105	31
SMS	1234	497

Table 1: Average number of similar words for the three types of morphological relationships.

3.2.2. Paradigmatic strength similarity

A SMS reference is optimal for each of the three types of morphological relationships when considered separately. In contrast, the fact that family members are always nearer to the entry than any member of the series is a very rough approximation. For instance, is *incompetency* more similar to *competitiveness* or to *insufficiency*? To overcome this problem, I built a second reference resource calculated from the same CELEX morphological structures. This resource treats all three types of morphological relationships uniformly. It uses the morphological structures indirectly by considering the analogies involving these structures instead of the morphological relations they define. Examples of such analogies are presented in (11) and (12). Recall that an analogy $A : B = C : D$ is a relation between four terms (A, B, C, D) such that A is to B as C is to

D. Analogical quadruples combine members of the same derivational family and members of the same derivational series in one relation, making the comparison of different types of morphological relationships possible. These analogies allow us to measure morphological similarity by counting the number of quadruples in which each pair of words appears. Two words are considered more similar if they appear together in more quadruples. This is equivalent to morphological relations frequency (i.e. the size of the derivational series), which is an index of morphological regularity.

(11) a. govern:governable = accept:acceptable

b. govern:governable = imagine:imaginable

c. govern:governable = educate:educable

(12) governable:acceptable = govern:accept
governable:acceptable = governance:acceptance
governable:acceptable = governability:acceptability

Linguistic relations between a pair of words are numerous, and each one can potentially give rise to analogies. For example, in the case of (11a), the analogy is

graphemic: the string `governable` is obtained by adding `able` at the end of `govern` in the same way that `acceptable` is obtained by adding `able` to the end of `accept`;

phonological: /ɡʌvərnəbəl/ is obtained by adding /əbəl/ at the end of /ɡʌvərn/ just as /æksɛptəbəl/ is obtained by adding /əbəl/ to the end of /æksɛpt/;

morphological: *governable* is derived from *govern* by an *-able* suffixation in precisely the same way as *acceptable* is derived from *accept*;

semantic: *governable* is the quality of what can be *governed* just as *acceptable* is the quality of what can be *accepted.*

Morphological similarity must naturally be based on morphological analogies (i.e., on the relations described by the morphological structures). Morphological structures abstract away from the orthographic and phonological variations, as in (11b) and (11c). In the first example, the elision of the final `e` in *imagine* prevents the analogy between the written forms. In the second, there is no analogy at the phonological level because the /eɪt/ ending of *educate* is truncated. When considered at the level of the morphological structures, the analogies in (11) become explicit, as in (13).

(13) a. `(govern)[V]:((govern)[V],(able)[A|V.])[A] =`
`(accept)[V]:((accept)[V],(able)[A|V.])[A]`

b. `(govern)[V]:((govern)[V],(able)[A|V.])[A] =`
`(imagine)[V]:((imagine)[V],(able)[A|V.])[A]`

c. `(govern)[V]:((govern)[V],(able)[A|V.])[A] =`
`(educate)[V]:((educate)[V],(able)[A|V.])[A]`

ϵ	(govern)[V]	ϵ
(	(govern)[V]	,(able)[A|V.])[A]

ϵ	(accept)[V]	ϵ
(	(accept)[V]	,(able)[A|V.])[A]

Figure 3: Formal analogy between morphological representations. The differences are framed. ϵ represents the empty string.

The analogies between strings, called formal analogies, have been studied in great detail by Lepage (2003), Yvon (2006), and Stroppa (2005). One way to check a formal analogy is to find a decomposition (or factorization) of the four strings such that the differences between the first two are identical to the differences between the second two. These can be represented as in Figure 3. The analogical quadruples formed by CELEX morphological structures can be obtained by means of the method proposed in (Lepage, 2004; Gosme and Lepage, 2011).

We use morphological analogies to estimate morphological similarity. The pair of words $X_1 : Y_1$ is more similar than the pair $X_2 : Y_2$ if the number of quadruples in which $X_1 : Y_1$ appears is larger than the number of quadruples in which $X_2 : Y_2$ appears. This amounts to estimating the similarity of two words by the size of their derivational paradigms. For example, the number of pairs connected by a suffixation in *-able* is an estimate of the similarity of *govern*:*governable*. The number of pairs of members of the derivational families of *govern* and *manage* with superimposable structures, as in (14), is an estimate of the similarity of *governable*:*manageable*.

(14) govern:manage
government:management
misgovernment:mismanagement

This measure of similarity is defined uniformly for all pairs of words and is applicable to all types of derivational relations. In what follows, it is called "paradigm strength" or PSS. Figure 4 shows the 20 words most similar to *governable* in the Large corpus with respect to PSS. In this example, PSS is fully compatible with intuition. The members of the derivational family tend to be more similar. The members of its series, the *-able* suffixed adjectives, come next. The order of the family members also confirms the adequacy of PSS with intuition: *govern*, the base of *governable*, is the most similar, followed by its immediate derivative *ungovernable* and three derivatives of *govern*. PSS and SMS are compared in Section 4.

govern_V **ungovernable**_A **government**_N **governor**_N **governance**_N manageable_A utterable_A impeachable_A endurable_A employable_A avoidable_A serviceable_A inhabitable_A favourable_A comfortable_A reliable_A **misgovernment**_N treatable_A translatable_A touchable_A

Figure 4: The most similar words to *governable* with respect to PSS.

3.3. Metrics

SMS and PSS are based on morphological representations created and revised by lexicographers and are only defined for a small fragment of the English lexicon. Other methods must be used if one wishes to estimate the morphological similarity of words that are not part of the Large or Small corpora. These estimates can be calculated from word forms or phonological transcriptions. Ideally, measures of morphological similarities should take into account the meaning of the words represented, for instance, in distributional semantic spaces, as in (Baroni et al., 2002; Schone and Jurafsky, 2000, 2001). More recently, Lazaridou et al. (2013) have applied methods of composition of the distributional semantic representations to the calculation of the meaning of morphological derivatives.

In the experiments reported below, we compare a variety of metrics. Two of them are basic models (or baseline methods) that perform basic parses of the words and four more sophisticated measures that are more relevant from a computational or cognitive perspective. The first is Levenshtein edit distance included because it is a *de facto* standard in NLP for word similarity. Notice that the longuest common subsequence (LCS) is also widely used but we dismissed it because it only consider the common letters in the compared words and their relative order. Therefore, words which are not alike at all are declared to be similar by this measure (for instance, *human* has an LCS of 5 with *chimpanzee* and only of 3 with *woman*). We also tested two measures (DPW and Proxinette) designed for the creation of linguistic resources from raw input, namely lists of words. These measures focus on the linguistic generalizations that emerge from the corpus. The forth method is PHACTS, a psycho-computational model of phonotactics acquisition. It is designed to assess the role of frequency and position in the phonotactic processing.

Baselines. Two baselines have been defined to assess the difficulty of the task. One is oriented toward the derivational families, and the other is oriented toward the derivational series. The first baseline, LCPref, estimates the similarity of the words according to the size of their longest common prefix. This method tends to bring closer words derived by the same prefixation and words derived by suffixation from the same stem. We only consider pairs of word that share at least two initial characters. The second baseline, LCSuff, estimates the similarity of words by the size of their longest common suffix. It brings closer words built by the same suffixation. Only pairs of words that share at least two final characters are considered. The baselines can be seen as basic phonotactic models only that represent the words by their initial or final parts. Given the English phonotactics, stems occur at the beginning of the words and derivational suffixes at the end. Therefore, we expect LCPref to select family members with high precision and LCSuff to do so for series members.

Levenshtein. The most popular measure used to estimate the similarity of written forms is the edit distance (Levenshtein, 1966), which counts the number of edit operations (addition, deletion, or substitution of characters) needed to transform one string into the other. For example, we must perform 2 operations to transform `adorable` into `admirable`: replace `o` with `m` and add one `i`. In this study, I have used the Levenshtein ratio from the Levenshtein Python library. This measure is defined as follows:

$$\text{Levenshtein.ratio}(w_1, w_2) = \frac{|w_1| + |w_2| - \text{dist}(w_1, w_2)}{|w_1| + |w_2|}$$

where $|w|$ is the length of w and *dist* is an edit distance where replacements have a cost of 2. I imposed a minimum threshold of 0.53 on this ratio.

The Levenshtein distance is a versatile NLP measure used in many applications ranging from spell-checking, approximate string matching to DNA analysis. This distance is however blind to phonotactics because the position of an edit operation does not affect its cost. It also does not take frequency into account: frequent edit operations cost the same as rare ones. Therefore, we expect Levenshtein distance to behave poorly because it is unable to learn and exploit the phonotactic regularities that occur in the lexicon.

De Pauw and Wagacha (DPW). De Pauw and Wagacha (2007) proposed a morphological shallow parsing method designed for poorly endowed languages such as Gĩkũyũ, a Bantu language spoken in Kenya. They applied this method to discover the particular prefixes that are used to mark the inflection classes in this language. The idea is to use a statistical classifier based on maximum entropy to estimate morphological similarity. The classifier is used in a non-standard manner because each word in the lexicon defines a class by itself. The similarity of word Y with word X is estimated by the probability that Y belongs to the class X.

The method is original because it makes no assumption about the morphological nature of the languages: words are characterized by all n-grams of characters that appear in their written forms. The n-grams have an additional tag that indicates whether they occur at the beginning, at the end, or in the middle of the word. This information can be described by adding a `#` at the beginning and end of the written forms. For DPW and the following two measures, only n-grams of size greater than or equal to 3 were used. Figure 5 shows the features that describe the form `comparable`. The DPW method has been implemented using the csvLearner machine learning tool developed by Assaf Urieli.[4] This tool is based on the MaxEnt library of the OpenNLP project. For

#comparable#
#comparable comparable#
#comparabl comparable omparable#
#comparab comparabl omparable mparable#
#compara comparab omparabl mparable parable#
#compar compara omparab mparabl parable arable#
#compa compar ompara mparab parabl arable rable#
#comp compa ompar mpara parab arabl rable able#
#com comp ompa mpar para arab rabl able ble#
#co com omp mpa par para rab abl ble le#

Figure 5: Features that describe the form `comparable`.

[4] github.com/urieli/csvLearner.

this reason, DPW was only applied to the Small corpus, the number of classes allowed by the classifier being limited.

DPW is a lexical parser that processes bare lists of words. The features used to train the statistical classifier enable it to capture all the phonotactic regularities present in the corpus. So we expect DPW to be quite effective in finding out morphological similarity. The only weak point is that it cannot be used for an English corpus that includes the headwords of a standard machine readable dictionary such as Wiktionary. In the experiments reported below, frequency was not taken into account in order to have an identical setting for all the measures.

Proxinette. Proxinette is a morphological similarity measure designed to reduce the search space for the derivational analogies. The reduction is obtained by bringing closer words that belong to the same paradigms, namely derivational families and series, because it is within these paradigms that an entry is likely to form analogies (Hathout, 2008, 2011a). Proxinette uses the same features as DPW, namely all the *n*-grams that appear in the canonical forms of the lexemes, but in a different way. Proxinette builds a bipartite graph with the words of the lexicon on one side and the features that characterize them on the other. Each word is linked to all its features, and each feature is connected to the words that own it. The graph is weighted so that the sum of weights of the outgoing edges of each node is equal to 1. Morphological similarity is estimated by simulating the spreading of an activation. For a given entry, an activation is initiated at the node that represents it. This activation is then propagated toward the features of the entry. In a second step, the activations in the feature nodes are propagated toward the words that possess them. The words that obtain the highest activations are the most similar to the entry. The edge weights and the way the graph is traversed brings closer the words that share the largest number of common features and the most specific ones (i.e., less frequent).

Proxinette is very similar to DPW and capture the phonotactic regularities in pretty much the same way, but does not rely on statistical learning. Therefore, it is highly scalable and can be applied to very large corpora (more than 100,000 words). Proxinette does not use frequency and is expected to perform as DPW.

PHACTS. PHACTS is a model of the formation of phonotactic knowledge in speakers' minds designed by Calderone and Celata (2011, 2012). The algorithm is based on the principles of Kohonen's (1995) self-organizing maps, or SOMs. SOMs are associative memories that represent training data described by a large number of features in a smaller dimension while preserving their topological relations. Input data with similar features are represented by adjacent positions in the reduced space. PHACTS was not conceived to produce a morphological parsing or a morphological similarity measure, unlike DPW and Proxinette, but rather to reproduce the formation of phonotactic representations of phonological words. It is completely blind to morphology (because phonological words are not morphologically annotated) and is conceived to work with corpora of naturalistic language data, i.e., including lexical frequencies (and not bare word lists). PHACTS uses the same features as DPW and Proxinette. Therefore, the initial space has as many dimensions as there are *n*-grams in the lexicon written forms. The words described by these features are projected onto a map of 25×35 neurons that is 835 dimensions. The map is created by iterative learning that mimics the exposure of speakers to the stream of phonological words. Once the learning is completed, the similarity of two words is estimated by the cosine of the vectors that describe them in the reduced space. The PHACTS estimates presented in this paper were calculated by Basilio Calderone for the Small and the Large corpora.

3.4. Evaluation procedure

We use various criteria to compare the morphological measures. Their ability to capture the three types of morphological relationships is measured by their precision, recall, and f-score with respect to the derivational families, derivational series, NP-similar words, and SMS reference. These indicators were calculated on the Small and Large corpora for all measures and for the PSS reference to characterize this particular projection of the morphological structures. As a reminder, recall is the proportion of similar words in the reference that are identified by the candidate measure; precision is the proportion of correct answers; and f-score is the harmonic mean of recall and precision:

$$R = \frac{|V \cap S|}{|S|} \qquad\qquad P = \frac{|V \cap S|}{|V|} \qquad\qquad F = \frac{2\,P\,R}{P + R}$$

where R is the recall, P is the precision, and F is the f-score. V is the set of the closest neighbors with respect to the candidate measure, and S is the set of similar words in the reference resource.

The second evaluation criterion is the ability of the measures to account for the intuition presented in Section 2, namely, to give a higher similarity to family members, a lower one to series members, and a marginal one to NP-similar words. When the neighbors of an entry are ordered by decreasing similarity, they should, according to this criterion, appear grouped in three successive sub-lists, the first containing family members, the second containing series members, and third containing the NP-similar words. To assess their ability to discriminate between the three types of morphological relationnships, I used Kendall's tau, which compares ordered lists by counting the proportions of pairs that appear in the reverse order with respect to a reference resource, in this case, SMS. More specifically, I calculated the proportion of inversions among pairs of F × S, F × NP and S × NP in the Small and Large corpora for all the measures presented above and for PSS. More formally, the proportion of inversions for two sets X (first type) and Y (second type) can be calculated as follows:

$$\tau = \frac{\sum_{(x,y)\in X\times Y} inv_{XY}(x, y)}{|X \times Y|}$$

$$inv_{XY}(x, y) = \begin{cases} 1 & \text{if } x \in X,\ y \in Y \text{ and } r(x) > r(y) \\ 0 & \text{otherwise} \end{cases}$$

where $r(x)$ is the rank of x in the list of candidate neighbors.

The third criterion compares the proposed measures with the PSS reference to estimate their overall capacity to correctly grasp and order morphologically similar words, regardless of the three varieties of morphological relationships. The ability of the measures to primarily return the most similar words can be evaluated using a measure borrowed from information retrieval, precision at rank *N* for different values of *N*, namely 1, 2, 5, 10, 20, and 100. Precision at *N* is suitable because it compares the list of *N* first neighbors and the list of the *N* most similar words in PSS. *P@N* were calculated with respect to PSS in the Small and Large corpora.

4. Results

4.1. Neighborhood examples

Figures 6 and 7 show the 20 words most similar to *comparable* in the Small corpus with respect to the baselines and the four measures that have just been described. Although these examples cannot be considered representative of the behavior of the measures in the entire corpus, several observations can be made. Figure 6 clearly shows that LCPref returns family members unless they are prefixed. Because families are small sets, most of the neighbors suggested by LCPref are errors. In contrast, LCSuff returns words that belong to the derivational series. It also catches the prefixed words of the derivational family. Because series are large sets, few errors occur among the neighbors proposed by LCSuff.

LCPref	LCSuff
comparatively_B	**incomparable**_A
comparative_A	*parable*_N
*compartment*_N	inseparable_A
comparison_N	unbearable_A
compare_V	bearable_A
*compatriot*_N	*arable*_N
compatible_A	arable_A
*compassionate*_A	vulnerable_A
*compassion*_N	venerable_A
*compass*_N	unfavourable_A
*company*_N	undesirable_A
*companionship*_N	unanswerable_A
companionable_A	tolerable_A
*companion*_N	recoverable_A
*compact*_N	preferable_A
*compact*_A	pleasurable_A
*computerize*_V	miserable_A
*computer*_N	memorable_A
*compute*_V	measurable_A
*computation*_N	invulnerable_A

Figure 6: The 20 words most similar to *comparable* in the Small corpus with respect to the baselines LCPref and LCSuff. Errors are in italics.

Figure 7 highlights a difference between PHACTS and the other three measures. The former almost exclusively returns members of the derivational series and makes very few errors for this type of relationship. The other three identify more varied similarities. This difference is surprising because DPW, Proxinette, and PHACTS use the same features. In view of these examples, Proxinette seems to favor the members of the derivational family. The Levenshtein measure tends to make more errors than the other three because it is completely blind to phonotactics (see Section 3.3 supra).

Levenshtein	DPW	Proxinette	PHACTS
incomparable_A	**incomparable**_A	**incomparable**_A	honourable_A
*parable*_N	*parable*_N	**incomparably**_B	comfortable_A
compare_V	arable_A	**comparatively**_B	conceivable_A
incomparably_B	*arable*_N	**comparative**_A	commendable_A
compatible_A	inseparable_A	*parable*_N	hospitable_A
companionable_A	**compare**_V	inseparable_A	formidable_A
comparative_A	unbearable_A	*compartment*_N	considerable_A
comfortable_A	companionable_A	**comparison**_N	charitable_A
*marble*_N	compatible_A	**compare**_V	noticeable_A
*arable*_N	*company*_N	unbearable_A	measurable_A
arable_A	bearable_A	bearable_A	creditable_A
incompatible_A	durable_A	*arable*_N	contemptible_A
payable_A	*compact*_A	arable_A	deplorable_A
*comrade*_N	adorable_A	*parabolic*_A	remarkable_A
*complex*_N	vulnerable_A	*unbearably*_B	*monosyllable*_N
*complex*_A	inexorable_A	companionable_A	favourable_A
*compile*_V	**comparison**_N	compatible_A	answerable_A
capable_A	tolerable_A	*compatriot*_N	marketable_A
remarkable_A	**comparative**_A	*compassionate*_A	preferable_A
measurable_A	*compact*_N	*compassion*_N	foreseeable_A

Figure 7: The 20 most similar words to *comparable* in the Small corpus with respect to Levenshtein, DPW, Proxinette and PHACTS.

4.2. Neighborhood size

Table 2 shows the average size of the neighborhoods defined by the similarity measures presented in Section 3.3. The size of the neighborhoods was arbitrarily set for DPW and PHACTS because no threshold was applied during the calculation. I used the 500 first neighbors in the Large corpus and the 300 first in the Small corpus to obtain neighborhoods of roughly the same size as those of Levenshtein and Proxinette.

	Large	Small
PSS	421	110
LCPref	440	214
LCSuff	1003	382
Levenshtein	569	309
DPW	–	300
Proxinette	488	276
PHACTS	500	300

Table 2: Average neighborhood size

We first observe that PSS provides only a partial view of the morphological relationships when compared to SMS (compare Table 2 with Table 1 above). Two-thirds of the words in the Large corpus and approximately four-fifths in the Small corpus do not participate in any analogy (see the first row of Table 6). Table 2 shows that LCSuff is the only measure that returns an average number of neighbors that approximates the similar words in SMS. LCSuff neighbors are essentially members of the derivational series that tend to form very large sets.

4.3. Evaluation

The following three tables present an evaluation of the measures with respect to each of the three types of relationships: family, series, and NP similars.

	Large			Small		
	P	R	F	P	R	F
PSS	0.013	0.541	0.026	0.025	0.589	0.049
LCPref	0.011	0.423	0.022	0.016	0.532	0.031
LCSuff	0.003	0.284	0.006	0.003	0.101	0.005
Levenshtein	0.015	0.694	0.030	0.019	0.820	0.036
DPW	–	–	–	**0.023**	0.928	**0.044**
Proxinette	**0.023**	**0.948**	**0.045**	0.022	**0.972**	0.043
PHACTS	0.004	0.154	0.007	0.005	0.217	0.009

Table 3: Precision (P), recall (R) and f-score (F) with respect to the families of SMS

	Large			Small		
	P	R	F	P	R	F
PSS	0.982	0.422	0.590	0.981	0.314	0.476
LCPref	0.074	0.017	0.028	0.081	0.019	0.031
LCSuff	**0.512**	**0.363**	**0.425**	**0.439**	**0.108**	**0.173**
Levenshtein	0.291	0.087	0.134	0.237	0.075	0.114
DPW	–	–	–	0.354	0.107	0.164
Proxinette	0.227	0.062	0.098	0.203	0.069	0.103
PHACTS	0.283	0.082	0.128	0.274	0.098	0.144

Table 4: Precision, recall and f-score with respect to the series of SMS

The results of the three previous tables are merged in Table 6. The evaluation of the capability of the measures to separate the families from the series and the NP-similar words is shown in Table 7. Tables 8 show the precision at different ranks in the Large and Small corpora. These indicators reflect the overall usefulness of the measures and their ability to correctly order the words that are similar to the entries. This assessment is conducted with respect to PSS.

	Large			Small		
	P	R	F	P	R	F
PSS	0.013	0.042	0.020	0.016	0.021	0.018
LCPref	0.072	0.096	0.082	0.085	0.105	0.094
LCSuff	0.031	**0.153**	0.051	0.034	0.055	0.042
Levenshtein	0.060	0.107	0.077	0.077	0.125	0.095
DPW	–	–	–	0.085	0.126	0.101
Proxinette	**0.074**	0.119	**0.091**	**0.092**	**0.168**	**0.119**
PHACTS	0.042	0.077	0.054	0.059	0.119	0.079

Table 5: Precision, recall and f-score with respect to the NP similars of SMS

	Large			Small		
	P	R	F	P	R	F
PSS	1.000	0.376	0.547	1.000	0.282	0.439
LCPref	0.085	0.028	0.043	0.083	0.030	0.044
LCSuff	**0.416**	**0.341**	**0.375**	**0.287**	0.102	0.151
Levenshtein	0.221	0.093	0.131	0.172	0.085	0.114
DPW	–	–	–	0.250	**0.114**	**0.157**
Proxinette	0.201	0.073	0.107	0.171	0.083	0.112
PHACTS	0.218	0.082	0.119	0.202	0.100	0.134

Table 6: Precision, recall and f-score with respect to SMS

5. Discussion

In Table 3, we see that PSS returns just over half of the family members, which is low in view of their size: 9 on average in the Large corpus and 3 on average in the Small corpus. It also shows that Proxinette and DPW are superior to all other measures, including PSS. Proxinette returns almost all the family members. DPW is marginally more precise than Proxinette in the Small corpus and obtains a slightly higher f-score. We also observe that LCPref, the baseline dedicated to families, receives low marks because families are small sets. PHACTS inability to recover the derivational families is explained by a different reason: PHACTS relies primarily on frequent features and has low sensitivity to rare ones. It cannot learn the patterns that characterize families because stems are much less frequent than affixes and because they are subject to numerous variations while affixes are not.

In Table 4, the evaluation with respect to the series presented is surprising because LCSuff, the dedicated baseline, obtains the best results for the three indicators in the two corpora, showing that the vast majority of the derivatives are formed by suffixation. However, LCSuff has lower performance than PSS. The precision values in the first row show that almost all PSS-similar words are members of the series. The last row confirms that PHACTS captures the similarities between series members better than between family members. DPW stands out as the most effective metrics, and Proxinette stands out as the least efficient one. Coupled with the results of families, DPW seems to be the best of the non-baseline metrics.

	Large			Small		
	F×S	F×NP	S×NP	F×S	F×NP	S×NP
PSS	0.076	0.104	0.449	0.047	0.061	0.560
LCPref	**0.119**	**0.126**	0.483	**0.099**	0.103	0.499
LCsuff	0.323	0.175	0.452	0.229	**0.101**	0.517
Levenshtein	0.257	0.208	0.410	0.208	0.170	0.401
DPW	–	–	–	0.217	0.190	0.397
Proxinette	0.211	0.202	0.463	0.129	0.120	0.472
PHACTS	0.367	0.296	**0.332**	0.396	0.309	**0.389**

Table 7: Proportion of inversions.

Large

	P@1	P@2	P@5	P@10	P@20	P@50	P@100
LCPref	**0.448**	**0.518**	0.441	0.331	0.231	0.144	0.106
LCSuff	0.064	0.110	0.168	0.220	0.265	**0.351**	**0.440**
Levenshtein	0.345	0.373	0.357	0.326	0.300	0.310	0.344
Proxinette	0.310	0.459	**0.470**	**0.414**	**0.347**	0.311	0.316
PHACTS	0.015	0.027	0.068	0.117	0.158	0.231	0.306

Small

	P@1	P@2	P@5	P@10	P@20	P@50	P@100
LCPref	**0.511**	**0.550**	0.429	0.302	0.206	0.132	0.104
LCSuff	0.035	0.063	0.101	0.141	0.184	0.257	0.323
Levenshtein	0.383	0.425	0.421	0.378	0.364	0.368	0.375
DPW	0.229	0.323	0.399	0.410	**0.436**	**0.485**	**0.518**
Proxinette	0.375	0.514	**0.518**	**0.434**	0.373	0.348	0.363
PHACTS	0.019	0.045	0.135	0.192	0.261	0.346	0.418

Table 8: Precision at rank 1, 2, 5, 10, 20, 50 and 100 with respect to PSS

Table 5 shows that NP similarity is poorly captured by all the measures, including PSS. Proxinette obtains the best results overall. The way Proxinette exploits features allows it to capture partial generalizations more extensively than other measures, which seem to rely on larger phonotactic contexts to determine word similarity. Note that the good recall of LCSuff is mainly due to conversion. It changes the category of the words without changing their shape. The transformation moves the suffixed derivatives into the category of NP-similar words. Few NP similars are present in PSS because analogy only captures paradigmatic relations. The ones found in PSS are relations between bases of the same derivation that accidentally share one or more affixes. The last row shows that NP similarity does not give rise to strong enough partial generalizations to allow PHACTS to identify them.

Table 6 provides an overall assessment with respect to SMS. The importance of the series is clear. LCSuff obtains the best results in the Large corpus. In the Small corpus, DPW outweighs all the other measures, including LCSuff.

The results of LCPref and LCSuff in Table 7 are skewed by the fact that they tend to return only one type of similar words. All of their neighbors are errors, as shown in Tables 3 and 4. To a lesser extent, the same is true for PHACTS. The first row shows that in PSS, family members are generally more similar to the entry than the members of the series and the NP-similar words. In this regard, PSS is fully compatible with intuition. The last column suggests that NP-similar words are confused with members of the series because for both types, words are similar if they share one or more affixes. The inability to distinguish series members from NP similars seems common to all the measures.

In Tables 8, notice first the good performance of the baselines. LCPref obtains the best results at ranks 1 and 2, whereas LCSuff stands out beyond rank 50 in the Large corpus. Further, we see that DPW is the most efficient measure in the Small corpus from ranks 20 to 100. Proxinette obtains good results below rank 20 in the Large corpus and below 10 in the Small corpus. The advantage is most likely due to its ability to find the members of the derivational families.

In light of these results, the identification of morphologically similar words appears to be a task that is both simple and complex. It is simple because basic heuristics such as LCPref and LCSuff obtain results that are among the best. It is complex because none of the measures considered here seems able to capture morphological relationship in its entirety. The results show that the different measures are complementary. A direction for future research is an intelligent combination of the neighborhoods defined by the different measures and guided by a statistical model built using machine learning techniques.

The overall performance of the measures is average. Depending on the applications, one will focus on recall or precision. When the neighbors are directly integrated into applications such as information retrieval, it is important to have good precision. Conversely, when they are used to reduce the search space, recall becomes critical. The results presented above, including the ones in Tables 6 and 8, show that there is a substantial margin of progression. This amelioration can only be achieved by actually incorporating meaning into the measures of morphological similarity (Hathout, 2009).

This study provides an idea of the contribution of phonotactic regularities to morphological similarity. It highlights the importance of considering these regularities in the lexicon as a whole, with the best results obtained by the statistical learning method DPW. The weighting of the graph edges used by Proxinette also takes into account the distribution of the *n*-grams into the lexicon. The results show that frequency and specificity are important factors in the selection of the regularities.

The results presented in the previous section also show that non-paradigmatic relationships are marginal with respect to paradigmatic ones. NP similarities are rare and do not seem to play a role in the lexicon morphological structure. These observations confirm the paradigmatic nature of this structure.

6. Conclusion

Morphological similarity has rarely been studied. This paper discusses various aspects of this concept, including its characterization from a linguistic point of view and how it can be calculated and evaluated.

This work paves the way for a range of other studies, such as a comparison of the average similarities for the different types of neighbors and the distribution of the similarity values in the lexicon. However, the central issue remains the integration of semantic information in the calculation of similarity.

Two longer term objectives emerge from this work: the recognition of morphological similarity in the linguistic descriptions and psycholinguistic research on morphology. The adoption of this concept in these two fields is directly dependent on the design of simple tools to identify similar words and compare their similarities.

7. Acknowledgments

I thank Basilio Calderone for the many interactions we had during this research and for his help in the calculation of PHACTS similarities in the Small and Large corpora. This work has also benefited from discussions with Ludovic Tanguy and Fabio Montermini.

References

Baayen, R. H., Piepenbrock, R., Gulikers, L., 1995. The CELEX lexical database (release 2). CD-ROM, linguistic Data Consortium, Philadelphia, PA.

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. ACL, Philadelphia, PA, pp. 48–57.

Calderone, B., Celata, C., 2011. Paradigm-aware morphological categorizations. Lingue e linguaggio 2011 (2), 183–207.

Calderone, B., Celata, C., 2012. PHACTS about activation-based word similarity effects. In: Proceedings of the EACL 2012 Workshop on Computational Models of Language Acquisition and Loss. ACL, Avignon, pp. 33–37.

De Pauw, G., Wagacha, P. W., 2007. Bootstrapping morphological analysis of Gĩkũyũ using Maximum Entropy Learning. In: Proceedings of the Eighth Annual Conference of the International Speech Communication Association (Interspeech). Antwerp, Belgique.

Firth, J. R., 1957. A synopsis of linguistic theory, 1930-1955. Oxford University Press.

Gosme, J., Lepage, Y., 2011. Structure des trigrammes inconnus et lissage par analogie. In: Actes de la 18^{e} conférence annuelle sur le traitement automatique des langues naturelles (TALN-2011). ATALA, Montpellier, France.

Harris, Z., 1954. Distributional structure. Word 10 (2-3), 146–162, traduction française dans *Langages* (20) 1970.

Harris, Z., 1979. Mathematical Structures of Language. Robert E. Krieger Publishing Company, Huntington, NY.

Hathout, N., 2008. Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy. In: Proceedings of the Coling workshop Textgraphs-3. ACL, Manchester, pp. 1–8.

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

Hathout, N., 2011a. Morphonette: a paradigm-based morphological network. Lingue e linguaggio 2011 (2), 243–262.

Hathout, N., 2011b. Une approche topologique de la construction des mots : propositions théoriques et application à la préfixation en *anti-*. In: Des unités morphologiques au lexique. Hermès Science-Lavoisier, Paris, pp. 251–318.

Kohonen, T., 1995. Self-Organizing Maps. Springer Verlag, Berlin / Heidelberg.

Lazaridou, A., Marelli, M., Zamparelli, R., Baroni, M., 2013. Compositional-ly derived representations of morphologically complex words in distributional semantics. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. ACL, Sofia, Bulgaria.

Lepage, Y., 2003. De l'analogie rendant compte de la commutation en linguistique. Habilitation à diriger des recherches, Université Joseph Fourier, Grenoble.

Lepage, Y., 2004. Analogy and formal languages. Electronic Notes in Theoretical Computer Science 53, 180–191, proceedings of the the 6th Conference on Formal Grammar and the 7th on the Mathematics of Language (FG/MOL-2001).

Levenshtein, V. I., 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet physics doklady 10 (8), 707–710.

Rubenstein, H., Goodenough, J. B., 1965. Contextual correlates of synonymy. Communications of the ACM 8 (10), 627–633.

Schone, P., Jurafsky, D. S., 2000. Knowledge-free induction of morphology using latent semantic analysis. In: Proceedings of the Conference on Natural Language Learning 2000 (CoNLL-2000). Lisbon, pp. 67–72.

Schone, P., Jurafsky, D. S., 2001. Knowledge-free induction of inflectional morphologies. In: Proceedings of the North American Chapter of the Association of Computational Linguistics (NAACL-2001). Pittsburgh, PA.

Stroppa, N., 2005. Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles. Thèse de doctorat, École nationale supérieure des télécommunications, Paris.

Yvon, F., 2006. Des apprentis pour le traitement automatique des langues. Habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris.