Multilingual lexicons for related languages*

Lynne J. Cahill and Gerald Gazdar
School of Cognitive and Computing Sciences
University of Sussex
Falmer
Brighton BN1 9QH
UK
Email: {lynneca, gerald}@cogs.susx.ac.uk

Abstract

The great increase in work on the lexicon by computational and theoretical linguists throughout the 1980s has concerned itself, almost exclusively, with monolingual lexicons. Meanwhile, applied work on multilingual lexicons, mostly for machine translation (MT), has employed monolingual lexicons linked only at the level of semantics. In this paper, we argue that the traditional MT lexicon architecture, while arguably adequate for unrelated languages, makes it impossible to capture important generalisations about related languages. These generalisations, if captured, can help to produce more robust multilingual natural language processing (NLP) systems for such languages. We discuss the types of relation we find in Dutch, English and German and describe a hierarchical lexicon structure which permits sharing of information across any level of lexical representation, illustrating it with examples from English and German.

1 Introduction

In contrast to linguistic work of the 1960s and 1970s, where the lexicon was assumed to be no more than a simple word list, current NLP work mostly places the lexicon at the centre of attention, assuming that almost all of the morphology, syntax, semantics and phonology of a language is to be captured within the lexicon rather than in extralexical components. The growth in the role of the lexicon has, not surprisingly, led to an increase in the development of flexible means of representing lexical information, with work from AI in knowledge representation contributing to the development of lexical knowledge representation languages (LKRLs). Thus the use of mechanisms which permit inheritance of various kinds, originating in the “semantic nets” proposed for general purpose knowledge representation within AI, is now common in LKRLs. LKRLs, as well as theories of how best to use them, have proliferated in the past decade. Daelemans and Gazdar (1992) and Briscoe et al. (1993) bring together much of this work on the application of inheritance networks to lexical description.

However, virtually all of the theoretically oriented work on lexical knowledge representation has concentrated on monolingual lexicons. Issues relating to multilingual lexicons have largely been left to those with pressing practical applications in hand (typically MT). The bilingual lexicons that have been constructed in this context might be better described as pairs of linked monolingual lexicons. Thus the very large multilanguage lexicon projects of the present and recent past such as MultiLex (1993) and EDR (1990) have concentrated on word and sense correspondences between essentially monolingual components, and on standardisation issues. The types of lexicons used in MT systems can be broadly divided into those which link words in two languages directly and those which link each to a conceptual representation of the meaning,

---

*This work is supported by an ESRC grant Multilingual lexical knowledge representation, number R000235724, to Gazdar & Cahill.
but in either case, the links are only at the level of semantic representation and there is no information provided about any similarities between the other levels of description in the lexical entries. Even the “Shake and Bake” algorithm for MT, which makes use of a more sophisticated lexicon which “relates translationally equivalent sets of lexical signs, carrying across the semantic dependencies established by the source language analysis stage into the target language generation stage” (Poznański et al, 1995) only operates on semantic links between the languages in question. The little theoretically oriented work on multilingual lexicons that does exist is mostly concerned with sense linkage in semantic description (see, e.g., Copestake et al. 1992).

The standard architecture for multilingual lexicons is a reasonable one when the languages involved are completely unrelated, as with English and Japanese, for example. But it is inappropriate for related languages since it forces the exclusion of a wide range of information that can contribute significantly to the robustness of multilingual NLP systems. Consider lexical incompleteness. If an English speaker who knows some Japanese looks up an uncommon English word in an English/Japanese dictionary and cannot find it, then there is no more to be done. But if a German speaker who knows some Dutch cannot find an uncommon German word in a German/Dutch dictionary, then they can probably make a very good guess as to what the Dutch equivalent might be, what its syntactic properties are, how it is inflected, and how it is spelt and pronounced. The standard architecture cannot support such inferences, and so cannot provide the level of robustness that would otherwise be attainable.

2 Closely related languages

Our project is concerned with multilingual lexicons for English, Dutch and German. In this section, we illustrate areas of lexical relatedness between these three languages, as well as noting the differences that exist.

2.1 Phonology and orthography

Since they are historically very closely related, it is no surprise that the three languages share many linguistic characteristics at all levels of description, not just semantic. The phonological and orthographic form of words which come from a single root are more or less similar depending on the diversifications which have taken place in each language. Thus the English word *bath, /bɑθ/* has the German equivalent *Bad, /ba:t/* and the Dutch *bad, /bat/*. The differences can all be attributed to regular phonological changes and different orthographic conventions.

2.2 Morphology

The three languages each have significant sets of subregular verbs, many of which are among the most frequent words of the languages. These subregular verbs exhibit alternations which are very similar in the three languages. This can be seen in the case of the three verbs meaning *swim*

<table>
<thead>
<tr>
<th></th>
<th>Present</th>
<th>Past</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>zwem /zwɛm/</td>
<td>zwem /zwɛm/</td>
</tr>
<tr>
<td>English</td>
<td>swim /swɪm/</td>
<td>swam /swɛm/</td>
</tr>
<tr>
<td>German</td>
<td>schwimm /ʃvɪm/</td>
<td>schwamm /ʃvamm/</td>
</tr>
</tbody>
</table>

In all three languages, the past tense form differs from the present tense stem only in the choice of vowel and, in English and German, even the vowels in question are virtually the same. Similar correspondences can be found in many classes of subregular verbs, and the differences can often

1The work that is closest in spirit to that discussed here is that of Kameyama (1988) who describes a genuinely multilingual unification grammar for nominal expressions in Arabic, English, French, German and Japanese.

2Additional inflections indicating person and number have been ignored.
be reduced to different realisations of identical sets of distinctive features. For example, the English /æ/ and the German /a/ in the example above have slightly different realisations, but are phonologically non-distinctive – i.e. if one were substituted for the other in one of the languages, the result would not be a different word, but would simply sound like a different accent.

An example of a morphological property which is shared by two of the languages and not the third arises with the past participle prefix /g-. This is identical in Dutch and German, but does not appear at all in English.

### 2.3 Morphophonology

The morphophonology of the three languages exhibit a range of similarities. For example, the regular past tense suffix for verbs has distinct phonological realizations determined by phonological context:

<table>
<thead>
<tr>
<th></th>
<th>Root</th>
<th>Past Suffix</th>
<th>Root</th>
<th>Past Suffix</th>
<th>Root</th>
<th>Past Suffix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>-</td>
<td>voiced</td>
<td>leer</td>
<td>/leər/</td>
<td>werk</td>
<td>/werk/</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>/da/</td>
<td></td>
<td>/ta/</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>alveolar</td>
<td>voiced non-alv.</td>
<td>add</td>
<td>/æd/</td>
<td>love</td>
<td>/lɔv/</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>/d/</td>
<td></td>
<td>/t/</td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>alveolar</td>
<td>non-alveolar</td>
<td>halt</td>
<td>/haːt/</td>
<td>sag</td>
<td>/zaːɡ/</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>/at/</td>
<td></td>
<td>/t/</td>
<td></td>
</tr>
</tbody>
</table>

In English, a distinction is made between the realization of the past tense suffix after an alveolar stop (where a vowel is present in the suffix) and its realization in other environments (where the suffix is just a single consonant). In addition, there is a distinction between voiced and voiceless realizations of the single consonant suffix, the voiced /d/ occurring after voiced consonants and vowels and the voiceless /t/ after voiceless consonants. In Dutch, we see that the same voicing distinction is made, but no distinction is made for roots ending in alveolar stops. In German, on the other hand, the alveolar distinction is made, but there is no voicing distinction.

### 2.4 Syntax

In syntax, there are many similarities between the languages. The subcategorisation frames of verbs, for example, often exhibit identical argument slots and similar, if not identical, argument types. Thus the verb *bet* (Dutch *widden*, German *wetten*) can have three argument slots, as shown below:

<table>
<thead>
<tr>
<th>Subcategorization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
</tr>
<tr>
<td><em>wed</em> PP[met] NP S'</td>
</tr>
<tr>
<td><em>Ik heb met John twintig ponden gewet dat hij niet zou komen</em></td>
</tr>
<tr>
<td>English</td>
</tr>
<tr>
<td><em>bet</em> PP[um] NP NP S'</td>
</tr>
<tr>
<td><em>I bet John twenty pounds that he wouldn't come</em></td>
</tr>
<tr>
<td>German</td>
</tr>
<tr>
<td><em>wett</em> NP NP S'</td>
</tr>
<tr>
<td><em>Ich habe John um zwanzig pfund gewettet, das er nicht kommen würde</em></td>
</tr>
</tbody>
</table>

---

3Illustrated here with the Dutch verbs *leren* (to teach) and *werken* (to work), the English verbs *add*, *love* and *laugh* and the German verbs *halten* (to hold) and *sagen* (to say). The German suffix here is the past tense component of the various person/number determined past tense suffixes and is the same as the past participle suffix.

4This is, of course, a direct result of the word-final neutralisation of voicing contrasts in German.
The basic character of the subcategorization is the same for all three languages although they use case, prepositions and constituent order differently. The differences, however, are not specific to the subcategorization of this lexeme, but follow from more general lexical parameters that apply to the languages in question.

3 Hierarchical multilingual lexicons

The previous section illustrated the fact that related languages such as English, Dutch and German exhibit many similarities across their syntax, morphology, morphophonology, phonology and orthography. Capturing the similarities, as well as the differences, will facilitate the construction of more robust multilingual NLP systems. We will now consider how this can be done.

The rationale of inheritance-based monolingual lexicons requires that information be pushed as far up the hierarchy as it can go, generalising as much as possible. Thus information which applies to all words of a language will appear right at the top of the hierarchy, information common to all nouns will appear above all the individual noun entries and so on. In a multilingual lexicon, the same rationale can quite simply be extended to carry across languages. Information which is common to all three languages can be stated at a higher points in the hierarchy than that which is unique to just one of the languages. In addition, in the framework presented here, we make use of orthogonal multiple inheritance, which allows a node in the hierarchy to inherit different kinds of information from different parent nodes. With this resource, each language’s nodes can inherit a mix of information from within the language’s own hierarchy and from the common hierarchy.

To start with, let us consider the structure of a hierarchical monolingual lexicon, and then consider how it can be generalised to the multilingual case. Our examples below are based on morphology and phonology, but exactly the same basic techniques can be employed for all other aspects of lexical description.

3.1 Monolingual lexicon structure

We assume a contemporary phonological framework in which all lexical entries are defined as having a phonological structure consisting of sequences of structured syllables, a syllable consisting of an onset (the initial consonant cluster) and a rhyme, and a rhyme consisting of a peak (the vowel) and a coda (the final consonant cluster). This structure is defined at the top of the (phonological) hierarchy, and thus applies, by default, to all words. The lexicon then allows us to define the values of the relevant onsets, peaks and codas at the individual lexical entries.

In addition, we assume a rather traditional morphological framework in which all lexical entries are defined as having a morphological structure consisting of a root together with one or more affixes. Thus the default structure for a word in English is a root and a suffix and this is stated at the top of the (morphological) hierarchy. Since the inheritance of information is nonmonotonic, these general structural definitions can be overridden at any point lower down in the hierarchy. An example hierarchy for a monolingual English lexicon is shown in Figure 1 below.

Here, abstract (non-terminal) nodes are marked with solid boxes, with concrete (terminal) lexical entries marked with dashed boxes. In this example of a morphological hierarchy, we can see that nouns, verbs, adjectives, and so on, all inherit from Word. Below the Verb node, we have a mix of terminal and non-terminal nodes, with regular verbs such as ask and talk inheriting directly from Verb and sub-regular verbs, such as bring and swim inheriting from Verb.A and Verb.B nodes respectively. Thus alternation information common to a class of subregular verbs

\footnote{It may also provide a formal account of how the three languages have diverged from their common origin, an enterprise of interest to historical linguists, but not our concern here.}
multiple orthogonal inheritance, at the (terminal) lexical entry nodes. We will now consider how these different aspects of linguistic description. These hierarchies are connected, thanks to

orthography, phonology, morphology, syntax and semantics. Oversimplifying somewhat, a full monolingual lexical description involves a set of essentially disjoint hierarchies corresponding to these different aspects of linguistic description. These hierarchies are connected, thanks to multiple orthogonal inheritance, at the (terminal) lexical entry nodes. We will now consider how this architecture of layered inheritance hierarchies can be generalized to the case of multilingual lexicons for related languages.

3.2 Extending the structure to multilingual lexicons

Given \( n \) sets of monolingual hierarchies for \( n \) distinct but related languages, we can move to an architecture for a multilingual lexicon that employs \( n + 1 \) hierarchies where (i) the additional set of hierarchies contain an abstract account of what the \( n \) languages have in common, and (ii) the \( n \) sets of language-particular hierarchies can be reduced to setting the parameters peculiar to the language in question. This architecture is illustrated schematically in Figure 2 below.

This lexicon fragment for English and German subregular verbs is divided into three interacting hierarchies. Each of the language hierarchies (labelled English and German) inherits only from the common hierarchy, not from the other. Thus nodes in the English hierarchy can inherit from any nodes in the English or the common hierarchies. Given orthogonal multiple inheritance, this allows individual lexical entries to combine information unique to English with information common to English and German. As Figure 2 implies, there can be links on a number of levels.

First, we can see a fairly straightforward hierarchy at the common level, with Verb, Noun and so on inheriting from Word, and two verb classes, Verb_A and Verb_B, inheriting from Verb. There are also nodes for Bring and Swim, which inherit from these subclass nodes. The English and German hierarchies similarly show simple inheritance networks with individual verb entries (shown by the dashed boxes again) inheriting from subclass nodes and these in turn inheriting from verb nodes. These verb nodes, however, instead of inheriting from language specific word nodes, inherit from a common Verb node, which in turn inherits from a common Word node. As well as this link at the top of each hierarchy, there are links at all levels below.

Each of the subclass nodes inherits from one of the common subclass nodes, Verb_A and
Verb_2, as well as inheriting from the relevant language-specific Verb node. The amount of information inherited may vary and its precise realisation may be further determined by language-specific information, but there is nevertheless a core of information at each of these nodes which is shared. For Verb_1, the verb subclasses in the different languages behave identically in that their affixation patterns are essentially regular and the only thing that differentiates them from regular verbs in their language is that they change their rhyme in their past tense and participle forms. In English, this rhyme change is realised as /ət/ and in German as /ux/. In the case of Verb_2, we similarly have a case of affixation being essentially regular, but the peak, this time, being different in the past tense and participle forms. In this class, though, the past tense and participle forms are also different. Here we have the vowel alternation /i/ ~ /a/ ~ /u/ in English (present ~ past ~ past participle) and /i/ ~ /a/ ~ /u/ in German. These are remarkably similar — identical in the present tense and nondistinct in the past tense. The information which is shared by these forms can therefore be inherited by both languages' subclass nodes, in the form of sets of distinctive features, which are then interpreted phonologically in the context of the individual language.

The Bring and Swim nodes in the common hierarchy are used for the phonological and orthographic information which is shared by the forms in both languages, and, in these cases,
the semantics also. However, it is important to note that in this framework, there is no necessity that the words in question share semantics. They may inherit common phonology, orthography, and morphology (even syntax on occasion) without sharing semantics, as in the case of the verbs become in English, which translates to German werden and German bekommen, which translates to English receive. These words derive from the same root form historically, but have undergone semantic drift which has left them more distant in meaning than in phonology, orthography and morphology.

To illustrate the fact that related lexemes in two languages may need to inherit information about different types of linguistic property from different points in the hierarchy, consider the English verb find, and its German equivalent finden. They have essentially identical semantics, their orthography is identical and their phonology is very close, only the vowels differ. However, their morphology differs: the English past tense and participle form is found, whereas the German has the forms fund and gefunden. The English verb inherits its morphology from the verb class node it shares with bind, but the German verb inherits its morphology from the same node as the English verb swim (with the necessary distinctions in precise realisation of the vowels in each case being handled by the phonology of the relevant language).

3.3 Lexical incompleteness

Let us now consider how this architecture can provide robustness when confronted with lexical incompleteness. Imagine that we want to know the German word for forbid. We may know that the English verb bid has the German equivalent bieten, and know their inflectional patterns which are very similar. We may also know that the English verb forsake has the German equivalent versagen, that the English verb forget has the German equivalent vergessen and so on. Now, in an architecture such as that proposed here, these separate bits of information would have been generalised, so that an English verb beginning with the syllable for- will inherit the form of that syllable from an abstract node for the syllable, a node which also has the effect of specifying that the German equivalent is the syllable ver-. With this information organised in this way, we can construct a hypothesised German form by simply adding the syllable ver- onto the verb bieten, giving the form verbieten, which is the correct translation for forbid. We can infer the morphological information by inheriting from the verb bieten wholesale and the phonological and orthographic form by inheriting from the for/-ver- syllable node, and the bieten node. The relevant syntactic information can be inferred in exactly the same kind of way. The lexicon already tells us the syntax for English forbid and it also tells us the corresponding syntax in German for words of the grammatical class to which forbid belongs. It is thus reasonable to infer that the latter will correspond to the syntax of verbieten.

None of the inferences sketched above may be correct, but they are all the best possible guesses that can be made given the way English and German work and given the way they usually relate to each other. Sometimes such guesses will be wrong because of lexical idiosyncrasy to be found in one or both languages. But this will be uncommon since lexical idiosyncrasy is a feature of the most commonly used vocabulary and is rarely found in less common lexemes. It is always the latter that give rise to problems of lexical incompleteness since rational designers of NLP lexicons ensure that they contain all the most commonly occurring lexemes. It is perhaps worth noting that Zipf's (1935) law tells us, in effect, that lexical incompleteness will always

---

8Note that these are not lexical entries, and so do not have dashed boxes, but are abstract nodes containing information shared by forms in both languages. Their node names may be misleading in this respect – they could equally well have been given arbitrary language-neutral names such as Node_24 and Node_65.

9In fact, in the case of this example, they are all correct.

10Zipf's law is a well-established empirical generalization about the frequency distribution of words that says that frequency is inversely proportional to rank. Thus, for example, if the tenth most frequent word in a corpus of text occurs once in every 100 words, then we would expect the thousandth most frequent word to occur only once in every 10,000 words. Plotting frequency against rank thus gives us a doubly exponential curve. Low
be a problem for NLP lexicons. We can reduce it by building a bigger lexicon but we can never eliminate it entirely. Reduction of the problem by increasing the size of the lexicon is also subject to dramatically diminishing returns (again, this follows from the shape of the Zipf curve). And, since building multilingual lexicons is an extremely expensive process, it is likely to make much better economic sense to deal with incompleteness problems by means of a linguistically intelligent lexicon design than it is by the brute force method of increasing the size of the lexicon by one or two orders of magnitude.

4 Conclusions

We have proposed an architecture for multilingual lexical representation which aims to encode and exploit lexical similarities between closely related languages at all levels of linguistic description. We have shown how such an architecture can capture generalisations about English and German phonology, orthography and morphology, and we have noted that the same architecture can be used in a similar way to represent syntactic commonalities across languages.

There are several benefits to be gained from such an architecture for language pairs such as English/German over the traditional MT lexicon architecture. Our architecture provides a more economical encoding of lexical information and one which is easier to maintain. Just as inheritance lexicons generally provide the benefit of reduced redundancy in the lexicon and therefore more economical and transparent storage, so the use of inheritance across related languages will clearly provide the same benefits for multilingual lexicons.

Finally, there is the benefit of more intelligent approaches to lexical incompleteness. As discussed in the previous section, the architecture allows one to exploit default information from both source and target languages together with information about the default commonalities across those languages. The guesses will not always be right, but they will at least be intelligent guesses, guided by genuine linguistic principles, and in many cases, they will be close enough for comprehension by a native speaker.

References


ranked words occur with very low frequencies. The converse also holds: high ranked words occur with very high frequencies. See Baayen (1991) for useful discussion and references to the relevant literature subsequent to 1935.