Bilingual Word Association Networks *

Uta Priss, L. John Old

Napier University, School of Computing,
u.priss@napier.ac.uk, j.old@napier.ac.uk

Abstract. Bilingual word association networks can be beneficial as a tool in foreign language education because they show relationships among cognate words of different languages and correspond to structures in the mental lexicon. This paper discusses possible technologies that can be used to generate and represent word association networks.

1 Introduction

The research for this paper is motivated by a word association "game" which traces the cognates of a word in different languages. This word association game is driven by a fascination for detecting conceptual structures that appear to be inherent in natural languages. The game is played by one player starting with a word and then the other player responding with "doesn't this relate to ...". Together the players will then spin a network of words that are associated as bilingual translations (for example, English/German) or by shared etymological root or lexicographic or semantic similarity.

As an example, figure 1 shows an English/German word association network that started with the word "two". The associations of "two" that come first to mind are words, such as "twelve", "twenty" and their German counterparts; then "twilight" (i.e. two lights), "twill" and "tweed" (made from two threads), "twin", and "twig" (branching into two). The German words "Zwieback" and "Zweifel" which both associate with "zwei" (the German word for "two") are interesting because they enlarge the network by leading to other words for "twoness" which do not start with the letters "tw" or "Zw". The literal translation of "Zwieback" is "two bake" which is a literal translation of "biscuit" (via Italian), although in the modern languages "Zwieback" and "biscuit" are not direct translations of each other. The English translation of "Zweifel" is "doubt", which happens to relate to "duo" - the derivative of the Latin/Greek word for "two". Thus "Zwieback" and "Zweifel" establish a connection from "two" to other words of "twoness". The network appears to have three centres relating to the number words "two/zwei", "duo" and "bi" (the other Latin word for "two") which are connected via a bilingual connection.

Because the game is associational, the result is not deterministic. Humans do not have strict rules about etymology in what Miller et al. (1990) call the "mental lexicon", i.e. the mental representation of lexical knowledge in the mind. Any word association

* This is a preprint of a paper published in U. Priss, S. Polovina, and R. Hill (Eds.): Proceedings of ICCS 2007, Springer Verlag, LNAI 4604, 2007, p. 310-320.

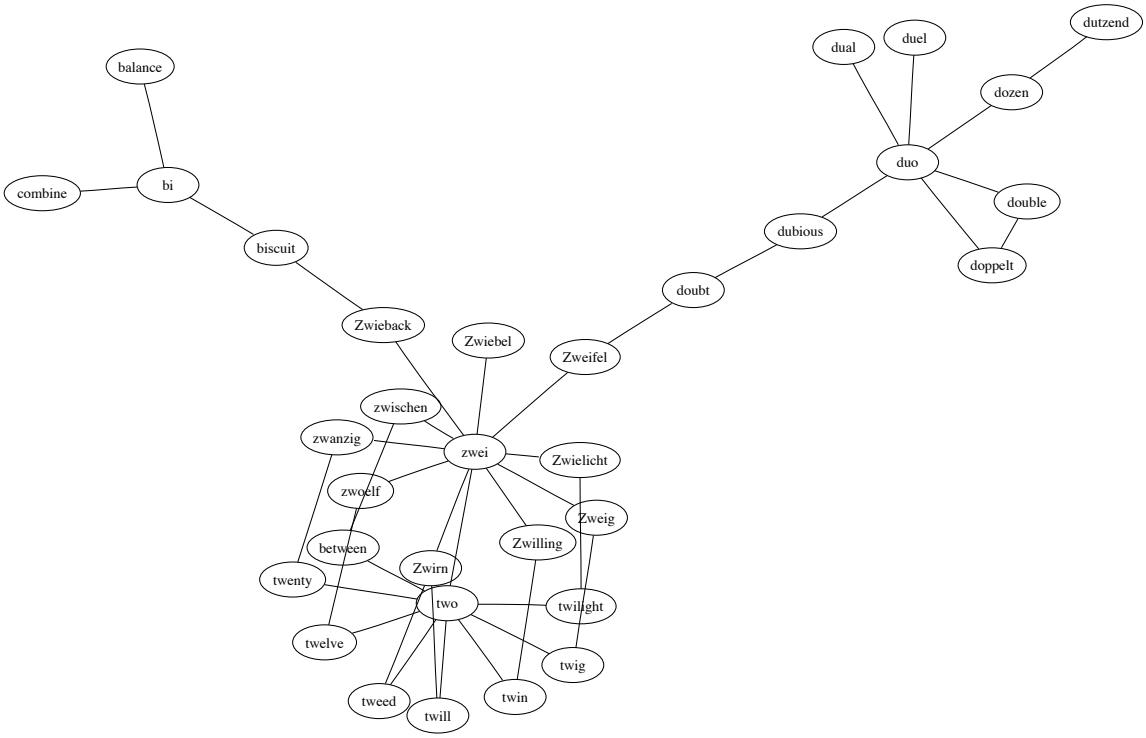

Fig. 1. A bilingual word association network

game is influenced by the mental lexicons of its players. The relations in a mental lexicon do not need to correspond to linguistically significant relations, such as etymology, but, instead, a mental lexicon is influenced by social, psychological, and cultural factors and can be idiosyncratic.

At some stage the players of a word association game may want to utilise dictionaries and thesauri to add further related words and to compare their intuitively derived associations with established lexical relations, such as etymology. In the example of figure 1, an etymological dictionary will reveal that "Zwiebel" (onion) is not etymologically related to "zwei"; and "duel" is not related to "duo" but instead to "bellum" (Latin for "war"). On the other hand, "bi", "duo" and "two" are related because they are all descendants of the Indo-European root "DWO".

So, is there a more serious background to such kinds of word association networks? Word association networks have been researched in the areas of linguistics and psychology for decades (cf. Nelson et al. (1998)). Word association tests have been used by psychologists and psychiatrists to evaluate patients. Though less popular these days as a projective/interpretive, psychoanalytic technique, they are still used to identify disorders. A stimulus or cue word is given and the response is recorded. The reaction time and type of response (for example concrete or multiword responses) can have diagnostic value. In contrast to the game described above, traditional word association studies are monolingual. But in principal there is no difference between monolingual and bilingual associations if the players have some bilingual linguistic competence.

In addition to the psychological relevance of word associations, Inkpen et al. (2005) explain that the detection of cognate words across different languages has many applications in natural language processing, for example, for sentence alignment and statistical machine translation models. An important application for bilingual word associations is the construction of tools that aid learners of foreign languages. Barrière and St-Jacques (2005) argue that a visual interface that displays word associations (or what they call "semantic context") promotes vocabulary learning. Word association tests show that native speakers store words in their mental lexicons with associations. Thus the use of explicit visualisations of word associations can help language learners to establish structures in their mental lexicons that are similar to those of native speakers. The aim for this paper is to suggest FCA-based[1] technologies for constructing and representing bilingual word association networks.

2 Definitions and technologies

Inkpen et al. (2005) suggest the following definitions: *cognates* are words from two languages that are perceived as similar and mutual translations of each other, such as "two/zwei". *Partial cognates* are words that have the same meaning in some, but not all contexts. For example, "twig" and "Zweig" are used similarly in some contexts, but in other contexts "Zweig" is better translated as "branch". Both "Zweig" and "branch" have metaphoric meanings ("a branch of a business"), which "twig" does not share. *Genetic cognates* are words that derive from the same word in an ancient language. Due

[1] Formal Concept Analysis, FCA, is a mathematical method for knowledge representation using formal contexts and concept lattices (Ganter & Wille, 1999).

to gradual changes in phonetics and meaning, genetic cognates can have very different forms and meanings in modern languages. The ancestor of genetic cognates is called a *Root*.

Linguists usually distinguish cognates from *false friends* (Inkpen et al, 2005), which have similar spellings but very different meanings. Linguists tend to exclude *lexical borrowings* from being genetic cognates because of their different historical development. Furthermore, linguists will normally distinguish genetic cognates from words that are related via *folk etymologies*, which are words that are lexicographically similar but not actually etymologically related. For example, the vegetable "Jerusalem artichoke" has no relationship with Jerusalem, but, in fact, "Jerusalem" is a folk etymology of the Italian word "girasole" for "sunflower". This means that the vegetable should really be called "sunflower artichoke" instead of "Jerusalem artichoke". For the purposes of constructing bilingual word association networks for language learners, these linguistic distinctions between cognates, false friends, borrowings and folk etymologies are not always important because these are linguistic structures which may not correspond to the conceptual structures in a mental lexicon. For the purposes of learning English, it would not matter if learners associate "Jerusalem artichokes" with Jerusalem instead of sunflowers if that helps them to learn the vocabulary.

The remainder of this section discusses technologies and lexical databases that can be used in the creation of bilingual word association networks. Priss & Old (2004) present a variety of applications of Formal Concept Analysis to lexical databases which are potentially useful for word association networks. In particular, the extraction of *semantic neighbourhoods*, i.e., fields of semantically related words, from lexical databases using *neighbourhood lattices* is relevant. Since *semantic* association networks can easily be derived from WordNet (Miller et al. 1990) and Roget's Thesaurus using neighbourhood lattices and have already been discussed elsewhere (Priss & Old, 2004), the focus of this paper is on *word* association networks. Of special interest are word association networks that establish genetic cognates (such as the one presented in figure 1). Inkpen at al. (2005) compare different algorithms that have been used to automatically identify cognates. These phonetic algorithms are usually based on orthographic similarity measures. Inkpen et al. conclude that in order to determine genetic cognates, the Soundex algorithm is most suitable among the phonetic algorithms based on orthographic similarity.

The Soundex algorithm was invented by Russell and Odell (c.f. Knuth, 1998). It codes a word into a letter followed by three numbers, where the letter is the first letter of the word and the numbers encode the remaining consonants. Consonants that are phonetically similar, such as b, f, p, v, are encoded by the same number. The algorithm was originally used by government officials, hospitals and genealogists to detect names that are similar and may be misspelled variants. For example, a patient might tell a hospital staff member that his name is "Rupert", which the staff member might erroneously record as "Robert" - both have the same Soundex code. Because the processes that lead to misspellings of names are similar to historical, phonetic changes, the Soundex algorithm can also be used to detect genetic cognates. For example, the consonant shifts that occurred during the First Germanic Sound Shift (also called Grimm's Law), which separates Proto-Germanic from other Indo-European languages, mostly pertain to con-

sonants that receive the same code in Soundex (for example, German "Vater", Latin "pater", English "father"). More details about the Soundex algorithm are provided in the Appendix.

The usefulness of the Soundex algorithm can be further improved by adapting its rules to the specific characteristics of the languages that are involved instead of using generic rules. Historical linguists might disagree with the use of the Soundex algorithm because it might yield false friends and because it lacks the contextual sensitivity that their linguistic techniques usually display. But, as discussed before, word association networks do not necessarily require the same amount of precision as historical linguistics.

Because of its simplicity, the Soundex algorithm is a powerful tool for detecting possible cognates. But unless it is combined with other technologies, its resulting association networks would be far too large. If one is interested in genetic cognates only, then the easiest approach might be to use an etymological dictionary. Unfortunately, as far as we know, it is quite difficult to obtain comprehensive etymological dictionaries that are of reasonable quality and in the public domain. The second author has complied a database of proto-language Roots and their descendants from a variety of available sources, which is used in some of the examples below (and called "ETYM" in this paper). But this database only provides Roots for a subset of the English and German words. Even in this limited database, the set of all descendants of a Root can be quite large (up to 360 words) and contain many words that are false friends in modern languages.

In order to reduce the size of the resulting networks, lexical databases, such as Roget's Thesaurus[2] or WordNet[3] can be used to identify semantic similarity among sets of phonetically similar words. Both Roget's Thesaurus and WordNet are available on-line, although in the case of Roget's Thesaurus only older versions are in the public domain (for example, from the Project Gutenberg website) because of copyright restrictions. In this paper, a relational database of Roget's Thesaurus (called "RIT") is used, which is based on Roget (1962) and not in the public domain. Another possibility is the use of bi-lingual dictionaries, which are in the public domain for many language pairs. Any of these could be combined with either the Soundex algorithm or etymological dictionaries to reduce the number of words in the resulting word association networks.

Another possible data source is word association data derived from psychological experiments. Two sets of word association data are easily available: the first set, "USF data", from the University of South Florida (Nelson, et al., 1998), is available on-line for download in the form of multiple ASCII text files. The USF data consists of about 5,000 cue words, or prompts, and about 10,000 target words, or responses. The second set of word association data, from the University of Edinburgh (Kiss et al., 1973), is available online in the form of the Edinburgh Word Association Thesaurus (EAT) and forms part of the MRC Psycholinguistic Database (Wilson, 1988). The EAT and the USF data differ in their distribution of response frequencies because word associations are culturally determined (in this case, UK versus US data). Word association data is loose and probabilistic. Responses to cue words can vary by subject, as can the fre-

[2] http://www.roget.org/

[3] http://wordnet.princeton.edu/

quency of the words or targets, given as responses. In contrast to the cue words, the target words are not a fixed set. Individual subjects may give unique (and sometimes puzzling) responses. However, usually the data is aggregated across participants. The associations of higher frequency counts have a reasonable amount of validity in their temporal, cultural contexts. The word association database (WAD) used in this paper consists of a set of 5,018 cue words and was constructed by the second author as an intersection of the British EAT and the US USF word association databases.

3 Examples

This section shows several examples of word association networks relating to the word "two" which demonstrate the use of different lexical databases and algorithms. The example in figure 2 is a neighbourhood lattice of the genetic cognates of "two" in Roget's Thesaurus, although at the category level. This is a standard construction described by Priss & Old (2004). The set of objects consists of the genetic cognates of "two". The attributes are all the categories in Roget's Thesaurus to which these words belong. The SQL statement that was used to select the data is as follows:

```
SELECT word, category FROM RIT WHERE word IN
   (SELECT descendants FROM ETYM WHERE root = 'two')
```

The labels of the attributes are omitted in the diagram in figure 2 because it is sufficient to observe the grouping of the words. This example is monolingual because Roget's Thesaurus is an English database ("zwieback" is an English word borrowed from German). Some of words seem to cluster according to similar phonetics (as in figure 1) because some words have similar meanings in addition to their similar phonetics. In summary, neighbourhood lattices from Roget's Thesaurus by themselves have some similarity with word association networks but do not appear to be sufficient to represent all of the structures that are expressed in figure 1. Furthermore, these lattices can be quite large and may need some manual editing pertaining to the level of detail to be used in Roget's Thesaurus and the manual creation of the graph layout.

In order to reduce the complexity of a neighbourhood lattice, a process of "restriction" can be applied. Figure 3 shows a "restricted neighbourhood lattice" of the lattice in figure 2 which contains only the polysemous cognates of "two" (i.e. the ones with more than one sense in Roget's Thesaurus) and only the corresponding categories from Roget's Thesaurus that contain more than one of the words from the set of formal objects. The following SQL statement was used:

```
SELECT r1.word, r1.sense INTO temptable
FROM RIT r1 WHERE r1.word IN
   (SELECT descendants FROM ETYM WHERE root = 'two')
AND EXISTS
   (SELECT * FROM RIT r2 WHERE r1.word = r2.word
   AND r1.sense != r2.sense));

SELECT t1.word, t1.sense FROM temptable t1
```

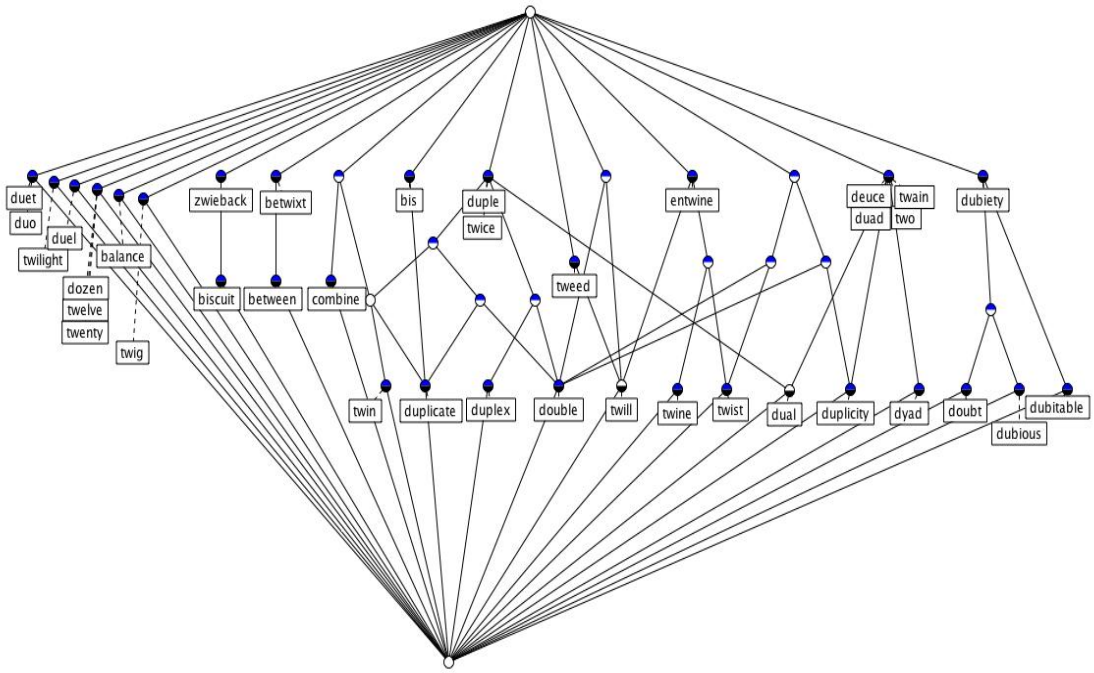

Fig. 2. A neighbourhood lattice for the genetic cognates of “two” from Roget’s Thesaurus

```
WHERE EXISTS
   (SELECT * FROM temptable t2 WHERE t1.sense = t2.sense
   AND t1.word != t2.word);
```

The restricted lattice contains the core structures from the original lattice in figure 2 and is sufficiently simple that it can be automatically constructed. But the process of restriction is not symmetric with respect to objects and attribute because a different result is obtained depending on whether the objects or the attributes are restricted first. Thus in order to automatically generate such diagrams a variety of heuristics might need to be developed depending on the number of concepts in the lattice, the depth of the lattice and possibly some other factors. The lattice in figure 3 again is monolingual, but the German translations could be added alongside their English counterparts. Because Roget's Thesaurus groups the words semantically, the German words would belong to approximately the same Roget's categories as the English ones.

Neighbourhoods and the process of restriction can also be applied to graphs in general, not just lattices. Figure 4 shows a restricted neighbourhood of "two" derived from the word association database WAD. Word association data is asymmetric – there is a difference depending on whether a word is given as the "cue" in the word association test, or whether it is the word that participants give in response to a cue (the "target"). This neighbourhood of "two" shows words that invoke "two" as a response, when they are given as cue; or are invoked as a response to "two" when it is given as the cue. The set is "restricted" so that it contains only the words that have other, mutual, associations, in addition to their association with "two". That is, they additionally invoke another, or are invoked by another, of the words associated with "two". Two is excluded for simplicity – it would be connected to every node in the network.

The arrows in figure 5 show the cue/target direction (which may be reciprocal). The strength of the arrows corresponds to the percentage of participants who produced that particular association. This diagram has been included to demonstrate that traditional word associations are quite different both from the bilingual word association networks, as in figure 1, and from the purely semantic neighbourhood lattices in Roget's Thesaurus.

The final example in figure 5 in this section combines data from Roget's Thesaurus, from the etymological database ETYM and from a bilingual dictionary with a very simple form of the Soundex algorithm. The idea for this example is that a user could manually select a list of words (in this case all the words from figure 1) and a Root, and then automatically generate a diagram. Thus, for the concept lattice in figure 5, the English and German descendants of the Indo-European Root "DWO" were derived from ETYM. The set of descendants was restricted to words where the Root describes the stem of the word instead of a prefix. This is possible because ETYM contains information about whether a Root pertains to the prefix or stem of a word. In addition all the words from figure 1 were included, i.e. the "false" descendants "Zwiebel" and "duel" were also included. Furthermore, translations were derived for most of the words using a bilingual dictionary.

The English descendants of the Indo-European Root are then chosen as formal objects and the German descendants as formal attributes. A simplified Soundex algorithm

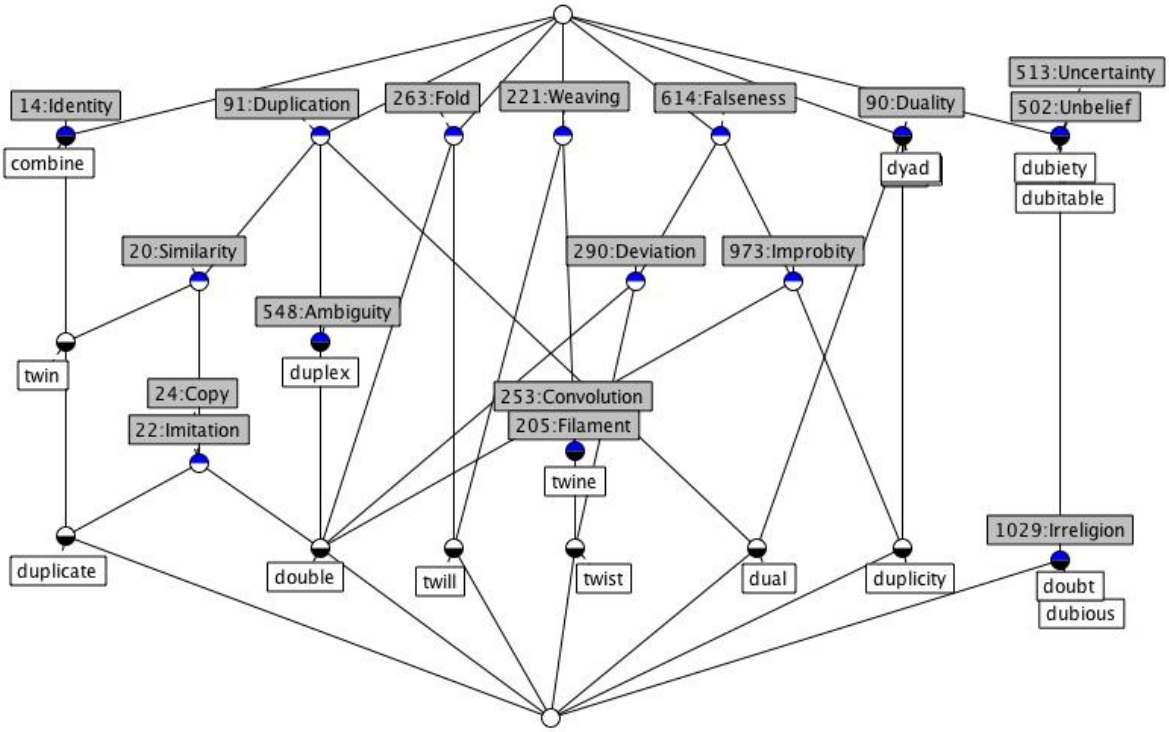

Fig. 3. A restricted neighbourhood lattice for the genetic cognates of "two" from Roget's Thesaurus

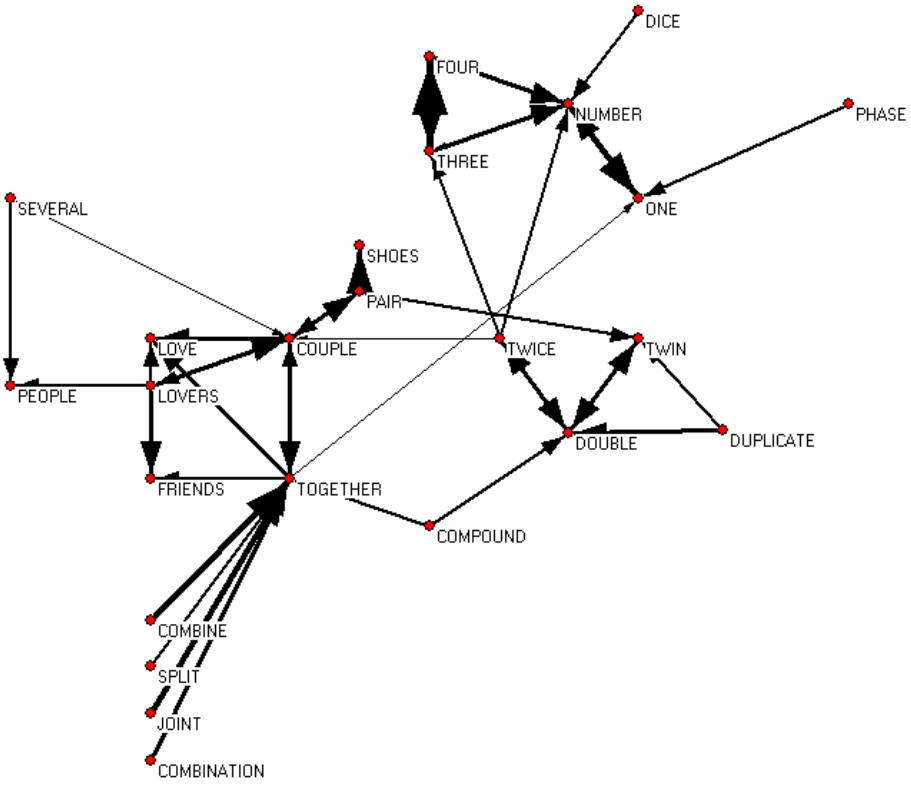

Fig. 4. Words relating to “two” derived from a restricted word association neighbourhood

is used by adding the objects and attributes “tw-”, “Zw-”, “bi-”, and “du-”, which facilitate clustering of phonetically similar words. For example, the English words starting with “tw” are assigned the additional attribute “tw-”, and so on. In the centre of the figure, there are a few words that appear to be less regular and that correspond to some of the more interesting aspects that were discussed in relation to figure 1. The pairs “Zwieback/biscuit” and “Zweifel/doubt” are cases where the German and English words have a different Soundex encoding. The pair “Zweikampf/duel” is also in this category. As mentioned before, “duel” is not a descendant of “DWO”, but “Zweikampf” is a descendant. The lattice does not clarify this fact. On the other hand, the pair “Zwiebel/onion” stands out because, in this case, the English word does not correspond to any of the given Soundex codes. It should be noted that only the words from figure 1 were manually added, but not their relationships. All relationships in figure 5 are automatically derived. Thus overall, the lattice in figure 5 presents a sufficient amount of the structures from figure 1. This example demonstrates that an approach that combines the databases, RIT, ETYM, a bilingual dictionary with a simplified Soundex algorithm appears to be most promising.

4 Conclusion

In order to automatically construct bilingual word association networks that can aid language learners, a variety of tools and technologies can be utilised. Table 1 summarises the different lexical databases and tools that facilitate the representation of genetic and semantic cognates and of conceptual structures in the mental lexicon. The previous section illustrates the use of some of these tools and databases in examples. The example in figure 5, which is based on the use of an etymological dictionary, a modified Soundex algorithm and Roget’s Thesaurus, displays structures that are most similar to the target structure in figure 1. Thus an algorithm based on these components seems most appropriate for representing bilingual word association networks. Further testing using other examples will need to be conducted in order to establish whether this strategy is successful in other areas of the lexicon.

genetic cognates	semantic cognates	mental lexicon
etymological dictionary Soundex	Roget’s Thesaurus WordNet bilingual dictionary	word association data

Table 1. Possible components in the creation of word association networks

The other examples from the previous section demonstrate that reliance on Roget’s Thesaurus or word association data by itself produces results that are quite different from the target in figure 1. In particular, networks derived from traditional word association structures are very different from the bilingual word association network in figure 1. Our intuition is that networks, as in figure 1, are more suitable for vocabulary

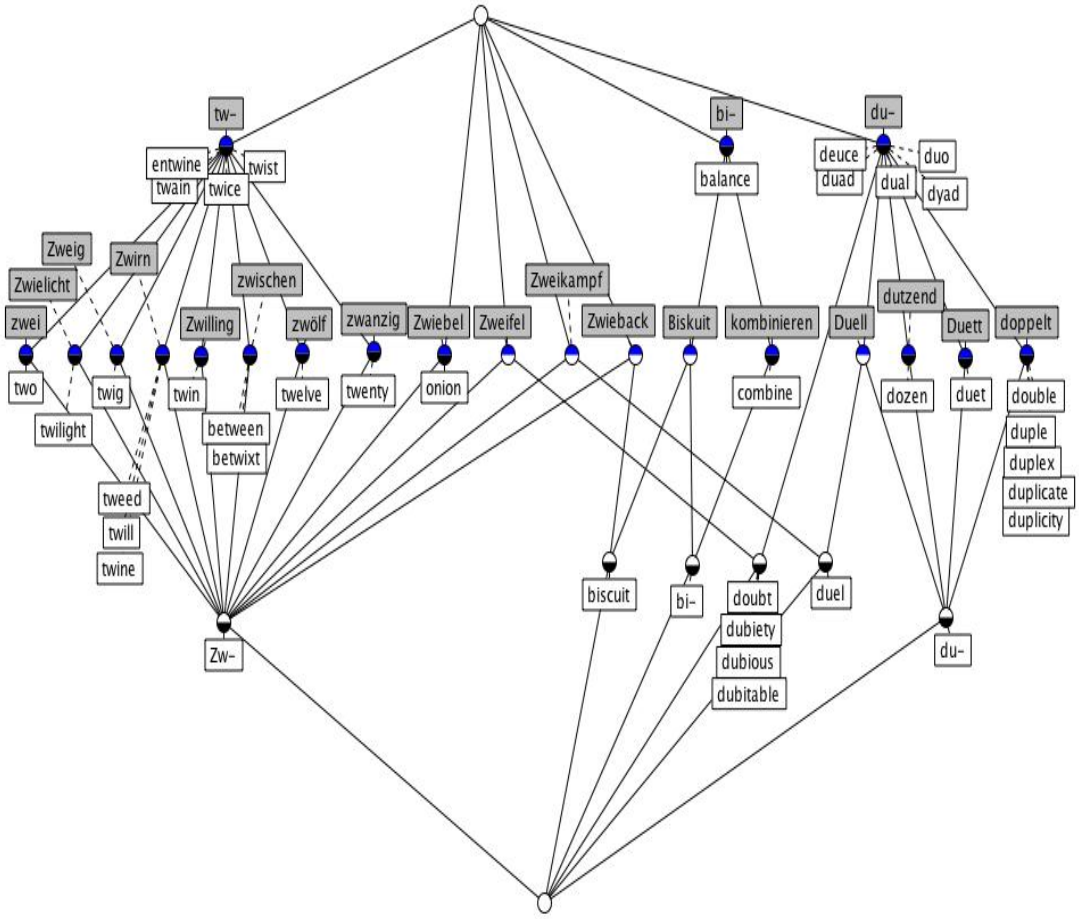

Fig. 5. A modified Soundex algorithm applied to the genetic cognates of “two”

learning than networks as in figure 4, even though networks as in figure 4 are based on psychologically derived data. The reason for this is that in modern language instruction, each lesson usually focuses on a topic. Thus some of the structures in networks as in figure 4 are implicitly contained in each language lesson anyway. Bilingual networks as in figure 1, however, allow a focus on similar but contrasting language elements. This kind of information is not usually available in modern language teaching resources and would be an additional resource.

References

1. Barrière, C.; St-Jacques, C. (2005). *Semantic Context Visualization to Promote Vocabulary Learning.* Association Computers in the Humanities & Association for Literary and Linguistic Computing Conference, p 10-12.
2. Ganter, B.; Wille, R. (1999). *Formal Concept Analysis.* Mathematical Foundations. Berlin-Heidelberg-New York: Springer, Berlin-Heidelberg.
3. Inkpen, D.; Frunza, O.; & Kondrak G. (2005). *Automatic Identification of Cognates and False Friends in French and English.* Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2005), p. 251-257.
4. Kiss, G. R.; Armstrong, C.; Milroy, R; Piper, J. (1973). *An associative thesaurus of English and its computer analysis.* In Aitken, A.J., Bailey, R.W. and Hamilton-Smith, N. (eds.), The Computer and Literary Studies. Edinburgh: University Press.
5. Knuth, D. (1998). *The Art of Computer Programming* Volume 3: Sorting and Searching (2nd Edition), Addison-Wesley Professional.
6. Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* Available at `http://w3.usf.edu/FreeAssociation/`
7. Miller, G.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K. J. (1990). *Introduction to WordNet: an on-line lexical database.* International Journal of Lexicography, 3, 4, p. 235-244.
8. Phillips, L. (2000). *The Double Metaphone Search Algorithm.* C/C++ Users Journal, Volume 18, Issue 6 (June 2000).
9. Priss, U.; Old, L. J. (2004). *Modelling Lexical Databases with Formal Concept Analysis.* Journal of Universal Computer Science, Vol 10, 8, 2004, p. 967-984.
10. Roget, Peter Mark (1962). *Roget's International Thesaurus.* 3rd Edition Thomas Crowell, New York.
11. Wilson, M.D. (1988) *The MRC Psycholinguistic Database: Machine Readable Dictionary.* Version 2. Behavioural Research Methods, Instruments and Computers, 20(1), p. 6-11.

Appendix - Soundex Algorithm

The encoding of the Soundex algorithm consists of a letter followed by three numbers: the letter is the first letter of the word, and the numbers encode the remaining consonants (up to three). Vowels and "h, w, y" are ignored unless they are the first letter. The vowels are coded loosely following Grimm's Law. For example, "b, f, p, v" are each assigned the numeric code "1". Duplicates, such as "tt", are assigned only one code or the second consonant is ignored.

The exact Soundex algorithm is as follows (see http://en.wikipedia.org/wiki/Soundex):

1. Retain the first letter of the string

2. Remove all occurrences of the following letters, unless it is the first letter: a, e, h, i, o, u, w, y
3. Assign numbers to the remaining letters (after the first) as follows:
   b, f, p, v = 1
   c, g, j, k, q, s, x, z = 2
   d, t = 3
   l = 4
   m, n = 5
   r = 6
4. If two or more letters with the same number were adjacent in the original name (before step 1), or adjacent except for any intervening h and w (American census only), then omit all but the first.
5. Return the first four characters, right-padding with zeroes if there are fewer than four.

There exist many soundex variants. The Double Metaphone algorithm (Phillips, 2000) provides significant improvements over the basic Soundex algorithm and is, for example, used to suggest spelling corrections in spell-checkers.