Res on Lang and Comput (2009) 7:137–170
DOI 10.1007/s11168-010-9068-8

Towards Unrestricted, Large-Scale Acquisition of Feature-Based Conceptual Representations from Corpus Data

Barry Devereux · Nicholas Pilkington · Thierry Poibeau · Anna Korhonen

Published online: 14 July 2010

Abstract In recent years a number of methods have been proposed for the automatic acquisition of feature-based conceptual representations from text corpora. Such methods could offer valuable support for theoretical research on conceptual representation. However, existing methods do not target the full range of concept-relation-feature triples occurring in human-generated norms (e.g. *flute produce sound*) but rather focus on concept-feature pairs (e.g. *flute – sound*) or triples involving specific relations only (e.g. *is-a* or *part-of* relations). In this article we investigate the challenges that need to be met in both methodology and evaluation when moving towards the acquisition of more comprehensive conceptual representations from corpora. In particular, we investigate the usefulness of three types of knowledge in guiding the extraction process: encyclopedic, syntactic and semantic. We present first a semantic analysis of existing, human-generated feature production norms, which reveals information about co-occurring concept and feature classes. We introduce then a novel method for large-scale feature extraction which uses the class-based information to guide the acquisition process. The method involves extracting candidate triples consisting of concepts, relations and features (e.g. *deer have antlers*, *flute produce sound*) from corpus data parsed for grammatical dependencies, and re-weighting the triples on the

B. Devereux (✉)
Centre for Speech, Language and the Brain, Department of Experimental Psychology,
University of Cambridge, Downing Street, Cambridge CB2 3EB, UK
e-mail: barry@csl.psychol.cam.ac.uk

N. Pilkington · A. Korhonen
Computer Laboratory & RCEAL, University of Cambridge,
William Gates Building, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK

T. Poibeau
Laboratoire LaTTiCe, CNRS UMR 8094 and École Normale Supérieure,
1 rue Maurice Arnoux, Montrouge 92120, France

basis of conditional probabilities calculated from our semantic analysis. We apply this method to an automatically parsed Wikipedia corpus which includes encyclopedic information and evaluate its accuracy using a number of different methods: direct evaluation against the McRae norms in terms of feature types and frequencies, human evaluation, and novel evaluation in terms of conceptual structure variables. Our investigation highlights a number of issues which require addressing in both methodology and evaluation when aiming to improve the accuracy of unconstrained feature extraction further.

Keywords Property norms · Conceptual representations · Semantic features · Corpus-based acquisition

1 Introduction

Concrete concepts like TIGER, APPLE, CHISEL and BICYCLE constitute a fundamental part of people's coherent mental representations of the world around them. A key question in cognitive science is how these semantic representations are organised and accessed. Most recent theories of conceptual representation in cognitive psychology take a componential approach to semantic representation, and assume a distributed, feature-based model of conceptual knowledge (e.g. Farah and McClelland 1991; Masson 1995; Cree et al. 1999; McRae et al. 1997; Pexman et al. 2002; Randall et al. 2004; Tyler et al. 2000). According to such theories, conceptual knowledge is distributed across a network of interconnected feature units (such as ⟨*has_eyes*⟩, ⟨*has_ears*⟩, ⟨*has_stripes*⟩) with concepts' meanings being represented as patterns of activation across these units.

Theories of distributed conceptual representation have been investigated using a wide range of methodologies, including connectionist modelling, empirical psychological experiments, and developmental and neuropsychological studies. They have been explored and tested with a variety of behavioural tasks, including similarity priming (McRae et al. 1997; Masson 1995), object naming (Pexman et al. 2002), feature verification (Cree et al. 2006; Randall et al. 2004), word recognition (Pexman et al. 2002), and categorization (Pexman et al. 2003).

The relative prominence of distributed, feature-based accounts of conceptual representation in the literature reflects the many perceived strengths of such a framework. For example, distributed accounts give a natural description of semantic similarity in terms of overlapping patterns of activation, yielding strong predictions about the mechanics of semantic priming that are consistent with the empirical evidence (McRae et al. 1997; Masson 1995). It has also been claimed that the distributed framework better reflects the physiological substrate of memory compared to other localist models (McClelland and Rumelhart 1985).

A key issue for all studies which aim to test distributed theories of concepts is the problem of accurately estimating the knowledge that people are likely to represent in such a system. Early computational work (e.g. McClelland and Rumelhart 1985; Hinton and Shallice 1991; Farah and McClelland 1991; Plaut 1997) tended to focus on small-scale representations which implemented relatively small numbers of feature

nodes and concept patterns. These toy models were explicitly constructed to test various specific predictions made by distributed accounts of conceptual knowledge; as such, they used hand-crafted representations (i.e. feature nodes and conceptual patterns defined on them) that were specifically tailored to test such predictions and did not attempt to encode any information about real-world concepts.

As connectionist models of cognition have grown more sophisticated (and computational power has increased) researchers have implemented models of conceptual representation which use empirically-derived real world conceptual knowledge (McRae et al. 1997; Cree et al. 2006; Grondin et al. 2009; Randall et al. 2004). For these models, the feature nodes together with the concept patterns represented on them are intended to reflect the actual conceptual knowledge of people, as derived empirically from property norming studies. The largest set of such norms that has been made available to date is that collected by Ken McRae and colleagues, which consists of norms for 541 living and non-living things from various categories (e.g. WEAPON, VEHICLE, BIRD, TREE). Features were collected by instructing participants to fill out forms on which they were asked to list properties of the concepts. Thirty participants listed features for each concept. McRae et al. normalised the resultant feature lists by mapping feature descriptions with the same meaning but different wordings to the same feature label. For example, the responses "used for transportation", "used for transport", "is used for transportation", "people use it for transportation", and "transportation" were all mapped to the same ⟨*used_for_transportation*⟩ feature label (see McRae et al. 2005, for details). In this way a relatively small list of normalized feature labels (on average, about 13 per concept) was constructed, with a production frequency value (i.e. the number of participants who produced that feature for that concept) associated with each feature in the list. Other researchers have gathered property norms using much the same methodology (e.g. Rosch and Mervis 1975; Randall et al. 2004; Vinson and Vigliocco 2008).

As well as being useful in deriving large scale connectionist models of real-world conceptual knowledge, property norms such as these have also been of great utility in psycholinguistic studies exploring theoretical accounts of how people represent and activate concepts. In one influential early study, Rosch and Mervis (1975) used property norming data to investigate to what extent category typicality judgements could be accounted for by semantic feature overlap. They found that concepts with features that were true of many members of the concept's category were judged more typical than concepts with features that were true of only a few members of the category (for example, the features listed for CHAIR are true of many of the other members of the FURNITURE category whereas the features for TELEPHONE are not; CHAIR was rated a more typical member of the FURNITURE category than was TELEPHONE). Property norm data therefore had a crucial role to play in developing the family resemblance account of category typicality (see Murphy 2002 for discussion of related theoretical work).

Property norms have also been used to generate experiments testing distributed theories of conceptual representation. For example, McRae et al. (1997) used measures of semantic similarity calculated from property norm data to predict semantic priming effects in tasks where participants made semantic decisions (e.g "does it grow?") to concept word targets. Pexman et al. (2002) investigated distributed conceptual

frameworks in a series of experiments that examined the role of number of features (NOF) in lexical decision (i.e. word/non-word judgement) and word-naming tasks. They found that the number of features in a concept, as measured by property norms, influenced reaction times—participants were faster to process words which had many semantic features. Similarly, Pexman et al. (2003) investigated the NOF effect with respect to various semantic decision tasks (e.g. judgements of bird/nonbird, living/non-living), again finding that NOF facilitated judgements.

Furthermore, feature production norms are useful for calculating important statistical properties of concepts (e.g. McRae et al. 1997, 2005; Cree et al. 2006; Grondin et al. 2009; Tyler and Moss 2001; Moss et al. 2007; Taylor et al. 2008). Conceptual statistics that can be calculated from property norms include *feature distinctiveness* (i.e. how many concepts a feature occurs for; ⟨*has_an_udder*⟩ is a relatively distinctive feature, occurring in only one or two concepts, whereas ⟨*has_legs*⟩ is a relatively shared feature, occurring in many concepts) and *feature correlation* (i.e. how likely it is for two features to occur together; ⟨*has_eyes*⟩ and ⟨*has_ears*⟩ are highly correlated, but ⟨*has_eyes*⟩ and ⟨*is_large*⟩ are not). Such statistics can then be used to generate predictions about different kinds of neuropsychological phenomena. For example, a common kind of semantic impairment exhibited by herpes simplex encephalitis patients is a semantic deficit for living things relative to non-living things (Tyler et al. 2000; Moss et al. 2002). The conceptual structure account (CSA; Tyler et al. 2000; Greer et al. 2001; Tyler and Moss 2001; Moss et al. 2007; Taylor et al. 2008) explains how category-specific semantic deficits emerge as a consequence of the differences in the conceptual structure of different domains of concepts. According to the CSA, distinctive features which are highly correlated with many other features are less susceptible to damage compared to distinctive features which are not highly correlated, because of the relative strength of the connections between correlated features in the distributed feature network. The distinctive features of living things tend to be poorly correlated (e.g. ⟨*has_a_mane*⟩ is not highly correlated with the other features of LION) whereas the distinctive features of non-living things tend to be highly correlated due to form-function relationships (e.g. ⟨*has_a_blade*⟩ and ⟨*used_for_cutting*⟩). The susceptibility of the distinctive features of living things to damage manifests as a deficit for living things in tasks which involve unique identification of a concept like drawing a picture of a lion, or distinguishing between a lion and a tiger. Thus, the CSA explains category-specific semantic deficits in terms of *conceptual structure* (that is, structural properties of concepts in the semantic space that can be estimated from property norms).

Feature-based representations of concepts such as those generated in norming studies have therefore had an important role to play in testing theories of conceptual knowledge (see McRae et al. 2005, for further discussion of the utility of property norms). The many and varied uses to which property norms have been put as outlined above illustrates the importance of reliable and scalable methods for generating feature-based conceptual representations. However, property norms come with several important caveats (see e.g. Murphy 2002, for a discussion). One issue is that participants tend to under-report features which are present in many of the concepts in a category (McRae et al. 2005; Murphy 2002, p. 32); for ELEPHANT for example, participants list salient features like ⟨*has_trunk*⟩ but not less salient features like ⟨*breathes*⟩

(McRae et al. 2005). The over-reporting of salient features probably reflects pragmatic concerns in participants' approach to the task: participants naturally produce the information about the concepts which they feel is most informative (e.g. the most useful in distinguishing between the concept and other similar concepts), rather than give an enumeration of features unbiased by a pragmatic desire to communicate useful information. Thus ⟨*breathes*⟩ is not listed for ELEPHANT although presumably all McRae et al.'s participants knew that elephants breathe air (indeed, WHALE is the only concept for which ⟨*breathes*⟩ is given as a feature in the McRae norms, making ⟨*breathes*⟩ a distinctive feature according to the norms). Earlier researchers (e.g. Rosch and Mervis 1975) addressed this problem by having judges decide whether a feature listed for one concept in a category was also true of other members of that category. However, for larger studies such as McRae et al.'s this becomes impractical, potentially requiring thousands of features to be verified for hundreds of concepts.

Another concern is the size of the currently available property norms. Although the largest collection of norms lists features for over 500 concepts, larger sets of norms would be useful to psycholinguists, given the number of confounding variables (word length, frequency, familiarity, imageability, etc.) that need to be controlled for in empirical studies of concepts and word meaning. The biggest practical problem in extending current norm collections is the time and cost required to produce them. For example, the norms for 541 concepts collected by McRae and colleagues have been collected over 15 years (McRae et al. 2005).

Recently, researchers have begun to develop methods which can automatically extract feature norm-like representations, comparable to the McRae norms, using corpus-based computational techniques (e.g. Almuhareb and Poesio 2004, 2005; Andrews et al. 2009; Baroni and Lenci 2008; Barbu 2008; Baroni et al. 2010). The automatic approach is cost-effective and can gather large-scale frequency data from textual corpora. As corpora contain words denoting concepts and their features in natural language, they provide ideal material for feature generation. To date, several methods have been proposed for feature extraction, mainly by the computational linguistics community (see Sect. 2 for a review). The current methods tend to target concept-feature pairs only, or are restricted to specific relations between concepts and their features (e.g. part-of relations). Unconstrained extraction of feature-norm-like concept-relation-feature triples from corpora is a very challenging task due to the wide range of potential feature and relation types. However, our goal is to investigate the extraction of such unconstrained representations for concepts, as such representations could then be used in place of manually-generated norms in psychological experiments (for example, by generating statistics from the triples that can be used to test cognitive accounts of conceptual knowledge, as discussed above). In this paper, we investigate the challenges that need to be met in both methodology and evaluation when aiming to move towards this more challenging task. With regard to methodology, we investigate the usefulness of three types of external knowledge in guiding feature extraction: encyclopedic, syntactic and semantic knowledge. We first conduct class-based analysis of the largest human-generated feature production norm set (McRae et al. 2005) which involves clustering features and concepts in semantic classes. We then introduce a novel method suitable for large-scale extraction which is the first method (to our knowledge) which extracts concept-relation-feature triples from grammatical

dependences produced by an automatic parser. The method guides the acquisition process via conditional probabilities derived from our class-based analysis which indicate the likelihood of different semantic concept and feature classes co-occurring. We apply this method to a large, automatically parsed Wikipedia corpus and evaluate its accuracy using a number of different methods: direct evaluation against the McRae norms and their expansion set, qualitative evaluation, and evaluation in terms of conceptual structure variables. Our investigation highlights a number of issues in both methodology and evaluation which are important for further development of this line of work.

This paper is structured as follows: in the following section we review the recent work in the computational linguistics and cognitive science literature on acquiring conceptual features. In Sect. 3 we describe the class-based semantic analysis of the McRae norms which provides high level information about the various kinds of features that tend to be part of the representation of different classes of concepts. Section 4 describes our novel method for feature extraction, including the learning of the dependency patterns used to extract candidate concept-relation-feature triples and how the distributional analysis was used to re-weight the likelihood of these candidate features for their respective concepts. Section 5 presents a series of evaluations of the features extracted from the Wikipedia corpus. These include a direct comparison with the McRae norms (the original norms, as well as an expansion set which supplements the McRae features with their synonyms) and complementary methods for evaluation which we argue are vital for investigating the true potential of feature extraction: manual evaluation of extracted features not included in the McRae norms, and novel evaluation in terms of conceptual structure statistics. Finally we discuss the strengths and limitations of our method and present ideas for further work in Sect. 6.

2 Background

A considerable body of work has been conducted within the computational linguistics and cognitive psychology research communities on measuring the *semantic similarity* or *semantic relatedness* between words or concepts. The corpus-based work has exploited the idea that semantically similar or related words tend to occur in similar contexts in texts. Since concepts and their features are semantically related, co-occurrence based techniques have been applied to feature extraction.

Andrews et al. (2009) have used co-occurrence statistics (like those gathered by word-space models such as LSA (Landauer et al. 1998) and HAL (Lund and Burgess 1996)) in their experiments aimed at distinguishing between experiential and distributional sources of knowledge about concepts. *Experiential data* is knowledge about concepts that is obtained through interaction with the world (as apposed to learned through language), and takes the form of properties or attributes (particularly sensory-motor and functional attributes) of concepts. For Andrews et al., experiential data is approximated by the kind of information that is gathered in property norming studies (although with the caveat that linguistic factors certainly influence such norms; see Murphy 2002, p. 106). The second source of information about concepts is *distributional data* about how words are distributed over texts in a corpus. This corresponds to the co-occurrence data employed in word-space models. The key theoretical claim

made by Andrews et al. (2009) is that people's conceptual representations are built up from a combination of both sources of information. This combination is synergistic: experiential data gives information that distributional data does not, and vice versa. In particular, Andrews et al. argue that distributional data is not grounded in the physical—it can be inferred from distributional models that the concepts LION and TIGER are similar, but these models cannot tell us what features lions and tigers have in common.

The integration of experiential and distributional data in Andrews et al's account is realised in a probabilistic framework via latent Dirichlet allocation which combines feature norm data with distributional data gathered from a corpus. In the combined model, each word is a distribution over latent variables which couple feature-clusters (i.e. latent variables extracted from the experiential data) and discourse topics (i.e. latent variables extracted from the distributional data). For evaluation, Andrews et al. calculated a distance measure between concept representations and correlated these measures with various behavioural datasets and showed that the combined model provides a better fit to the behavioural data than either a model which uses the experiential data alone or a model that uses distributional data alone. However, in spite of the model's success, it does not use an important source of information about concepts that is available to people but which is neither experiential (i.e., extra-linguistic) nor purely distributional in nature. For example, Andrews et al. state that distributional data acquired from corpora can be used to infer that CAT and DOG are similar but not that they are pets; however the knowledge that dogs and cats are pets can clearly be inferred from utterances likely to be available in large corpora (e.g. in fragments such as "pets such as dogs and cats" and "dogs make the best pets"). As Andrews et al. point out, extracting this kind of experiential, world knowledge information about concepts from corpus data is challenging.

Computational linguists have proposed various solutions to this problem. Almuhareb and Poesio (2004, 2005) have proposed a method for feature extraction which focuses on extracting information about attributes and values of concepts. In their model, each concept is described as a vector of "relations" and a relation is either a value (*red*) or an attribute (*has wheels*). Almuhareb and Poesio (2004) acquire this information from the Web using simple linguistic patterns based on two types of relations appearing in WordNet (Fellbaum 1998): hyponymy (i.e. *is-a*) relations and meronymy (i.e. *part-of*) relations. For example, a linguistic pattern might be "*the X of the Y*", where *X* is an attribute of *Y*; this pattern is instantiated in the phrase "*the price of the car*"). Clustering techniques are then used to merge the closest words to avoid data sparsity. The method is evaluated against the dataset of Lund and Burgess (1996) which provides attribute-value descriptions for 34 concepts representing three broad classes (animals, body parts, and geographical locations). Evaluated against these classes, the method achieves 97.30% accuracy for attributes and 64.86% for values.

Almuhareb and Poesio (2005) compare this pattern-based model against a model based on grammatical relations (GRs) produced by a parser. They introduce a new dataset for these experiments which groups 402 nouns into 21 clusters using the WordNet hierarchy (the 21 clusters correspond to the 21 unique beginners in WordNet; each cluster is labeled with its unique beginner). They query the Web for 10,000 snip-

pets for each unique beginner, and apply their linguistic patterns to this data. The GR-based model assumes that nouns which are syntactically related to one of the unique beginners can potentially be relevant attributes or values. The GRs are produced using the RASP parser (Briscoe and Carroll 2002). The clustering-based evaluation against the WordNet dataset shows that the pattern-based model outperforms the GR-based model. The pattern-based model performs well because the scope is restricted: the method of Almuhareb and Poesio is applicable to (and evaluated for) two types of relations between concepts and features only (*is-a* and *part-of*) and the patterns for each relation are developed manually.

Barbu (2008) combined linguistic patterns with a co-occurrence based method to extract six types of properties: superordinate, part, stuff, location, quality and action properties. Superordinate, part, stuff and location properties are learnt using a pattern-based approach where patterns are developed manually. Quality and action properties are learnt using a method that quantifies the strength of association between the nouns representing the concepts and the adjectives and verbs co-occurring with them in a corpus. The methods were evaluated on the BNC and the Web-based UkWak corpus. According to the evaluation conducted against 44 concepts from the McRae norms (the 'ESSLLI set', where the original McRae features are expanded to include synonyms which were originally lost due to the normalization process; see Sect. 5.1 and Baroni et al. (2008)), the method performs particularly well on superordinate relations, attaining 87% recall and 85% pattern precision. The scores for other relations are lower (e.g. 0% recall and 55% pattern precision for the part relation). This method is similar to the method of Almuhareb and Poesio in the sense that it assumes a range of relation types in advance, the scope is restricted to those relation types only, and specific patterns are defined manually for each relation type.

Baroni and Lenci (2008) present two methods for the extraction of feature-based conceptual descriptions: a co-occurrence based method and a linguistic pattern-based method. In the first method a co-occurrence matrix is constructed and singular value decomposition (SVD) is used to reduce this matrix down to the 21,000 most commonly occurring words in the BNC and to 125 dimensions. The second method is Structured Dimension Extraction and Labeling (Strudel). Instead of fully predefined patterns (like those employed in earlier work) Strudel makes use of shallower "connector patterns"— consisting of sequences of part-of-speech tags—to look for nouns, adjectives and verbs which appear near a target concept. The method assumes that "the variety of patterns connecting a concept and a potential property is a good indicator of the presence of a true semantic link (as opposed to simple collocational association)". Properties are thus scored based on the number of distinct patterns connecting them to a concept, rather than on the overall number of corpus co-occurrences.

Baroni et al. (2010) evaluated Strudel against the attribute value model of Almuhareb and Poesio (2004), SVD, and a dependency vector-based method. The four methods were applied to the 2 billion word UkWaC corpus, and evaluated against the ESSLLI test set. Of the models tested, the Strudel method yields the highest precision (23.9%). A number of concept categorization tasks are also used to evaluate the accuracy with which the methods identify different relation types in the McRae norms. The results show that Strudel is indeed the most accurate method, although the different methods have different strengths in terms of the relation types they discover.

An interesting discovery is that Strudel tends to focus on activities or situations people interact with while the McRae norms focus on descriptions of the physical properties of objects. This difference is reflected in categorization where Strudel shows a two-way distinction between things people use and things that move on their own. This contrasts with the traditional living/non-living distinction and shows that corpus-based techniques can be used to enrich existing models with new information.

Strudel appears to be best method of the ones compared so far. Unlike other similar methods, it has amongst its strengths the fact that it generalizes over various feature types and relies on less restricted patterns. However, its precision is not impressive, indicating that many of the concept feature tuples generated by the method are noisy. It is unlikely that further development of the shallow connector patterns will lead to significant improvements in accuracy. We believe that due to the complex, unconstrained nature of the task (the fact that concepts exhibit a wide range of feature and relation types) some knowledge-based guidance will be needed for effective filtering of noise.

Our paper investigates the usefulness of three types of knowledge for the task: encyclopedic, syntactic, and lexical-semantic. We extract features from a parsed corpus which includes encyclopedic knowledge (Wikipedia) and our technique for filtering will make use of class-based information about the types of semantic features different concept types tend to co-occur with. Unlike previous methods, our method does not rely on manually defined patterns or pattern templates. It is also the first method (to our knowledge) which targets the full range of concept-relation-feature triples (e.g. *flute produce sound*), which are more similar to what is found in human-generated norms than concept-feature pairs (*flute – sound*).

In addition, our paper investigates evaluation methods since this is critical when aiming to improve accuracy. Recent approaches have been evaluated against the ESSLLI sub-set of the McRae norms which expands the set of features in the McRae norms with their synonyms. As highlighted by Baroni et al. (2010), the McRae norms do not constitute an adequate gold standard, and the aim of the corpus-based methods is not only to replicate the McRae norms but to enrich them with additional novel information. We investigate the difference the ESSLLI set makes (when compared with the original McRae feature norms) and conduct qualitative analysis of the features which appear errors when evaluated against the norms. We also present novel evaluation in terms of conceptual structure variables, and discuss how to improve evaluation in the future.

3 Semantic Analysis of Concepts and Features

3.1 Overview

An important fact about conceptual knowledge is that the kinds of features which describe a concept are not independent of the kind of object that the concept represents. For example, an animal concept such as LION will tend to be represented in terms of features which describe body parts and behaviours (⟨*has_teeth*⟩, ⟨*has_eyes*⟩, ⟨*has_mane*⟩, ⟨*does_roar*⟩, etc.) whereas artifacts such as tools are more likely to be represented in

terms of features which describe their composition (e.g. ⟨*made of_metal*⟩) or what their intended purpose is (e.g. ⟨*used for_cutting*⟩). The probability of a particular feature being part of a concept's representation is therefore likely to be dependent on the semantic category that the concept belongs to (⟨*used for_cutting*⟩ should have low probability for concepts in the animal category, for example). The goal of the analysis we present in this section is to quantify this type of information based on the data available in the McRae norms. Our aim is to identify higher-order structure in the distribution of semantic classes for features and concepts, with the ultimate aim of investigating whether this information can be used to guide feature extraction (Sect. 4).

More formally, we assume that there is a 2-dimensional probability distribution over concept and feature classes, $P(C, F)$, where C is a concept class (e.g. *Animal*) and F is a feature class (e.g. *Body Part*). Knowing this distribution gives a way of evaluating how likely it is that a candidate feature f is true for a concept c, assuming that we know that $c \in C$ and $f \in F$. Unfortunately, we do not know the distribution P. However, we can regard the McRae norms as being a sample drawn from it, provided the concept terms and feature terms appearing in the McRae norms can be assigned to suitable concept and feature classes.

3.2 Recoding the McRae features

In order to find suitable classes we did separate cluster analyses on the concept terms and the feature terms in the McRae norms, using WordNet-based lexical similarity metrics (see Sect. 3.3). Before clustering the norms in this manner, it was necessary to recode them to a more uniform representation that was more appropriate for our computational work. Rather than use the published version of the McRae norms, we used a modified version of the norms that have been made suitable for use with native British English speakers and to which aspects of the normalization process described by McRae et al. have been applied more consistently (Randall et al. 2009). In the original norms, complex features such as ⟨*eats_bread*⟩ were decomposed into two or more features (e.g. ⟨*beh_eats*⟩ and ⟨*eats_bread*⟩); in the modified norms this decomposition step was applied to features which had not yet been decomposed, creating several new features. Several concepts that are very unfamiliar to British English speakers were removed (e.g. GOPHER, CHICKADEE), as were superordinate concepts (e.g. TOY, BUILDING). 517 concepts remain in this anglicised version of the norms.

Each concept-feature pair in the norms (e.g. TIGER ⟨*has_stripes*⟩) was then automatically recoded to a triple of the form *concept relation feature-head* where *concept* was the singular of the concept noun (e.g. 'tiger'), *relation* was the root form of a verb (e.g. 'have') and *feature-head* was always a (singular) noun or an adjective (e.g. 'stripe'). Feature-heads containing more complex information than could be captured with a single noun or adjective were split into two or more triples (for example, the norm feature ⟨*is a_musical instrument*⟩ for ACCORDION was recoded to the two triples *accordion be instrument* and *accordion be musical*). Where "beh" and "inbeh" appeared in features in the McRae norms (indicating behaviour features of animate and inanimate concepts) this was replaced with the verb "do". Prepositions and determiners were also

Table 1 Examples of recoded McRae norms

Original	Encoded
DOG ⟨*beh_barks*⟩	*dog do bark*
CAR ⟨*used for_transportation*⟩	*car use transportation*
ALLIGATOR ⟨*lives_in Florida*⟩	*alligator live florida*
BAGPIPES ⟨*used in_marching bands*⟩	*bagpipes use marching*
	bagpipes use band

removed when constructing the triples. A small number of polysemous concept words which are disambiguated in the McRae norms (e.g. "BAT (animal)" and "BAT (baseball)") are given the same concept lemma in recoding (e.g. BAT). Although recoding the norms involves a loss of information to some extent, it also enables us to clearly distinguish between the relation and feature-head parts in each norm, which allows us to perform clustering on the feature-head terms alone. Table 1 presents examples of features appearing in the anglicised McRae norms and their corresponding encoded forms.[1]

3.3 Clustering

We considered several similarity metrics in the cluster analyses, in order to find a suitable classification of concepts and features. The five similarity measures were those of Wu and Palmer (1994); Leacock et al. (1998); Resnik (1995); Lin (1998) and Jiang and Conrath (1997). Each of these methods uses the WordNet ontology as the basis for calculating similarity: concepts which fall close together under specific superordinate classes in WordNet will tend to be highly similar. The latter three measures use the principle of *information content* in the calculation of similarity scores between concepts: if the most specific common subsumer synset (in the WordNet hierarchy) of the two concepts being compared has low information content (i.e. is a very general concept) then the similarity scores will tend to be lower (see Resnik 1995, for details). Information content values are calculated from word counts in corpora; in particular, we used information content scores calculated on the British National Corpus (Burnard 2000) using "Resnik counting" (Resnik 1995). For a detailed overview and evaluation of each of these metrics, see Budanitsky and Hirst (2006). WordNet-based measures of similarity such as these are appropriate for our purposes as we are interested in generating suitable superordinate classes for which we can calculate distributional statistics; with these measures, concepts falling into a particular semantic category (e.g. REPTILES) will tend to cluster together.

[1] Note that this recoding also allows for a fairer lexical comparison of the McRae norms with the norms generated by our computational method (Sect. 5). Without such and encoding scheme, a feature such as "avocados have stones" would be scored as an incorrect production, given that "avocado has a stone" is what appears in the McRae norms. Encoding both the extracted feature and the feature in the McRae norms as *avocado have stone* allows us to identify this point of agreement.

Table 2 Example members of three concept clusters and three feature clusters

Concept clusters			Feature clusters		
Reptiles	Fruit/veg	Vehicles	Body parts	Plant parts	Events/activities
alligator	cucumber	ambulance	ear	bark	cluck
crocodile	honeydew	helicopter	foot	berry	drip
iguana	mushroom	car	fuzz	blade	emergency
rattlesnake	plum	rocket	nose	grape	flow
tortoise	tangerine	jet	small	prune	funeral
turtle	tomato	missile	tongue	trunk	grip

The set of concepts and the set of feature-head terms appearing in the recoded norms were each clustered independently using agglomerative clustering. In order to evaluate the quality of the clusters, each item in the dataset needed to be tagged with a category label. Labels were assigned automatically to the concept and feature terms using the labels from the WordNet hypernym path associated with each concept and feature, using the most frequent senses of the words in WordNet. A node depth of seven was selected for features, while a depth of eight was selected for concepts; these choices of node depth give a good compromise between having many very specific labels and a small number of very general labels. Preliminary experimentation with the different metrics indicated that the Lin metric gave slightly better results when evaluated against these node labels. We therefore selected the Lin metric as our semantic similarity measure and in the results that follow the Lin similarity metric was used. In total, 50 concept clusters and 50 feature clusters were generated. Table 2 presents three of the concept clusters and three of the feature clusters with six representative members of each cluster (we have given intuitive labels to these clusters, purely for explanatory purposes).

3.4 Calculating Conditional Probabilities

Having attained reasonably high quality clusters, we can now calculate the 2-dimensional probability distribution $P(C, F)$ and the conditional probability $P(F|C)$ of a feature cluster given a concept cluster, using the data available in the McRae norms. Table 3 gives the conditional probability for each of the three feature clusters given each of the three concept clusters that were presented in Table 2. In this table we can see, for example, that $P(BodyParts|Reptiles)$ is higher than $P(BodyParts|Vehicles)$: given a concept that is known to be a member of the *Reptiles* cluster the probability of a *Body Part* feature is relatively high whereas given a concept known to be a member of the *Vehicle* cluster the probability of a *Body Part* feature is relatively low.

Our cluster analysis therefore supports our initial hypothesis that the likelihood of a particular feature for a particular concept is not independent of the semantic categories that the concepts and features belong to. We will investigate whether this type of high-level information can be used to guide the feature extraction process. In the

Table 3 Conditional probabilities $P(F|C)$ for $C \in$ {Reptiles, Fruit/Veg, Vehicles} and $F \in$ {Body Parts, Plant Parts, Events/Activities}

	Reptiles	Fruit/veg	Vehicles
Body parts	0.164	0.031	0.023
Plant parts	0.009	0.130	0.014
Events/activities	0.100	0.060	0.140

next section, we propose (i) a base extraction method for generating features from parsed corpus data and (ii) a method which re-ranks the resulting features on the basis of the class-based analysis presented in this section. The purpose of (ii) is to decide whether a candidate feature is likely or unlikely to be a true feature of a concept. For example, if two of the candidate triples extracted for ALLIGATOR are *alligator have jaw* and *alligator have grape* we use the distributional data described in this section to give greater weight to the *alligator have jaw* possibility. Since the purpose of our experiments is merely to investigate the usefulness of semantically-driven information in guiding the purely lexical and syntax-driven feature extraction, we will maximise the accuracy by including in our clusters only McRae features and concepts. If this information proves useful for the task, future work should investigate creating more general clusters less dependent on the McRae norms.

4 Method

4.1 Wikipedia Corpora

Previous work on feature extraction has focused on general text corpora (e.g. those based on newspaper text). We chose Wikipedia as our corpus as it is a large, freely available, comprehensive and highly organised encyclopedia that includes basic information on many everyday topics. Almost all of the concepts in the McRae norms have their own Wikipedia articles, and the information given in those articles often includes information similar in nature to that which is listed by participants in norming studies (for example, the article *Elephant* describes how elephants are large, are mammals, and live in Africa). Indeed, one can view Wikipedia as an online collaborative property-norming study, where features of concepts listed by many individuals are collated into articles rather than lists of norms. Furthermore, Wikipedia tends to give canonical descriptions of things and events, rather than simply mentioning them in an incidental context. General corpora may tend to under-represent this kind of context-independent information, but they have the benefit that they are considerably larger than Wikipedia. By investigating the usefulness of Wikipedia for feature extraction, we will therefore investigate the usefulness of a smaller amount of more focused (encyclopedic) corpus data for the task.

The XML dump of Wikipedia was filtered to remove non-encyclopedic articles (disambiguation pages, articles with titles beginning with *List of*, etc). Inline references were deleted, as these typically did not contain parsable content (e.g. book

citations). The resulting data were then preprocessed with Wikiprep (Gabrilovich and Markovitch 2007), removing tables, Wikipedia infoboxes and other unparsable elements and WikiMedia mark-up. Finally, article sections that were very unlikely to contain parsable text (reference sections, picture galleries, bibliography sections, etc) were removed and each article was converted to plain text.

Two subcorpora were created from the resultant set of 1.84 million articles. The first of these simply used the Wikipedia articles that corresponded to each of the McRae concepts (so for the concept ELEPHANT, the *Elephant* article was included). We refer to this as the Wiki500 corpus as it contains approximately 500 articles (1.1 million words). Secondly, a subcorpus was created that consisted of the subset of articles which contained one of the McRae concept words in the title and which had a title that was less than five words long. The subset was limited to articles which had titles that were less than five words long in order to avoid articles on very specific topics which were unlikely to contain basic information about the target concepts (e.g. *Coptic Orthodox Church of Alexandria* for the concept CHURCH). This yielded a corpus of 109,648 plaintext articles (36.5 million words), which will be referred to as the Wiki110K corpus.

4.2 Parsing the Corpora

Most previous research on our task has used shallow-parsed corpora which are only parsed to the extent of part-of-speech tagging and lemmatization. However, many computational techniques aimed at extracting semantic information from corpora perform better when given as input corpora analysed for deeper, syntactic structures. For example, in our task, such corpora would enable us to extract concepts-relation-feature triples which occur in the same grammatical unit (as opposed to merely in the same document or sentence) and which are therefore related to each other in a meaningful way.

We parsed Wikipedia using the Robust Accurate Statistical Parsing (RASP) system (Briscoe et al. 2006). RASP is a domain-independent system which parses arbitrary English text with state-of-the-art levels of accuracy and greater depth of analysis than is possible using extant deterministic finite-state partial parsers or even treebank-based statistical parsing systems. The system has been used to automatically annotate over 1 billion words of English text in the context of published work developing lexical databases, question-answering systems, text classifiers, information extraction systems, and so on.

RASP includes a tokenizer, tagger, lemmatizer, and a wide-coverage unification-based tag-sequence parser. We use the standard scripts supplied with RASP to output the set of grammatical relations (GRs) for the most probable analysis returned by the parser or, in the case of parse failures, the GRs for the most likely sequence of subanalyses. The GRs are head-based dependencies that have been suggested as a more appropriate representation for general parser evaluation than phrase-structure trees (Carroll et al. 1998). They take the form of (GR-type optional-subtype head dependent optional-initial-GR). In practice, only non-clausal subjects take the optional initial GR field and only some modifier and complement relations take the optional subtype

field. The following example illustrates how GRs are used to represent relevant dependencies:

Kim flew to Paris from Geneva

```
(ncsubj flew Kim _) (iobj flew to) (iobj flew from)
  (dobj to Paris) (dobj from Geneva)
```

There are a total of 17 GR types included in RASP, which are organized as a subsumption hierarchy. For more information about the RASP GR output, see Briscoe (2006). GRs were extracted from all the sentences in the Wikipedia corpora and used as input to our feature extraction system.

4.3 Baseline Feature Extraction Method

We implemented as a baseline a co-occurrence-based model, based on the "SVD" (Singular Value Decomposition) model described by Baroni and colleagues (Baroni and Lenci 2008; Baroni et al. 2010). This model combines aspects of both the HAL (Landauer et al. 1998) and LSA (Lund and Burgess 1996) models in constructing representations for words based on their co-occurrences in texts. A word-by-word co-occurrence matrix was constructed for both our corpora, storing how often each target word co-occurred with each context word. The context window was defined by sentence boundaries: two words co-occur if they appear in the same sentence (note that in Baroni et al's implementation a context window of 5 (Baroni and Lenci 2008) or 20 (Baroni et al. 2010) words either side of the target word is chosen instead; we chose a sentence-based context window as it is analogous to what was used as context in our experimental method described in the next section). Context words were defined to be the 5,000 most frequent content words (i.e. words not occurring in a stop-list of function words) in the corpora. Target words were the concept names in the recoded McRae norms, supplemented with the 10,000 most frequent content words in the corpora (with the exception of the top 10 most frequent words).

Following Baroni and Lenci (2008), the dimensionality of the target-word $\times$ context-word co-occurrence matrix was reduced to 150 columns by singular value decomposition. That is, the singular value decomposition of the co-occurrence matrix was computed and the 150 left singular vectors that accounted for most of the variance, multiplied by the corresponding singular values, were used as the 150-dimensional representation of each target term. Similarity between pairs of target words was calculated as the cosine between them, and for each concept word we chose the 200 most similar target words to be the feature-head terms extracted by the model.

4.4 Our Novel Feature Extraction Method

As discussed in Sect. 2, fairly accurate features can be acquired from corpus data when the scope of the acquisition is limited to certain concept types or lexical patterns only. As our aim is to investigate the range of issues that need to be addressed when moving towards unrestricted, large scale feature extraction, we do not restrict the scope in any way, and target the full range of concept-relation-feature triples occurring in human-generated norms. Obviously, this makes our task a lot more challenging and we

do not predict a very high level of accuracy. However, we will be able to investigate the potential benefit of external knowledge (encyclopedic, syntactic and semantic) for guiding the extraction of realistic, human-like feature norms from corpus data — which is important, given the state of the art in this research area (i.e., relatively low accuracy of even the best available large-scale feature extraction systems).

4.4.1 Candidate Feature Extraction

Our method for extracting concept-relation-feature triples consists of two stages. In the first stage (which we call *candidate feature extraction*), we extract large sets of candidate triples for each target concept from the corpus. In the second stage (*re-ranking and filtering*) we aim to optimally filter out poor quality candidate triples whilst retaining the triples which are most likely to be true semantic features.

The candidate feature extraction process was deliberately designed to extract a large number of candidate triples from parsed data in order to maximise recall. For each target concept noun, all sentences containing that noun are retrieved from the corpus. The extraction process therefore operates with a context window limited to the current sentence (more specifically, the set of GRs extracted by RASP for the sentence; see Sect. 4.2). The set of grammatical relations for the sentence "Golden tabby tigers have light gold fur, pale legs and faint orange stripes." (which occurred in the *Tiger* Wikipedia article) is given below:

```
(ncsubj have tiger+s _)    (dobj have and)
(conj and fur)             (conj and leg+s)
(conj and stripe+s)        (ncmod _ stripe+s faint)
(ncmod _ stripe+s orange)  (ncmod _ leg+s pale)
(ncmod _ fur light)        (ncmod _ fur gold)
(ncmod _ tiger+s Golden)   (ncmod _ tiger+s tabby)
```

The extraction method proceeds in a similar way to that of Pado and Lapata (2007). Using breadth-first search from the target concept word, we construct an acyclic graph (i.e., mathematically a tree) of the GRs that spans the sentence and which has the target concept word as its root node. The nodes are labelled by the words occurring in the sentence and the (undirected) edges are labelled by the GR types. The graph constructed for the example above is presented in Fig. 1.

Our method considers the set of paths through the tree that exist between the target concept root node and all the other nodes in the sentence which are either an adjective or a noun; these adjectives and nouns are the potential feature heads in the concept-relation-feature triples. For each such path, if there exists a verb (i.e. a word in the sentence tagged as a verb by RASP) in the path between the target concept and the feature head, we extract the candidate triple *concept verb feature-head*. We use information about auxiliaries present in the GRs to avoid extracting auxiliary verbs as relations (so for the sentence "Tigers can eat people" we extract *tiger eat people* but not *tiger can people*). For the example sentence and tree presented above, three of the candidate triples that are extracted are *tiger have fur*, *tiger have stripe* and *tiger have leg* as the verb *have* occurs in the path between TIGER and *fur*, TIGER and *stripe*, and TIGER and *leg*, respectively. The first stage of our system therefore proceeds by extracting all possible candidate triples that can be extracted from the set of paths in

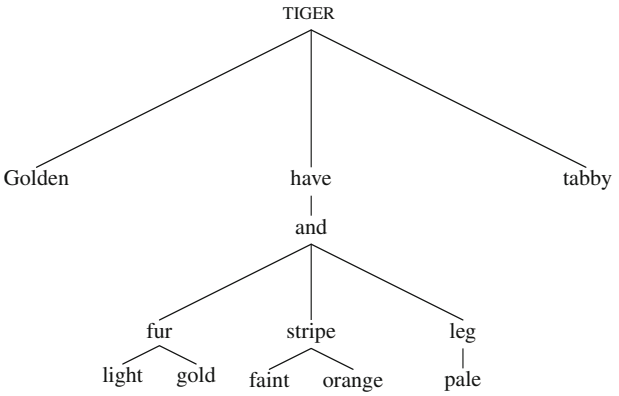

Fig. 1 The GR graph constructed for the sentence "Golden tabby tigers have light gold fur, pale legs and faint orange stripes"

the tree. The candidate feature extraction stage of our method is therefore maximally "greedy" with respect to the set of candidate triples that it can extract from the GR graph. The second stage of the method evaluates the quality of these extracted candidates on the basis of semantic information, with the aim of filtering out the poor quality features in a way which improves precision without a significant decrease in recall. We discuss this re-ranking stage in more detail in the following section.

4.4.2 Re-Ranking Based on Semantic Information

In the first stage a number of candidate concept-relation-feature triples are extracted for each concept along with their production frequencies (i.e. the number of times across the corpus each triple was extracted). The production frequency of a triple can be taken as a measure of the evidence that the triple is true for the concept; we assume that the more often a triple is extracted for a concept, the more likely it is that triple corresponds to a feature related to the concept (all else being equal). However, using the production frequency value alone as the measure of the likelihood of the feature term is problematic, because concept terms and candidate feature terms can co-occur in our GR paths for reasons other than because the feature term denotes a true semantic feature of the concept. For example, one of the extracted triples for TIGER is *tiger have squadron* (because of the RAF squadron called the Tigers).

Our semantic analysis of the McRae norms (Sect. 3) has shown that there is higher order structure in the distribution of different classes of features for different concept classes. We use the semantic analysis to improve the quality of the candidate extractions by using the conditional probabilities of the appropriate feature cluster given the concept cluster as a weighting factor. More specifically, to get the distributional data for an extracted *concept relation feature-head* triple, we find the clusters that the concept word and the feature-head word belong to. In some cases the feature-head word of the extracted triple appears in the norms and therefore the feature's cluster membership can be looked up directly from the earlier cluster analysis (Sect. 3.3). For cases where the extracted feature-head term is not in the McRae norms and therefore

was not included in our clustering, we assign the feature-head to the nearest cluster (i.e. the feature cluster with which it has the highest average similarity). Given the concept and feature clusters determined for the concepts and features in the triples, we then reweight the triples' production frequency values by multiplying them by the conditional probability of the feature cluster given the concept cluster. This re-ranking can help to eliminate incorrect triples that occur frequently in the data and can also boost the evidence for correct triples.

For example, two of the candidate features extracted for TIGER are *tiger have squadron* and *tiger have fur*. These features are extracted with equal frequency, though it is clear that only the latter is a correct semantic feature of TIGER. In the semantic analysis, the concept TIGER is assigned to a cluster that almost exclusively contains animals as members. The feature *fur* is assigned to the *Body Part* cluster identified in Sect. 3.3 and *squadron* is assigned to a cluster containing human collections (including *team*, *navy*, *clergy*, *army*, *congregation*, etc). Given the *Animal* concept cluster, the conditional probability of the *Body Part* cluster is more than three times the probability of the *Human Collection* cluster, reflecting the fact that collections of people are unlikely to be features of animals. Therefore, after the re-weighting of the features using the conditional probabilities, *tiger have fur* moves higher in the ranking of features while *tiger have squadron* is discounted.

The re-ranking of features in this manner serves as a way of dealing with a problem that is inherent to the semantic feature extraction task: that two words co-occur together in a grammatical unit does not guarantee that one of the words is a semantic feature of the other. As Baroni et al. (2010) note, TIGER and *year* co-occur together often (in the phrase "the year of the tiger") although *year* is not a feature of TIGER. Baroni et al's method discounts incorrect features such as *year* for TIGER by counting the number of types, rather than the number of tokens, of lexico-syntactic patterns that are matched by the concept-feature pair. In our method, on the other hand, the candidate triple *tiger have year* would be discounted because durations of time are unlikely to be features of animals. The re-ranking method outlined above can thus be compared to the method used by Baroni et al. (2010), which is essentially a different and complementary way of dealing with the same issue.

The method presented in this section makes use of the semantic information inferred from the McRae norms. A real-world method would ideally need to be less reliant on the norms—however, we leave the development of such a method for later as our current aim is merely to investigate the usefulness of semantic information in guiding the acquisition process.

5 Experimental Evaluation

5.1 Methods of Evaluation

We chose several different methods for evaluating the quality of the extracted triples. First of all, the precision and recall for the extracted triples was calculated, with respect to the recoded McRae norms "gold standard". Although this evaluation assumes that the McRae norms can indeed be treated as a gold standard, it is important to note that the McRae norms deviate from a true gold standard in several important

respects. Firstly, using the McRae norms as a gold standard incorrectly presumes that extracted features will have the same lexical form as semantically identical features in the norms. For example, the feature triple *avocado have stone* appears in the recoded McRae norms whilst the triple *avocado contain pit* is extracted by the method. Though semantically identical, a direct lexical comparison of these two triples will result in *avocado contain pit* being classified as an error. This particular issue is a consequence of the normalization process reported by McRae et al. (2005), where semantically identical but lexically different features generated by participants were mapped to the same normalized feature (so for example, *loud*, *noise* and *noisy* would all have been normalized to *is loud*; note that it is possible that some of McRae et al's participants may have actually written down "contains pits" for AVOCADO, and that this may have been counted as an occurrence of AVOCADO⟨*have_a_stone*⟩ during the feature normalization process). Ideally, for the purposes of evaluation, we would like to compare the extracted features to the various lexical forms of each feature that participants actually produced in the norming study, but these are not available in the published norms.

A second problem with using the McRae norms as a gold standard is that they are incomplete in the sense that they do not list all correct features of concepts. Although almost all[2] the features listed in the McRae norms are true of their respective concepts, there are many features which are also true but which are not present in the norms. As noted in Sect. 1, the feature ⟨*breathes*⟩ is listed in the norms for only one concept (WHALE) although presumably all the participants in McRae et al's study knew that all mammals breathe. When comparing our extracted features to the McRae norms, therefore, we should aim for high recall (since the features in the McRae norms are true and should ideally be extracted by the method) but not necessarily aim to maximise precision (since extracted features that are not in the norms may still be correct).

We took these issues into account by conducting several different methods of evaluation. To deal with the fact that lexically different but semantically identical features will not be classed as hits for the model, we followed the approach taken in the ESSLLI 2008 Workshop on semantic models (Baroni et al. 2008, Task 3). The gold standard for the ESSLLI task was the top 10 features for 44 of the McRae concepts: for each feature an expansion set was given, which corresponded to words which were synonyms (taken from WordNet) of the feature head word that appeared in the McRae norms. For example, the feature *lives on water* was expanded to the set {*aquatic*, *lake*, *ocean*, *river*, *sea*, *water*} (see Baroni et al. 2008, for details). We also use the relational verb in our expansion sets, giving an expansion set of triples with which to evaluate (i.e. *concept live water*, *concept live ocean*, and so on).

In the precision and recall results presented in Sect. 5.2, we include an evaluation for these 44 concepts using these feature expansion sets (the "ESSLLI 44" set). For comparison, we also include an evaluation using the top 10 features of these 44 concepts but without using the feature expansion sets ("McRae 44"), as well as an evaluation using the full set of unexpanded features for all concepts ("McRae 517").

Given that features produced by the method but which are not in the McRae norms are not necessarily errors (e.g. *tiger breathe air*) we also conducted a manual evaluation

[2] There are a small number of cases where the feature in the norms is false (e.g. TIGER ⟨*lives_in_Africa*⟩).

of some of the most highly ranked extracted features which did not appear in the McRae norms. To foreshadow our results (see Sect. 5.3), we found that a large proportion of these features were in fact either true or plausible for their corresponding concepts.

Finally, in Sect. 5.4 we introduce a novel method for evaluating the quality of the extracted features, based on analysis of the extracted feature-based semantic representations of concepts in terms of their conceptual structure properties. As mentioned in our introduction, conceptual structure statistics such as feature distinctiveness, sharedness and correlation strength have an important role to play in testing distributed theories of conceptual knowledge. Therefore, we are ultimately interested in the accuracy of the conceptual structure statistics that can be calculated from the extracted features. If the conceptual structure statistics calculated for the computationally-extracted features resemble those obtained from human-generated norms, it provides evidence, albeit indirect, that the extracted features are capturing important aspects of the semantics of concrete concepts. Furthermore, this novel evaluation method is not based on a direct comparison with the McRae norms which are, as we have noted above, a problematic standard to use for evaluating extraction quality.

5.2 Precision and Recall

Our initial evaluation of the extracted features involved comparison with the recoded McRae norms in terms of precision, recall and F-score. The *recall score* for a concept is defined as the number of extracted features for the concept that appear in the recoded norms divided by the total number of features for that concept in the norms. High recall indicates that a high proportion of the McRae features are being extracted by the model. The *precision* score for a concept is defined as the number of extracted features for the concept that appear in the norms divided by the total number of features extracted for the concept. As discussed above, we aim to maximize recall and regard precision to be a less important measure for evaluating the quality of the feature extraction. We also report the F-score as a measure of the correspondence between the extracted features and the McRae norms.

Table 4 presents a summary of the results when we evaluate based on the feature-head term alone (that is, for the purposes of calculating precision and recall we disregard the relation verb and require only a match between the feature-head terms in the extracted triples and the recoded norms). This kind of analysis is how large-scale models of feature extraction have typically been evaluated in the past. The table presents the precision, recall and F-score (averaged across concepts) for four different sets of extracted triples. The first of these is the SVD baseline model (Sect. 4.3). Secondly, we present the results for the full set of extracted triples extracted by our experimental method, prior to any filtering or re-ranking. The final two sets of extracted triples aim to illustrate the effects of re-ranking and filtering of the extracted triples. "Method—top 25% unweighted" gives the results when all but the top 25% most frequently extracted triples for each concept are filtered out. Note that the filtering criterion here is raw extraction frequency, without weighting by conditional probabilities as described in Sect. 4.4.2. "Method—top 25% weighted" are the corresponding results when the

Table 4 Results for the baseline model and the extraction method, when matching on features but not relations

Extraction set	Corpus	Test set	Precision	Recall	F-Score
SVD Baseline	Wiki500	McRae 517	0.0184	0.2734	0.0343
		McRae 44	0.0166	0.3326	0.0316
		ESSLLI 44	0.0235	0.4712	0.0448
	Wiki110K	McRae 517	0.0099	0.1475	0.0184
		McRae 44	0.0092	0.1843	0.0175
		ESSLLI 44	0.0140	0.2798	0.0266
Method—unfiltered	Wiki500	McRae 517	0.0316	0.3861	0.0529
		McRae 44	0.0169	0.4578	0.0315
		ESSLLI 44	0.0195	0.5376	0.0365
	Wiki110K	McRae 517	0.0106	0.6238	0.0198
		McRae 44	0.0047	0.7697	0.0092
		ESSLLI 44	0.0051	0.8402	0.0101
Method—top 25% unweighted	Wiki500	McRae 517	0.0565	0.1950	0.0746
		McRae 44	0.0372	0.2551	0.0586
		ESSLLI 44	0.0456	0.3141	0.0728
	Wiki110K	McRae 517	0.0237	0.4400	0.0389
		McRae 44	0.0126	0.6126	0.0238
		ESSLLI 44	0.0143	0.6922	0.0271
Method—top 25% weighted	Wiki500	McRae 517	0.0980	0.3167	0.1263
		McRae 44	0.0561	0.3803	0.0884
		ESSLLI 44	0.0647	0.4371	0.1026
	Wiki110K	McRae 517	0.0344	0.5395	0.0566
		McRae 44	0.0150	0.6351	0.0292
		ESSLLI 44	0.0172	0.7078	0.0324

See the text for a description of the columns

features are weighted by the conditional probability factors prior to filtering; that is, using all but the top 25% re-ranked features. The effectiveness of using the semantic class-based analysis data in our method can thus be assessed by comparing the filtered results with and without feature weighting. For each of these four sets of extracted triples, we present the results using both the Wiki500 and Wiki110K corpora, and using the "McRae 517", "McRae 44" and "ESSLLI 44" test sets described above. Table 5 compares the F-scores for each extraction set using the Wilcoxon signed-rank test calculated over concepts, for each extraction set, test set and corpus[3]. In Table 6, we present general statistics for both the original anglicised McRae norms and the final version of our extraction method (i.e. "Method—top 25% weighted").

Looking at the results for the baseline implementation, the first point of note is that the results are better using the smaller Wiki500 corpus compared with the larger

[3] We thank an anonymous reviewer for this suggestion.

Table 5 Comparison of F-scores reported in Table 4 (p-values from Wilcoxon signed-rank test over concepts for each pair of extraction sets)

Extraction Sets	McRae 517			McRae 44			ESSLLI 44		
	2.	3.	4.	2.	3.	4.	2.	3.	4.
Wiki500									
1. SVD	0.000	0.000	0.000	0.168	0.004	0.000	0.003	0.002	0.000
2. Method—Unfiltered		0.000	0.000		0.000	0.000		0.000	0.000
3. Method—top 25% unweighted			0.000			0.000			0.000
4. Method—top 25% weighted									
Wiki110K									
1. SVD	0.759	0.000	0.000	0.010	0.068	0.054	0.000	0.822	0.599
2. Method—Unfiltered		0.000	0.000		0.000	0.000		0.000	0.000
3. Method—top 25% unweighted			0.000			0.086			0.185
4. Method—top 25% weighted									

Table 6 Number of concepts, number of unique features (relation + feature head), the average number of features per concept, and the average number of concepts per feature, for the anglicised McRae norms and the feature set extracted by our method (top 25% of weighted features)

Extraction set	Concepts	Features	Features/concept	Concepts/feature
Anglicised McRae Norms (all)	517	2341	13.54	2.989
Anglicised McRae Norms (ESSLLI)	44	268	9.95	1.638
Method (all)	509	190609	752.51	2.010
Method (ESSLLI)	44	42321	1272.75	1.323

Statistics are give for all concepts ('all') and for the 44 concepts used in the ESSLLI evaluation subset ('ESSLLI')

Wiki110K corpus (recall that 200 features were extracted for each concept in the baseline implementation, regardless of corpus size). This is not surprising, given that the smaller corpus was constructed by manually selecting the Wikipedia articles corresponding to the concepts in the norms. This smaller corpus thus minimizes sources of noise such as word polysemy that are more apparent in the larger corpus (and in corpora generally). For example, the word "tiger" almost always refers to the animal in the Wiki500 corpus, but can have other meanings in the larger corpus and in corpora generally (the RAF squadron called the Tigers, etc).

Comparing the results for the baseline model and the unfiltered experimental method, we see that the results are quite similar for the Wiki500 corpus. Note that on average 264 features per concept were extracted for the unfiltered method for the Wiki500 corpus, compared with the 200 features-per-concept extracted for the baseline (hence, precision scores are slightly higher and recall scores are slightly lower for the baseline implementation). As mentioned in Sect. 4 our extraction method is deliberately greedy, extracting many candidate features per sentence: it is therefore not surprising that its performance is comparable to a purely co-occurrence-based

method. The main innovation of our method is that it uses the information about the GR-graph of the sentence to also extract the verb which appears in the path linking the concept word and the feature term in the sentence, something which is not possible in principle for a purely co-occurrence-based approach.

The results for the unfiltered model using the Wiki110K corpus give the maximum recall achieved by our method. For the ESSLLI 44 test set, recall is as high as 0.840: 84% of the features in the test set are being extracted by the model. It must be noted here that precision is also very low (because of the large number of features being extracted) although, as mentioned above, we are less interested in precision, particularly for the unfiltered model. In these data we can also note the advantage of expanding features to their sets of synonyms: for the McRae 44 test set, which is the same as the ESSLLI 44 test set but without the synonym expansions, recall drops to 77%.

For the results of the filtered feature sets, where all but the top 25% of features were filtered out, we see an advantage due to the semantic re-ranking for all test sets and corpora, with the re-ranked frequencies yielding higher precision and recall scores relative to the method using the raw, unweighted extracted frequencies. For example, with the unweighted features, recall drops from 62 to 44% in the unweighted case, compared to 62% to 54% when weighting factors are used (using the Wiki110K corpus and the full McRae 517 test-set). In terms of the Wilcoxon test on the F-scores, the filtered method does significantly better than the unfiltered method for all test sets tested on the Wiki500 corpus, and for the McRae 517 test set tested on the Wiki110K corpus.

In Table 7 we present the corresponding results when the extracted triples are compared to the norms using the full relation + feature-head pair (i.e. both the feature and the relation verb have to be correct). Since relation verbs are not extracted by our baseline implementation there are no results for the baseline to be reported here. Note that previous researchers have typically only compared extracted features to the feature-head term. To our knowledge our research is the first to try and compare extracted features to the full McRae norm (i.e. matching on both the relation term and the feature-head term). Unsurprisingly, matching on both relation and feature-head reduces recall (and precision) compared to the case where only the feature-head term needs be matched. For example, for the Wiki110K corpus and the ESSLLI 44 test set, recall falls from 71 to 39% for the filtered re-ranked model. However, given that the relation verb is completely unconstrained (it can be any verb in English) and that we do not have expanded synonym sets for verbs (i.e. if 'consume' is the extracted verb it will not match 'eat' in the norms) a fall in recall of only around 50% is perhaps impressive, indicating that the verb agrees with what is in the recoded norms about 50% of the time. We again note that the reranked, filtered model gives better precision, recall and F-score results than the unweighted model. In terms of Wilcoxon signed-ranked test on F-scores, performance is better for the reranked model for every combination of corpus and test set evaluated (Table 8).

A qualitative inspection of the extracted triples reveals that there are a number of correct features that are being extracted but which do not appear in the norms. These relations are thus marked incorrect in terms of the gold standard, lowering precision and recall scores. As we mentioned in our introduction, treating the McRae norms as a

Table 7 Results for the extraction method, when matching on both relations and features

Extraction set	Corpus	Test set	Precision	Recall	F-Score
Method—unfiltered	Wiki500	McRae 517	0.0112	0.1440	0.0188
		McRae 44	0.0052	0.1659	0.0100
		ESSLLI 44	0.0060	0.1909	0.0114
	Wiki110K	McRae 517	0.0043	0.2922	0.0085
		McRae 44	0.0019	0.3937	0.0038
		ESSLLI 44	0.0022	0.4596	0.0044
Method—top 25% unweighted	Wiki500	McRae 517	0.0160	0.0616	0.0219
		McRae 44	0.0101	0.0818	0.0170
		ESSLLI 44	0.0128	0.0977	0.0212
	Wiki110K	McRae 517	0.0089	0.1904	0.0142
		McRae 44	0.0051	0.2846	0.0097
		ESSLLI 44	0.0060	0.3369	0.0115
Method—top 25% weighted	Wiki500	McRae 517	0.0327	0.1160	0.0438
		McRae 44	0.0156	0.1318	0.0266
		ESSLLI 44	0.0194	0.1591	0.0327
	Wiki110K	McRae 517	0.0139	0.2568	0.0227
		McRae 44	0.0064	0.3323	0.0122
		ESSLLI 44	0.0075	0.3869	0.0143

See the text for a description of the columns

Table 8 Comparison of F-scores reported in Table 7 (p-values from Wilcoxon signed-rank test over concepts for each pair of extraction sets

Extraction Sets	McRae 517			McRae 44			ESSLLI 44		
	2.	3.	4.	2.	3.	4.	2.	3.	4.
Wiki500									
1. SVD	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2. Method—Unfiltered		.002	.000		.016	.000		.001	.000
3. Method—top 25% unweighted			.000			.000			.000
4. Method—top 25% weighted									
Wiki110K									
1. SVD	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2. Method—Unfiltered		.000	.000		.000	.000		.000	.000
3. Method—top 25% unweighted			.000			.004			.003
4. Method—top 25% weighted									

gold standard is unsatisfactory as there are many true semantic features which will not be reported by participants in feature elicitation tasks. In the next section, we present a manual evaluation of these features which aims to determine the extent of this issue.

5.3 Manual Evaluation Analysis

A key motivation for developing NLP technology for learning the features of concepts from corpus data is the need to enrich existing models of human conceptual structure with additional features missing in the available norms. To evaluate the ability of the method to learn this type of novel data, 10 concepts (representing different concept types, both animate and inanimate) were selected at random from the norms. For each of the 10 concepts, we selected the top 20 extracted triples (after reranking and using the Wiki110K corpus) which where missing from the McRae norms and thus classified as errors in our precision and recall evaluation above. Two native English-speaking judges, cognitive scientists who were naïve as to the purpose of the study, were requested to evaluate whether these were genuine errors or valid data missing in the gold standard. They were told that triples described semantic features of concepts and consisted of a concept noun, a relation (verb lemma), and a feature term (adjective or noun lemma). They were instructed that correct triples need not be true of all instances of the concept (e.g. *tiger live circus*) and may be missing prepositions (e.g. *accordion wear chest* instead of *accordion worn on chest*).

The judges were asked to select between four possibilities for each triple: *correct* ('c') when the triple represented a correct, valid, feature; *plausible* ('p') when the triple was plausible in a very specific set of circumstances and/or was correct but very general; *wrong but related* ('r') when the triple was wrong, but there existed some kind of relationship between the concept and the relation and/or feature; or *wrong* ('w') when the triple was simply wrong.

Table 9 shows the lists of top 20 "incorrect" triples for the concepts ANT and CABBAGE, and Table 10 shows the results of the evaluation for the 10 concepts that were rated. Interannotator agreement was not particularly high (Cohen's $\kappa = 0.264$) which is not surprising given the subjective nature of the rating options (particularly the choice between "correct" and "plausible", and between "wrong" and "wrong but related"). With "correct" and "plausible" both coded as "correct" and "wrong" and "wrong but related" both coded as "incorrect", interannotator agreement increases to $\kappa = 0.347$, which is considered a fair level of agreement (Landis and Koch 1977).

Overall, 38% of the triples were considered correct or plausible, and 62% were judged either wrong or wrong but related. In other words, many of the 'errors' were not really errors but actually valid triples not occurring in the norms. This demonstrates the potential of the method in enriching existing models of conceptual representation.

The genuine errors provide interesting material for further research. The most common error type concerns an incorrect relation (given the feature). For example, *filling* might be a plausible feature of CABBAGE if the relation was *used as*, not *tuck*. This is a consequence of how relation verbs are extracted in the method; any verb (other than auxiliaries) in the path from the concept node to the feature node has the potential to be the verb in a triple. By learning constraints on where in the paths and with what GR edges relation verbs can occur, the accuracy of the extracted verbs may be improved. Furthermore, the verbs often underspecify the relational link between concept and feature, because our triple-based encoding does not include prepositions (e.g. *be building* should be *be in building*).

Table 9 Manual evaluation of the top 20 "incorrect" features for the concepts ANT and CABBAGE

ANT				CABBAGE			
relation	feature	Judge A	Judge B	relation	feature	Judge A	Judge B
eat	insect	w	c	produce	soil	r	c
have	queen	c	c	call	white	r	c
cultivate	fungus	c	w	be	plant	c	c
be	predator	w	w	be	red	c	c
import	red	r	w	be	cool	w	w
be	food	p	p	find	purple	w	c
be	size	w	r	be	soup	c	c
hatch	queen	r	c	include	vegetable	c	c
be	frog	w	w	make	soup	p	c
consume	frog	w	w	cook	pork	p	c
be	building	c	w	shred	coleslaw	c	c
be	jack	w	w	call	vegetable	r	c
have	butterfly	w	w	diet	soup	p	c
be	male	c	c	roll	meat	w	p
be	hospital	w	w	benefit	food	r	c
make	plant	w	r	resemble	name	w	w
be	animal	c	r	include	food	r	c
be	soldier	c	c	wrap	filling	w	c
have	tree	w	r	fill	meat	p	p
ward	predator	w	p	tuck	filling	w	r

Note that these lists do not include extracted triples that are also found in the recoded McRae norms

Table 10 Manual evaluation results

Judgement	Judge A		Judge B		Average	
	Count	%	Count	%	Count	%
Wrong	107	53.5	43	21.5	75.0	37.50
Wrong but related	43	21.5	55	27.5	49.0	24.50
Plausible	13	6.5	26	13.0	19.5	9.75
Correct	37	18.5	76	38.0	56.5	28.25

Also some of the features were clearly incorrect, like *jack* with ANT. In future work, we aim to conduct more large scale manual evaluation studies, perhaps using online resources such as Amazon's Mechanical Turk[4]. This will allow us to examine the most common patterns of error and thus guide future improvements of the method.

[4] We thank an anonymous reviewer for this suggestion.

5.4 Evaluation in Terms of Conceptual Structure Variables

As we noted in our introduction, of particular interest to the distributed, feature-based theories of conceptual knowledge is how relationships which exist between the features that compose concepts influence semantic processing. Statistics capturing such relationships have proven useful in testing theories of distributed semantic representation, including the conceptual structure account (Randall et al. 2004; Tyler et al. 2000). Researchers have calculated several variables from property norm data which capture various aspects of the structural organization of the semantic space(e.g. McRae et al. 1997, 2005; Randall et al. 2004; Taylor et al. 2008). In this section, we propose a novel method for evaluating feature extraction methods which is based on testing whether conceptual structure statistics calculated from the extracted features exhibit similar qualities to those calculated on the McRae norms.

Various kinds of conceptual structure statistics can be calculated. These include:

Number of features. The number of features present in a concept's representation (i.e. the number of features with non-zero production frequency).

Feature sharedness/distinctiveness. Highly shared features occur in many concepts (e.g. ⟨*has_legs*⟩), highly distinctive features occur in relatively few concepts (e.g. ⟨*has_an udder*⟩). The reciprocal of the number of concepts that the feature occurs in is taken to be a measure of the feature's distinctiveness (so a feature occurring in two concepts has a distinctiveness of 0.5). In particular, we define a feature to be *distinctive* if it occurs in only one or two concepts and *shared* if it occurs in more than two concepts. For each concept, we can then calculate the number of shared features, number of distinctive features, mean distinctiveness of features and the proportion of shared features.

Feature correlation. A measure of the degree of interconnection between a pair of features in terms of their co-occurrence in concepts. For example, the features ⟨*has_eyes*⟩ and ⟨*has_ears*⟩ occur together in concepts more often than do the features ⟨*is_gray*⟩ and ⟨*has_teeth*⟩. Feature correlation for a pair of features is calculated as the Pearson correlation of their production frequencies across all the concepts. We only calculate correlation strength for pairs of shared features, as some researchers have argued that correlations for distinctive features may be spurious (Cree et al. 2006). For each concept, we can then calculate how correlated overall its constituent features are. We calculate two such measures. The first is *intercorrelational density*; it is defined as the sum of the proportions of shared variance for each pair of shared features in the concept where the shared variance is greater than 6.5% (McRae et al. 2005). However, as this measure is a sum rather than a mean, it additionally reflects the number of pairs of correlated features in the concept rather than how correlated, on average, a concept's features are (see Taylor et al. 2008, for a discussion). We therefore also define a *mean correlational strength* measure, which is the *mean* Pearson correlation of all significantly correlated pairs of features in the concept.

This yields a total of seven conceptual structure variables: number of features, number of shared features, number of distinctive features, mean distinctiveness of features, proportion of shared features, intercorrelational density and mean correlational strength.

Different kinds of concepts tend to have different conceptual structure properties (they tend to differ in the number of features they have, in the mean distinctiveness of their features, and so on). For example, the features of living things tend to be more shared than the features of non-living things. It is interesting to consider whether the conceptual structure statistics calculated on the features extracted from a computational method agree with those that are calculated on the McRae norms. Evidence that they do would indicate that the extraction method is, like the McRae norms, capturing important structural properties of the conceptual space. It would also point to the potential usefulness of these features to research on conceptual representation in cognitive psychology.

Furthermore, we propose that conceptual structure statistics give researchers a way of evaluating computational models of feature extraction that does not depend on a direct comparison to the McRae norms (i.e. we avoid treating the set of features in the McRae norms as a gold standard, which, as we have argued above, is problematic). Instead, we test whether the conceptual space generated from the norms and the conceptual space generated for the extracted features exhibit the same structural properties. Indeed, the extracted features could be superficially quite different to the McRae features (with different terms used to represent them; e.g. the feature ⟨*has_a_stone*⟩ for the concept AVOCADO in the norms and the triple *avocado contain pit* in the extracted features) but the conceptual structure statistics calculated for the concepts could still be quite similar.

In calculating the conceptual structure statistics, we selected the top 10% of the reranked extracted features across all concepts.[5] The resultant feature space consisted of 59,325 feature dimensions. Each of the seven conceptual structure variables were calculated for both these extracted features and for the original anglicised McRae norms. Variables were log-transformed where this improved normality and retransformed means are reported below. A small number of concept words with multiple meanings (e.g. BAT, FAN) were excluded from the analysis.

In the case of NOF, there is a significant correlation between the measure for the model and for the norms ($r = 0.172$, $p < 0.001$). In the McRae Norms, living things have significantly higher NOF than nonliving things. However, for the extracted triples there are more features for non-living things than living things ($M_{living} = 59.5$, $M_{nonliving} = 118.8$; $t(452.9) = 6.04$, $p < 0.001$).

For NODF (number of distinctive features), there is a significant correlation between the norms and the extracted features ($r = 0.248$, $p < 0.001$). In the McRae Norms, nonliving things have significantly higher NODF. The same pattern emerges for the extracted features, with a significant difference between living and nonliving in the predicted direction ($M_{living} = 26.8$, $M_{nonliving} = 55.0$; $t(462.2) = 6.01$, $p = 0.001$).

For NOSF (number of shared features), however, there was no significant correlation between the extracted features and the norms ($r = 0.001$, $p = 0.983$),

[5] Selecting the top 10% of reranked features removes the low quality features and makes computing the conceptual structure statistics computationally tractable. This step is analogous to the removal of features with production frequencies less than 5, which was a step taken by McRae et al prior to calculating conceptual structure statistics.

nor is the difference between living and nonliving domains in the expected direction ($M_{living} = 33.2$, $M_{nonliving} = 63.8$; $t(440.2) = 5.93$, $p = 0.001$). This is a result of there being more features for nonliving compared to living things, which consequently means there will also be more shared features, all else being equal.

For mean distinctiveness, there is a significant correlation between the extracted features and the norms ($r = 0.188$, $p < 0.001$). As in the McRae norms, mean distinctiveness of nonliving things is higher than for living things, a difference which is marginally significant ($M_{living} = 0.45$, $M_{nonliving} = 0.47$; $t(356.8) = 1.84$, $p = 0.066$).

For the proportion of shared features in a concept's representation, there is a significant correlation between the extracted triples and the norms ($r = 0.174$, $p < 0.001$). The proportion of shared features is higher for living things than for nonliving things, as is the case for the McRae norms, though this difference is not significant ($M_{living} = 0.56$, $M_{nonliving} = 0.54$; $t(371.7) = 1.48$, $p = 0.139$).

For mean correlational strength, there is a significant correlation between the extracted features and the McRae norms ($r = 0.112$, $p = 0.014$). For the living and non-living domains the difference is significant but in the opposite direction to that predicted by the McRae norms ($M_{living} = 0.22$, $M_{nonliving} = 0.24$; $t(373.1) = 3.71$, $p < 0.001$).

For intercorrelational density, there is a significant negative correlation with the McRae norms ($r = -0.117$, $p = 0.008$) and a significant difference between domains in the unexpected direction ($M_{living} = 16.4$, $M_{nonliving} = 73.4$; $t(474.8) = 7.31$, $p < 0.001$). As mentioned above, intercorrelational density is calculated as the sum rather than the mean of pairwise feature shared variance, and so concepts with more features will tend to have higher intercorrelational density (because there are simply more pairs of features in the concept to be summed over).

Overall, there is a significant positive correlation between the McRae norms and the extracted triples for five of the seven conceptual structure variables. This is important as it indicates that the semantic representations generated from the extracted features are capturing some aspects of the conceptual structure that is present in the norms. On the other hand, some of the correlations are quite weak, and for intercorrelational density the correlation is negative. Also, we do not see expected differences between living and non-living domains (which have been found in many sets of property norms). We therefore do not claim that the current set of extracted features are of good enough quality to be used in psychological experiments (for example, experiments investigating conceptual structure or perceptual/functional accounts of conceptual representation). What we wish to highlight here is the potential usefulness of conceptual structure statistics as a means for evaluating models: further improvements to the extraction method should yield better quality conceptual structure statistics.

6 Discussion

Our paper has investigated large-scale, unconstrained acquisition of human-like feature norms from corpus data. Unlike previous similar investigations, our work was not restricted to a subset of concepts or relation types or concept-feature tuples only. Instead, we targeted the full set of concepts, relations and features, and extracted

norm-like concept-relation-feature triples. High accuracy extraction of this type of challenging data from corpora is unrealistic given the state of the art (see Sect. 2). The goal of our experiment was to investigate issues in both methodology and evaluation which need to be addressed when aiming towards higher accuracy, unconstrained feature extraction in the future. In particular, we have examined the usefulness of three types of external knowledge for guiding unconstrained feature extraction: encyclopedic, syntactic, and lexical-semantic. We have also compared different approaches to evaluation: direct evaluation against existing (both original and extended) norms, qualitative analysis, and evaluation against conceptual structure variables.

Our investigation showed that encyclopedic information can be useful for feature extraction: we obtained relatively good results although our Wikipedia corpora are considerably smaller than standard text corpora: the larger subset of Wikipedia which we examined consists of 36.5 million words, while the UkWak corpus (recently employed for feature extraction by Baroni et al. (2010)) consists of 2 billion words. Computational linguistic techniques tend to perform better on large corpora. Given this, future work should investigate the usefulness of the entire Wikipedia for feature extraction, and the benefit of combining Wikipedia with a large general corpus such as UkWak.

We also showed the benefit of parsing for feature extraction: our base extraction method operating on parsed data performs better than the baseline method which makes use of collocation data. Furthermore, using GRs allows us to extract relation verbs from the GR graph. Previous work (Baroni et al. 2010; Almuhareb and Poesio 2005) have suggested that parsed data are not useful for feature extraction. Our results demonstrate the usefulness of grammatically meaningful data for semantically meaningful feature extraction when the scope of acquisition is focused on maximally useful GRs. This finding is in line with much of the recent research in computational lexical semantics. However, future work should investigate the usefulness of a wider range of GR information for the task. For example, our current extraction method is quite unconstrained, requiring only that a path exists between the concept term and a feature term. Some paths will be more indicative of semantic relationships than others; in future work we aim to learn which kinds of paths are the best indicators that the candidate feature term is indeed a semantic feature of the concept word.

Further to our GR-based extraction method, we investigated the usefulness of lexical-semantic information for the task. We showed that semantically meaningful clusters of concepts and features can be constructed from existing norms via WordNet-based similarity measures, and that information about co-occurring concept and feature clusters can be used to guide the acquisition process: our re-ranking method which involved re-weighting of features based on distributional data obtained from the semantic analysis improved the results further. While our experiment has demonstrated the usefulness of lexical-semantic information for feature extraction, future work should investigate optimal ways of integrating such information in the task. We focused on clustering McRae concepts and features only, but real-world application of the method would ideally employ other knowledge sources and cluster relation verbs as well. Advanced techniques for clustering verbs exist that can be used here (e.g. Vlachos et al. 2009; Ó Séaghdha and Copestake 2008; Sun et al. 2008).

In sum, our work has shown that encyclopedic, syntactic and semantic information can all be useful for feature extraction. Whilst additional research is required before these sources of information can be exploited in the task in an optimal manner, our investigation has revealed an important avenue for future work—one which is likely to lead to considerable improvements in accuracy. The task of unconstrained extraction of concept-relation-feature triples from corpus data is in particular so challenging that significant improvements in accuracy are unlikely without the guidance of external (e.g. semantic) information. Yet this challenging task is the one which needs to be addressed when hoping to truly enrich existing norms.

In addition, our paper has investigated evaluation which is a critical but difficult issue for feature extraction. Most recent approaches have been evaluated against the ESSLLI sub-set of the McRae norms which expands the set of features in the McRae set with their synonyms. As our evaluation shows, the ESSLLI sub-set constitutes a considerably better gold standard than the original McRae set, even though it limits us to evaluating features for only 44 concepts. Yet even the ESSLLI set does not facilitate adequate evaluation. The McRae norms, even when supplemented with information about synonyms, are not complete in the sense that there are true features which are not included in the norms—which is precisely the reason why one wishes to acquire additional features from corpus data. Our qualitative analysis shows that many of the errors against the norms are in fact correct or plausible features. Moreover, the new evaluation we have conducted in terms of the conceptual structure variables acts as a task-based evaluation showing that the data we have extracted from corpora is indeed a lot more useful than what we could infer on the basis of gold standard evaluation. Future work should aim for larger-scale qualitative evaluation using multiple human judges as well as look into other task-based evaluations. One possibility here is evaluation using fMRI or EEG data that has been collected for concepts. In such an evaluation, the goal is to train a classifier to correctly predict the activation associated with a particular concept (following Mitchell et al. 2008). The input to the classifier is a corpus-based semantic model; i.e. semantic features for each concept extracted from a corpus. Better semantic models should have greater predictive accuracy, and the predictive accuracy of the classifier when different models are given as input therefore serves as a means for evaluating the quality of the extracted features produced by the different models. Like the conceptual structure evaluation we have proposed, this approach does not require comparison with a "gold standard" set of features such as the McRae norms. This kind of evaluation has been recently presented by Murphy et al. (2009).

Whilst the development of optimal methodology and evaluation methods for feature extraction is challenging, the investigation we have reported in this paper opens up a number of avenues that could be pursued further in order to advance the state of the art in this research area.

Acknowledgments This research was supported by EPSRC grant EP/F030061/1 and the Royal Society, UK. We are grateful to Ken McRae and colleagues for making their norms publically available, and to Aurelie Herbelot for help with the Wikipedia parsing. Our work uses the Natural Language ToolKit for Python (Bird et al. 2009).

References

Almuhareb, A., & Poesio, M. (2004). Attribute-based and value-based clustering: An evaluation. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP 2004)* (pp. 158–165). Barcelona, Spain.

Almuhareb, A., & Poesio, M. (2005). Concept learning and categorization from the web. In *Proceedings of the 27th annual meeting of the cognitive science society* (pp. 103–108). Stresa, Italy.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*(3), 463–498.

Barbu, E. (2008). Combining methods to learn feature-norm-like concept descriptions. In *Proceedings of the ESSLLI 2008 workshop on distributional lexical semantics* (pp. 9–16). Hamburg, Germany.

Baroni, M., Evert, S., & Lenci, A. (Eds.). (2008). *Proceedings of the ESSLLI 2008 workshop on distributional lexical semantics*. Hamburg, Germany.

Baroni, M. and Lenci, A. (2008). Concepts and properties in word spaces. *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science (Special issue of the Italian Journal of Linguistics), 20*(1), 55–88.

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science, 34*(2), 222–254.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media Inc.

Briscoe, E. (2006). An introduction to tag sequence grammars and the RASP system parser. Technical Report 662, University of Cambridge, Computer Laboratory.

Briscoe, E. & Carroll, J. (2002). Robust accurate statistical annotation of general text. In *Proceedings of the 3rd international conference on language resources and evaluation* (pp. 1499–1504). Las Palmas, Gran Canaria.

Briscoe, E., Carroll, J., & Watson, R. (2006). The second release of the RASP system. In *Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 77–80). Sydney, Australia.

Budanitsky, A., & Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics, 32*(1), 13–47.

Burnard, L. (2000). *Users reference guide for the British National Corpus*. Technical report. Oxford: Oxford University Computing Services.

Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 32*(4), 643–658.

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science, 23*(3), 371–414.

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General, 120*(4), 339–357.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Gabrilovich, E. & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on artificial intelligence (IJCAI'07)* (pp. 1606–1611). Hyderabad, India.

Greer, M. J., van Castern, M., McLellan, S. A., Moss, H. E., Rodd, J., Rogers, T. T., & Tyler, L. K., (2001). The emergence of semantic catgeories from distributed featural representations. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 358–363). Edinburgh, Scotland.

Grondin, R., Lupker, S. J., & McRae, K. (2009). Shared features dominate semantic richness effects for concrete concepts. *Journal of Memory and Language, 60*(1), 1–19.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98*(1), 74–95.

Jiang, J. J., Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the international conference on research in computational linguistics (ROCLING X)*, (pp. 19–33). Tapei, Taiwan.

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Leacock, C., Miller, G. A., & Chodorow, M. (1998). Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics, 24*(1), 147–165.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on machine learning (ICML-98)* (pp. 296–304). Madison, Wisconsin.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers, 28*(2), 203–208.

Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(1), 3–23.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*(2), 159–197.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*(4), 547–559.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126*(2), 99–130.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*(5880), 1191–1195.

Moss, H. E., Tyler, L. K., Devlin, J. (2002). The emergence of category-specific deficits in a distributed semantic system. In E. M. E. Forde & G. W. Humphreys (Eds.), *Category specificity in brain and mind*. Hove, UK: Psychology Press.

Moss, H. E., Tyler, L. K., & Taylor, K. I. (2007). Conceptual structure. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 217–234). Oxford, UK: Oxford University Press.

Murphy, B., Baroni, M., & Poesio, M. (2009). EEG responds to conceptual stimuli and corpus semantics. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP 2009)* (pp. 619–627). East Stroudsburg, PA.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: The MIT Press.

Ó Séaghdha, D., & Copestake, A. (2008). Semantic classification with distributional kernels. In *Proceedings of the 22nd international conference on computational linguistics (COLING-08)* (pp. 649–656). Manchester, UK.

Pado, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics, 33*(2), 161–199.

Pexman, P., Holyk, G., & Monfils, M. (2003). Number-of-features effects and semantic processing. *Memory & Cognition, 31*, 842–855.

Pexman, P., Lupker, S., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review, 9*, 542–549.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes, 12*(5), 765–806.

Randall, B., Moss, H. E., Rodd, J. M., Greer, M., & Tyler, L. K. (2004). Distinctiveness and correlation in conceptual structure: Behavioral and computational studies. *Journal of Experimental Psychology: Learning, Memory & Cognition, 30*(2), 393–406.

Randall, B., Taylor, K. I., Devereux, B., Acres, K., & Tyler, L.K. (2009). The differential engagement of shared and distinctive correlated feature information in picture categorisation and naming: evidence for the Conceptual Structure Account. In *Proceedings of AMLaP* (p. 33). Barcelona, Spain.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on artificial intelligence (IJCAI-95)* (pp. 448–453). Montréal, Canada.

Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573–605.

Sun, L., Korhonen, A., and Krymolowski, Y. (2008). Verb class discovery from rich syntactic data. In *Proceedings of the ninth international conference on intelligent text processing and computational linguistics (CICLing-2008)* (pp. 16–27). Haifa, Israel.

Taylor, K. I., Salamoura, A., Randall, B., Moss, H., & Tyler, L. K. (2008). Clarifying the nature of the distinctiveness by domain interaction in conceptual structure: Comment on Cree, McNorgan, and McRae (2006). *Journal of Experimental Psychology: Learning, Memory & Cognition, 34*(3), 719–725.

Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences, 5*(6), 244–252.

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language, 75*(2), 195–231.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods, 40*(1), 183–190.

Vlachos, A., Korhonen, A., & Ghahramani, Z. (2009). Unsupervised and constrained dirichlet process mixture models for verb clustering. In *Proceedings of the workshop on geometrical models of natural language semantics* (pp. 74–82). Athens, Greece.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. In *Proceedings of the 32nd annual meeting of the association for computational linguistics* (pp. 133–138). Las Cruces, New Mexico.