ColloCaid: A Real-time Tool to Help Academic Writers with English Collocations

Robert Lew[1], Ana Frankenberg-Garcia[2], Geraint Paul Rees[2], Jonathan C. Roberts[3], Nirwan Sharma[3]

[1]Faculty of English, Adam Mickiewicz University in Poznań, [2]School of Literature and Languages, University of Surrey, [3]School of Computer Science, Bangor University

E-mail: rlew@amu.edu.pl, a.frankenberg-garcia@surrey.ac.uk, g.rees@surrey.ac.uk, j.c.roberts@bangor.ac.uk, n.sharma@bangor.ac.uk

Abstract

Writing is a cognitively challenging activity that can benefit from lexicographic support. Academic writing in English presents a particular challenge, given the extent of use of English for this purpose. The ColloCaid tool, currently under development, responds to this challenge. It is intended to assist academic English writers by providing collocation suggestions, as well as alerting writers to unconventional collocational choices as they write. The underlying collocational data are based on a carefully curated set of about 500 collocational bases (nouns, verbs, and adjectives) characteristic of academic English, and their collocates with illustrative examples. These data have been derived from state-of-the-art corpora of academic English and academic vocabulary lists. The manual curation by expert lexicographers and reliance on specifically Academic English textual resources are what distinguishes ColloCaid from existing collocational resources. A further characteristic of ColloCaid is its strong emphasis on usability. The tool draws on dictionary-user research, findings in information visualization, as well as usability testing specific to ColloCaid in order to find an optimal amount of collocation prompts, and the best way to present them to the user.

Keywords: writing assistant, collocation, academic writing, English for academic purposes

1 Background and Rationale

As lexicography moves forward into the digital age (Lew & De Schryver 2014), stand-alone dictionaries are gradually giving way to sophisticated and specialized lexicographic devices integrated in digital tools which may be optimized for specific tasks. One task that requires extensive lexicographic assistance is writing. The present contribution introduces the ColloCaid tool, which will be able to suggest collocational choices in real time during the process of writing, with a focus on academic English. ColloCaid recognizes that there are no native users of academic language (Frankenberg-Garcia 2017; Hyland 2006; Kosem 2010), and is therefore foreseen to be of value to both native and non-native writers who do not have sufficient command of academic English collocations.

Existing automated collocation-extraction tools tend to adopt a one-size-fits-all strategy. This is true of the domains they address; for example, in addition to other functions, Grammarly, Read & Write Gold, and Write Away provide collocation suggestions for general English. It is also true of the type of collocations they deal with; Wanner, Verlinde and Alonso Ramos (2013) argue that the assumption that all collocation errors can be corrected in the same way is mistaken. They claim instead that tools should focus on collocations comprising the parts of speech which pose writers most problems. Those few existing tools that do deal with specific domains and genres are undoubtedly useful for the writer, however they address a limited range of collocation errors. For example, although Cambridge's Write and Improve provides non-native writers with feedback on set writing tasks, as far as

collocations are concerned this feedback is limited to highlighting missing or incorrect prepositions. In Spanish ArText provides feedback on texts from the domains of Public Administration, Medicine and Tourism. Its feedback on collocations centers on the over or underuse of connectors such as *por lo tanto* (therefore) and *sin embargo* (however).

In contrast, the emphasis of the present project is on providing carefully curated content based on relevant and extensive resources focusing on general academic English. Starting from the generally accepted notion (Hausmann 2004; Martin 2008) of a collocate comprising a base (sometimes called a node) and collocate (sometimes called a collocator), up-to-date academic vocabulary lists are first referenced to identify the relevant sets of collocational bases, then a number of state-of-the-art corpora are explored to identify the salient collocates of these collocational bases.

2 Curated Collocational Data

2.1 Master Word List

For noun bases, we plan to include their typical pre-modifiers, verbs that take those noun bases as subjects and objects, as well as any characteristic prepositions. For verb bases, adverbial modifiers and prepositions would be added. Finally, adjective bases would be supplied with their salient pre-modifying adverbs. To supplement the 'positive evidence', the tool should be able to identify inappropriate collocational choices attested in learner corpora and other sources.

The rationale underlying the decision to concentrate on these types of bases and collocates is that writers are more likely to start with a noun in mind and then look up a verb collocate than start with a verb and then search for a noun collocate. For example (see Figure 1), a writer might wish to comment on a certain *measure*, provoking the questions (and potential collocates): 'What preposition should I use?' (*a measure of/for*), 'How do I characterize the measure?' (*a reliable/objective/quality measure*), 'How do I say that this measure was used?' (*we adopted/introduced/developed a measure*), 'What does the measure do?' (*a measure captures/indicates/represents something*). Conversely, it is unlikely that a writer would think of the verb *develop* then wonder 'What to develop?' (*a theory, a measure, a system*). Nonetheless, it is possible that he or she might wonder how to qualify the verb in an idiomatic way. For example, the idea that CO_2 *emissions contribute to global warming* might prompt the questions: 'To what degree?' (*significantly, substantially*), or when a model is found to *account for patterns in data*, one might wonder what adverb to use to qualify the degree of fit of the model (*fully/largely/partially account*). Similarly, adjectives might also provoke collocational doubts during the writing process, for example, when two groups or conditions turn out to be *different*, typically a question arises: 'How different?' (*substantially, significantly*).

Even with this restriction on the parts of speech to be considered for inclusion in the master word list, the number of potential bases would be impractically large to be wholly relevant to the user or to permit any thoroughgoing lexicographic treatment. To address this problem, the results of three widely recognized studies of EAP lexis were applied to draw out those node words which would likely be relevant to EAP writers. The first, the Academic Vocabulary List (AVL, Gardner & Davies 2014), comprises 3,000 core lemmas that occur across a range of academic disciplines in the 120-million-word academic sub-corpus of the Corpus of Contemporary American English (COCA, Davies 2008-). Some 2,700 of these AVL lemmas fell within the part-of-speech categories specified above. Durrant (2016) found that only 427 AVL items were found frequently in over 90% of disciplines in university student writing as represented by the BAWE corpus (Alsop & Nesi 2009). Of these 174 were nouns, 136 verbs and 79 adjectives. Applying this AVL-BAWE filter gave a workable number of potential node words. Further validation is provided by the Academic Keyword List (AKL, Paquot

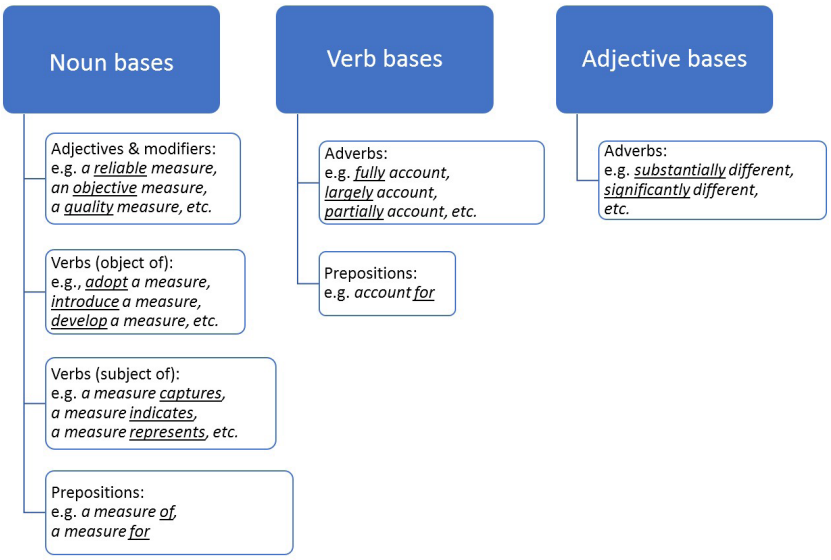

Figure 1: Types of collocational nodes included in ColloCaid, with examples.

2010). Cross-referencing the 353 nouns, 233 verbs and 180 adjectives contained in the AKL with the results of the AWL-BAWE filtered list, provided another means of drawing out potentially useful node words. A final means of filtering relied on the itemized list of 526 noun bases, 96 verb bases and 83 adjective bases of the Academic Collocation List (ACL, Ackermann & Chen 2013) found in the appendix of the *Longman Collocations Dictionary* (Mayor 2013). Table 1 shows the crossover among these three sources.

Table 1: EAP vocabulary considered in ColloCaid.

	AVL-BAWE lemmas	AKL lemmas	ACL lemmas	Total EAP lemmas considered	Lemmas attested in all three lists	Lemmas in at least two lists (ColloCaid)
Nouns	172	353	525	643	125	282
Verbs	129	233	95	283	38	136
Adjectives	86	180	83	231	24	94
Total	387	766	703	1157	187	513

Lemmas attested in at least two of the lists were considered as bases in the master list, with priority given to those 187 lemmas present in all three lists. Ultimately, the decision about the inclusion of the 513 node words in the final ColloCaid tool would depend on their collocational behavior. The following section sets out how this behavior was examined.

2.2 Collocates and Examples of Use

Collocational bases (see previous section) were looked up using the Sketch Engine (Kilgarriff et al. 2014; Kilgarriff et al. 2004).

As has been seen from the discussion of vocabulary lists above, corpora of student writing, namely BAWE and LOCNESS, were used to select collocation bases that novice writers were likely to use. However, corpora of professional academic writing representing 'expert performances' (Bazerman 1994: 131) are a more appropriate source of collocation information. Two such corpora in the Sketch Engine, made available with the kind permission of Pearson Longman and Oxford

University Press, were consulted: the Pearson International Corpus of Academic English (PICAE, Ackermann et al. 2011) and the Oxford Corpus of Academic English (OCAE). With around 70 million words of expert academic writing, the OCAE is more than double the size of PICAE, and was therefore given priority.

The Word Sketch tool was employed to list common collocates by syntactic set, arranged by their logDice scores (see Figure 2), currently the default measure of collocability in the Sketch Engine (Kilgarriff & Kosem 2012). By inspecting word sketches for a random sample of high and low-frequency bases from the different part-of-speech categories, a set of logDice and frequency thresholds corresponding with our intuitive judgements about collocations for EAP writers was found. The thresholds arrived at were a logDice score of ≥ 5 for all parts of speech with minimum co-occurrence frequency of 10 for lexical collocates and 100 for prepositions. This stage offers the opportunity to further curate the data. Collocates which are too general to be of relevance to the user e.g. *own* and *good,* in the modifier measure, are filtered out; as are base-collocate pairs which are obviously restricted to a small number of disciplines, for example, *entrepreneurial* found in the modifier relation for *ability.* The collocation *entrepreneurial ability* is likely not of interest to users working outside business studies and related disciplines, while it is highly likely that users working in these disciplines would have mastered this collocation.

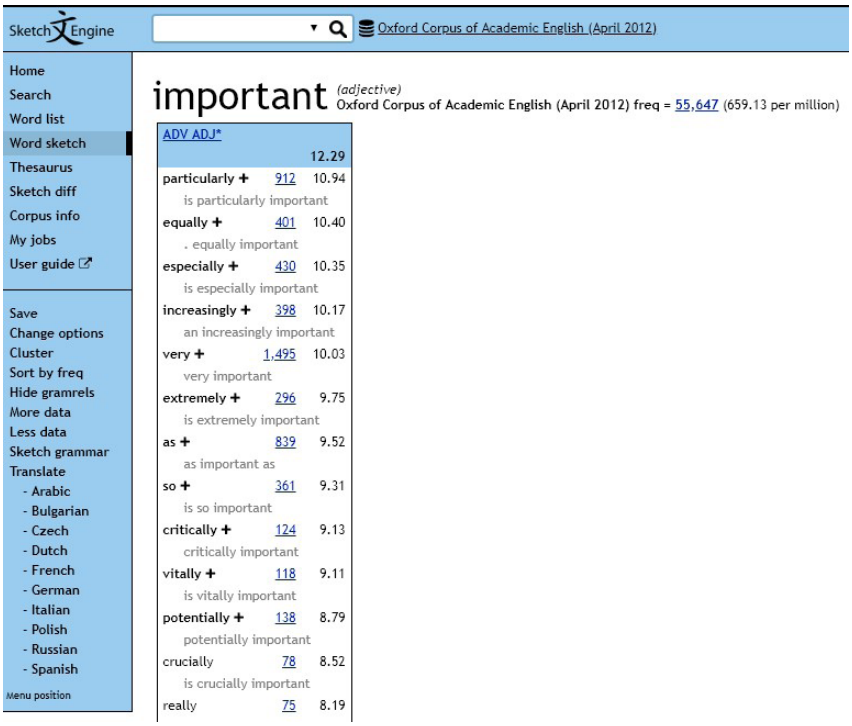

Figure 2: Query in the Oxford Corpus of Academic English for the collocational node *important* using Sketch Engine.

The collocates selected as above are systematically entered into a spreadsheet (see Figure 3), one collocate per row. The spreadsheet includes the base form along with its syntactic class (POS), type of collocational relation, the collocate, its raw frequency of co-occurrence with the base in the Oxford Corpus of Academic English, and the corresponding logDice score. Following the finding reported in Frankenberg-Garcia (2014; 2015) that one example alone may not be sufficient to aid language production, corpus citations are used to supply three examples per each collocate-base pair. In addition to the revision of base-collocate pairs in Word Sketch outlined above, the extraction of citations from

KWIC lines offers another opportunity to filter out those collocations which are predominantly used in a restricted set of disciplines. For example, from Word Sketch alone there was nothing about the collocation *unauthorised access* which suggested its usage is restricted to a particular field. However, while collecting citations from KWIC lines it became apparent that all instances of this collocation were related to computer science.

The examples included are based on corpus citations but are rarely verbatim excerpts. Elements not central to the core meaning expressed in the citation, primarily certain prepositional phrases and adjectives, are removed so as not to distract the user. To protect the identity of the source of the citations proper nouns are deleted or replaced with pronouns, e.g. *It is sometimes said that Watson and Crick discovered DNA* becomes *It is sometimes said that they discovered DNA*; numbers and dates are rounded, e.g. *1982* becomes *1980*; numerical references to figures and tables are changed, e.g. *Table 7* becomes *Table 1*; and in-text citations in author-date styles, e.g. *(Surname, 2018)*, are changed to a documentary-note style, e.g. *[1]*.

BASE	POS	RELATION	COLLOCATE	CO-	ASSOC	EXAMPLE1
equal	j	ADV ADJ*	roughly	108	11.42	the latter two groups had roughly equal rates of breal
equal	j	ADV ADJ*	exactly	62	10.85	total costs and our total revenues are exactly equal
equal	j	ADV ADJ*	nearly	57	9.89	three experiments were performed using nearly equa
equal	j	ADV ADJ*	almost	69	9.26	men and women are apparently almost equal now in t
equal	j	ADV ADJ*	formally	10	8.35	this occurs where treatment is formally equal
equal	j	ADV ADJ*	necessarily	14	7.81	attachment does not necessarily equal ownership
equal	j	ADV ADJ*	relatively	20	5.48	all household members have relatively equal access tc
important	j	ADV ADJ*	particularly	912	10.94	the sensitivity of a test is particularly important
important	j	ADV ADJ*	equally	401	10.4	the two stages are equally important and interlinked
important	j	ADV ADJ*	especially	430	10.35	especially important were the localization of brain anc
important	j	ADV ADJ*	increasingly	398	10.17	public relations is becoming an increasingly important
important	j	ADV ADJ*	very	1 495	10.03	it is very important for an economy to be stable
important	j	ADV ADJ*	extremely	296	9.75	they see work and its consequences as extremely imp
important	j	ADV ADJ*	critically	124	9.13	determine which points of critically important informa
important	j	ADV ADJ*	vitally	118	9.11	the link between the two concepts is therefore vitally

Figure 3: Excerpt from a collocations database underlying ColloCaid.

3 ColloCaid as a User-friendly Tool

Writing relies on cognitive processes such as user-attention, working memory and content retrieval from long-term memory in order to utilize different types of knowledges (domain, linguistic, pragmatic and procedural) for text production (Alamargot & Chanquoy 2001). In addition to this, features of the task environment such as the nature of audience, collaborators, already-composed text and the medium of writing add to the cognitive workload of the user. In the context of a digital learning environment this includes prompts and associated information offered through writing assistants, such as ColloCaid, which require further information processing resulting in new decision-making demands while performing the writing task. Therefore, from a learning perspective, this new information from a digital tool should be integrated and displayed to the user in a manner which does not disrupt the primary task of writing. Furthermore, after using this information, the user should be able to resume the writing task. Using a learning-centric approach we aim to prototype interactive tools which integrate lexicographic information into existing forms of text-editors, thus providing collocation information in context of the writing task. We then intend to evaluate these prototypes in order to understand how the new information is appraised by the users for improving their texts as well as developing writing expertise. The insights from these evaluations will help in further improving the design of ColloCaid and similar tools, and potentially offer opportunities to explore novel interaction and information-visualization techniques which may be appropriate for user-learning and improving

writing using text-editors (Roberts et al. 2017).

3.1 Example Scenario

Most free-to-use and commercial text editors offer a similar set of features (spellcheck, word-repetition, grammar check, etc.) to support users during the writing task. This provides us with a real-world context for using a task-centered design approach for integrating the collocation information into the existing user-workflow. We illustrate this approach using a task-based scenario which may inform the design of ColloCaid (Lewis & Rieman 1994). The tool would monitor the progress of the user as is standard in most text-editors which prompt the user when an error is encountered and/or a suggestion is recommended. In a similar manner, as soon as the user types one of the nodes, the tool would prompt the user that possible collocations may be available for that node (shown using dashed line under the node *research* in Figure 4). When the user interacts with this highlighted node, the tool would offer collocation suggestions, as in a simulated example in Figure 4, where the writer is given general patterns with the node *research* (in this case, the noun), a word which several studies of EAP lexis have highlighted as important; syntactic disambiguation needs to be dealt with in the occasional cases where there exist identically spelled nodes representing more than one syntactic category. In our case, a pop-up window would appear, indicating (here with pluses) that finer detail is available (this process is called drill-down).

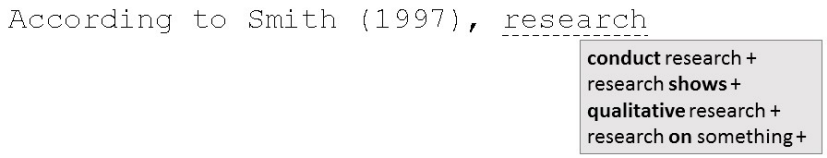

Figure 4: A pop-up general prompt triggered by the collocational node *research*.

To continue our example, let us assume the writer wants to report here on the research so far, therefore she clicks on the 'research **shows'** combination; to this, ColloCaid might respond by presenting a more detailed list of collocational choices, as in Figure 5, for example. It is important not to flood the user with too much information, a good general guide being considerations of working memory capacity (Miller 1956).

Figure 5: A narrower list of semantically related collocates are presented, following the writer's selection.

At this point, the choices available at the top of the prompt have been narrowed down to collocates that talk about research indicating something, possibly accompanied by related salient meanings (here towards the bottom). Again, plus symbols indicate that further information is available for each and every row. In this case, these would be the terminal nodes in the form of examples (Figure 6), which further

guide the user's writing. Here the user chose to use *suggest*, and examples are offered that illustrate this particular combination. In line with the recent finding (Frankenberg-Garcia 2015) that three examples are more helpful than a single one in supporting the writing process, three examples are given.

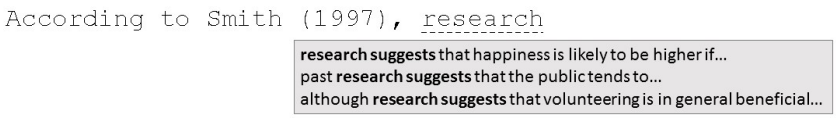

Figure 6: Examples given for the selected lexical combination.

4 Conclusion

The present project aims to develop an intuitive lexicographic resource integrated with digital writing environments to help academic English writers write more idiomatically in terms of their collocational choices. This paper has discussed the process of deciding which data the ColloCaid tool should cover, how this data is curated, and how it might be presented on screen in a way that is useful to the end-user. Thus far, the focus has been on 'positive evidence'. Lexicographically, this has involved reference to existing studies of academic lexis and corpora of expert academic writing, while from a visualization perspective it has focused on existing research on the on-screen visualization of text. The next step in the development process involves complementing this evidence. Lexicographically, this means adding information about those collocations which tend to present problems for EAP writers, and, from both a lexicographic and visualization perspective, conducting end-user studies to evaluate the tool. In completing the development process, it is anticipated that ColloCaid will provide useful contributions to the fields of human computer interaction, data visualization and lexicography. More importantly, it is hoped that the tool will make a positive practical difference to EAP writers of many proficiency levels, language backgrounds, and academic career stages, helping them to concentrate on the content of their writing and agonize less over the writing process.

References

Writing Tools

Grammarly. https://www.grammarly.com
Read & Write. https://www.texthelp.com/en-us/products/read-write/
WriteAway. http://writeaway.nlpweb.org/

Other References

Ackermann, K. & Chen, Y.-H. (2013). Developing the Academic Collocations List (ACL) – A Corpus-driven and Expert-judged Approach. *Journal of English for Academic Purposes* 12. 235-247.
Ackermann, K., de Jong, J., Kilgarriff, A. & Tugwell, D. (2011). The Pearson International Corpus of Academic English (PICAE).
Alamargot, D. & Chanquoy, L. (2001). *Through the models of writing*. *Studies in writing*. Dordrecht, Netherlands ; Boston: Kluwer Academic Publishers.
Alsop, S. & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora* 4 (1). 71-83.
Bazerman, C. (1994). *Constructing Experience*. Carbondale: Southern Illinois University Press.
Davies, M. (2008-). The Corpus of Contemporary American English.

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes* 43. 49-61.
Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL* 26 (2). 128–146.
Frankenberg-Garcia, A. (2015). Dictionaries and encoding examples to support language production. *International Journal of Lexicography* 28 (4). 490-512.
Frankenberg-Garcia, A. (2017). Assessing the productive collocation repertoire of writers for the development of dedicated writing assistant tools. *Electronic Lexicography in the 21st Century (eLex 2017)*. Leiden.
Gardner, D. & Davies, M. (2014). A new Academic Vocabulary List. *Applied Linguistics* 35 (3). 305-327.
Hausmann, F. J. (2004). Was sind eigentlich Kollokationen. In K. Steyer (ed.), *Wortverbindungen - mehr oder weniger fest*, Berlin: de Gruyter. 309-334.
Hyland, K. (2006). *English for Academic Purposes: An Advanced Resource Book*. London/New York: Routledge.
Kilgarriff, A. *et al.* (2014). The Sketch Engine: Ten years on. *Lexicography* 1. 7-36.
Kilgarriff, A. & Kosem, I. (2012). Corpus tools for lexicographers. In S. Granger, M. Paquot (eds.), *Electronic lexicography*, Oxford: Oxford University Press. 31–55.
Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004*, Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud. 105–116.
Kosem, I. (2010). Designing a model for a corpus-driven dictionary of Academic English. Ph.D., Aston University.
Lew, R. & De Schryver, G.-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography* 27 (4). 341–359.
Lewis, C. & Rieman, J. (1994). *Task-Centered User Interface Design: A Practical Introduction*.
Martin, W. (2008). A unified approach to semantic frames and collocational patterns. In S. Granger, F. Meunier (eds.), *Phraseology: An interdisciplinary perspective*, Amsterdam: John Benjamins. 51-66.
Mayor, M. (2013). *Longman Collocations Dictionary and Thesaurus*. Harlow: Pearson Education.
Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63 (2). 81-97.
Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Continuum.
Roberts, J. C., Frankenberg-Garcia, A., Lew, R., Rees, G. P. & Pereda, J. (2017). Visualisation and graphical techniques to help writers write more idiomatically. *IEEE Conference on Visualization (VIS)*. Pheonix, Arizona.
Wanner, L., Verlinde, S. & Alonso Ramos, M. (2013). Writing assistants and automatic lexical error correction: word combinatorics. *eLex 2013, 2013*, 427-487.

Acknowledgements

This research was supported by the Arts and Humanities Research Council [grant number AH/P003508/1].