Profiling idioms: a sociolexical approach to the study of phraseological patterns

Sara Moze and Emad Mohamed

Research Institute of Information and Language Processing (RIILP),
University of Wolverhampton, United Kingdom
{S.Moze, E.Mohamed2}@wlv.ac.uk

Abstract. This paper introduces a novel approach to the study of lexical and pragmatic meaning called 'sociolexical profiling', which aims at correlating the use of lexical items with author-attributed demographic features, such as gender, age, profession, and education. The approach was applied to a case study of a set of English idioms derived from the Pattern Dictionary of English Verbs (PDEV), a corpus-driven lexical resource which defines verb senses in terms of the phraseological patterns in which a verb typically occurs. For each selected idiom, a gender profile was generated based on data extracted from the Blog Authorship Corpus (BAC) in order to establish whether any statistically significant differences can be detected in the way men and women use idioms in every-day communication. A quantitative and qualitative analysis of the gender profiles was subsequently performed, enabling us to test the validity of the proposed approach. If performed on a large scale, we believe that sociolexical profiling will have important implications for several areas of research, including corpus lexicography, translation, creative writing, forensic linguistics, and natural language processing.

Keywords: Idiom, Phraseology, Sociolexical Profiling, Gender, Corpus Linguistics.

1 Introduction

The field of lexicography has undergone dramatic changes over the past three decades, mainly as a direct result of the latest technological developments in Computer Science and the emergence of Corpus Linguistics as the predominant methodology in the compilation of modern dictionaries and lexical resources. This has important implications for the study of language and meaning, as scholars are now able to study large numbers of observed uses of each word in order to discover how it is normally used. Furthermore, the Internet now offers plentiful evidence for word use and word meaning, with billions upon billions of words of text in machine-readable form being readily available to linguists and Natural Language Processing (NLP) specialists alike. The opportunities are seemingly limitless – and with a large portion of the general population being increasingly active on social media, the amount and type of linguistically relevant information that can be harvested from blogs, tweets, and other

publicly available online texts can help shed new light onto previously undiscovered patterns of linguistic behaviour.

Another multidisciplinary field that has greatly benefitted from Corpus Linguistic research is Forensic Linguistics, with corpus stylometry techniques being increasingly used to study the linguistic habits and patterns of both individuals and groups. Authorship profiling and attribution, in particular, are two such tasks where major breakthroughs have been made (cf. Grieve, 2007; Juola, 2008; Stamatatos, 2008; Argamon et al., 2009; Savoy, 2015). Computational methods are now being extensively used to determine which sets of textual features can be used to distinguish between the writing styles of different authors, enabling researchers to study their stylistic signatures in a similar way a detective would analyse a suspect's fingerprint, and to determine how common characteristics such as, for instance, age, gender, level of education, profession, and socioeconomic background are reflected in the choices we make as authors of texts (cf. Oakes, 2014).

In this paper, we propose a new approach to the study of lexical items, which we call ***sociolexical profiling.*** In a similar way to how forensic linguists approach the study of authorial fingerprints, we advance that researchers can profile lexical items (including phraseological units) by correlating their use with author-attributed demographic features. As far as we are aware, this approach has not been attempted anywhere before. In the present study, we decided to apply the technique to the study of English idioms in order to test the validity of our approach.

Existing English dictionaries provide a wealth of information on a word's semantics and, to some extent, phraseological units in which it participates, however, they very often fail to explain how the user is supposed to distinguish one sense from another, or provide them with sufficiently detailed information on a word's collocational preferences. To address these inadequacies, Hanks (2013) advocates for a new generation of 'context-sensitive' dictionaries, i.e. corpus-driven pattern dictionaries, which disambiguate between the different senses of a word based on the phraseological patterns in which it typically occurs. Whilst pattern dictionaries provide an innovative solution to the word sense disambiguation problem, enabling lexicographers to list verb senses in a systematic and empirically well-founded way, there are currently still some limitations as to the type of semantic information they provide. More specifically, a word's meaning typically also incorporates fine-grained semantic and pragmatic distinctions that cannot be captured from its local linguistic context. In dictionaries, usage information is typically limited to domain and register labels, with semantic prosody only sporadically being encoded (e.g. through the use of labels such as 'pejorative' and 'offensive'), and general monolingual dictionaries typically do not encode detailed demographic information, profiling lexical items or word senses in terms of the gender, age, profession, or level of education of their prototypical users. The reason for this is relatively simple – speaker profiles are not meaning-determining features, hence they are not covered in dictionaries. We propose that using sociolexical profiling to enrich existing dictionaries could benefit a range of users from language professionals (e.g. translators, writers), linguists, to NLP researchers, and if such information can be integrated in a user-friendly, visually appealing way (e.g. interactive pop-ups, graphs etc.), it could help attract a wider pool of non-expert dictionary users.

The paper is structured as follows: Section 2 provides a detailed description of the Pattern Dictionary of English Verbs (PDEV), the primary lexical resource used to compile the initial list of idiom candidates to be examined in the study. Section 3 is dedicated to the practical experiments undertaken in this paper: the corpus (Section 3.1) and methodology (Section 3.2) used in the case study are described, and the results of both the quantitative and qualitative analysis are presented and summarised (Section 3.3). Finally, the main findings of the study are discussed (Section 4), alongside potential applications and future directions for our research.

2 Pattern Dictionary of English Verbs

The Pattern Dictionary of English Verbs (PDEV) (Hanks, in progress) is an online lexical resource that aims to describe the full variety of phraseological patterns exhibited by English verbs. PDEV is a corpus-driven resource compiled using the methodological apparatus of Corpus Pattern Analysis (CPA) (Hanks 2004, 2013), which allows linguists to disambiguate between a word's senses by mapping meaning onto specific lexicogrammatical patterns exhibited by a verb in a given context. Underpinned by the Theory of Norms and Exploitations (TNE) (Hanks, 2013), CPA aims at identifying 'norms', i.e. semantically motivated syntagmatic patterns of normal usage, including literal and domain-specific uses, conventional metaphors, phrasal verbs, and idioms, and exploring the way these patterns are creatively exploited in language through detailed, labour-intensive lexical analysis of large corpus samples.

The core idea behind TNE is that whilst words are hopelessly ambiguous, lexicogrammatical patterns are unambiguous and can therefore serve as a powerful tool for word sense disambiguation. TNE focuses on real language use rather than preconceived speculations about language typical of introspection-driven theories of meaning, thus providing a window into the every-day phraseology, an area of study often overlooked by traditional linguists who, for a very long time, favoured atypical and marginal linguistic phenomena over prototypical patterns of language use. Due to this focus, CPA and TNE are particularly well-suited to both lexicographic projects and meaning-focused Natural Language Processing (NLP) tasks.

In CPA-based pattern dictionaries, the different senses of a verb are presented as combinations of specific syntactic structures, lexical collocates, and semantic types representing the typical nominal slot fillers for each syntactic argument. Consider the verb *bark* shown in Fig. 1. According to PDEV, this verb exhibits four patterns, which correspond to four separate verb senses. Pattern 1, which occurs in more than two thirds (67.68%) of the annotated sample, describes a situation where an animal, usually a dog, fox, seal, or baboon, emits one or more sharp cries. This is also the most cognitively salient sense of *bark*, meaning that this is the primary, core meaning associated with the verb. This meaning of the verb is extended in pattern 2, in which *bark* is coerced into being a reporting verb by the monotransitive construction, in which either a noun phrase (e.g. *The platoon commanders barked <u>their orders</u> to dismount*) or a quote (e.g. *`BY TOMORROW THEN!' he barked back, and slammed down the receiver*) occurs in the direct object slot. In this case, parts of the verb's core semantics relating to the utterance of sharp noises is preserved, whilst the aggressive

aspect of barking is foregrounded to indicate that the human's behaviour is perceived to be highly unpleasant (negative semantic prosody). The capitalised words displayed between double square brackets ([[Human]], [[Dog]], [[Speech_Act]]) are not lexical items, but 'semantic types', i.e. mnemonic labels that best describe the semantic features shared by the nouns that typically occur in a given argument slot. Pattern 4 is not relevant for our discussion, as it refers to an etymologically unrelated sense of the verb (homonymy). Pattern 3, on the other hand, features a well-known idiom, i.e. *bark up the wrong tree* (to pursue a misguided course of action or line of thought), which clearly originated from pattern 1. This type of patterns are the focus of the present paper and case study. Each observed sense of the verb *bark*, including the idiom *bark up the wrong tree*, can only be activated by this specific combination of obligatory syntactic arguments (subject, direct object, adverbial) and their corresponding semantic types or nominal slot fillers called lexical sets (e.g. *the wrong tree*).

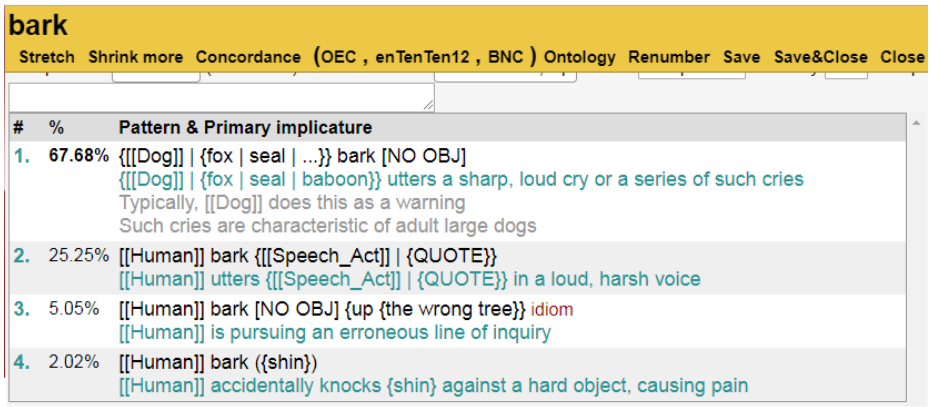

Fig. 1. The PDEV entry for the verb *bark*.

For each verb in PDEV, a random corpus sample consisting of at least 250 concordance lines is extracted from the British National Corpus (Leech, 1992) using Sketch Engine (Kilgarriff et al., 2014), and tagged with numbers corresponding to the pattern each concordance line exemplifies. When analysing high frequency and phraseologically complex verbs, e.g. light verbs such as *take*, *make*, or *blow*, the sample is normally augmented to 500, 1,000, or more lines so as to ensure that all phraseological patterns exhibited by the analysed verb are covered. Patterns are identified mainly through lexical analysis of corpus lines, complemented by information obtained through the 'word sketch' functionality in the Sketch Engine, which allows users to automatically generate a list of statistically relevant collocates and syntactic structures from a selected corpus. Patterns are then recorded and described in the CPA Editor (Baisa et al., 2015), our customized dictionary writing system, using CPA's shallow ontology of semantic types (Ježek and Hanks, 2010), which is shared across all CPA projects (for a detailed commentary on cross-linguistic adaptations of the ontology,

see Nazar and Renau, 2015). For each pattern, one or more implicatures are also added; the pattern's primary implicature (dark blue font) functions as its definition, whilst secondary implicatures (light grey font), which are optional, are added to semantically complex patterns that require further contextual information (situational, historical, or cultural) to be interpreted correctly (e.g. '*Typically, [[Dog]] does this as a warning'* in Pattern 1 – see Fig. 1). Domain (e.g. Medical, Law, Biology, Journalism) and register labels (e.g. Slang, Informal, Archaic) are added to each pattern individually, as are idiom and phrasal verb labels. Finally, each pattern is linked to the corresponding semantic frame in FrameNet (Ruppenhofer et al., 2006), with the aim of linking the two complementary lexical resources. PDEV entries also include quantitative information: for each separate pattern, a percentage is listed based on the pattern's frequency of occurrence in the manually annotated corpus sample.

Currently, PDEV contains over 1,700 completed dictionary entries, covering about one third of all lexical verbs in the English language. Similar pattern dictionaries are currently being compiled for Spanish (Renau and Nazar, in progress), Italian (Jezek et al., 2014), and Croatian, *inter alia*.

3 Case study: profiling PDEV idioms

3.1 Corpus and methodology

Corpus. The primary objective of this study was to investigate whether there are significant differences in the way men and women use idioms in every-day communication. For this purpose, we decided to use a subset of the Blog Authorship Corpus (BAC) (Schler et al, 2006; Argamon et al., 2009) comprising 681,288 blogs written by 19,230 bloggers, with a total of over 140 million words. BAC includes metadata about the bloggers' gender, age, occupation, and zodiac sign, which enabled us to generate gender usage statistics for our selected PDEV idioms. The subcorpus was selected so as to include an equal number of male and female bloggers, all aged between 13 and 47, as shown in Table 1 below.

Table 1. Age distribution in the BAC subcorpus.

Age group	Number of blogs
13-17	8,240
23-27	8,086
33-47	2,994

In order to test our core assumption that significant differences can be observed between speakers of different genders, we ran a preliminary experiment in which we trained a machine learning classifier to predict the gender of the blog writer. A precision and recall score of 0.72 was obtained in the experiment, indicating that the difference between the two genders is not random. In the experiment, we used the FastText classifier (Joulin et al., 2017) with a training set of 500,000 documents and a test set of 81,285 documents. The classifier used lemmatized bigrams as features and

was not optimized for prediction, as our only goal was to determine whether there are quantifiable linguistic differences between male and female bloggers.

Idiom selection. In order to maximize the number of extracted idioms, we decided to focus on PDEV verbs with the highest lexicogrammatical complexity, as we assumed that these are more likely to participate in idiomatic expressions. A frequency list of completed verbs in PDEV was generated based on the number of recorded patterns per verb, and the top 80 verbs were selected for the purpose of this study. The selected verbs are listed below:

abandon, absorb, act, admit, advance, align, answer, appear, ascend, ask, assess, assign, back, bang, battle, beat, beg, bite, blast, blow, boil, book, break, breathe, brush, build, burn, burst, call, clip, crash, cross, cry, die, dig, drain, eat, exchange, fail, filter, fire, fly, follow, grasp, grind, hack, hand, hang, hit, land, laugh, lead, live, lock, lose, mount, open, pack, pitch, plant, plough, point, pour, ride, rip, rush, scratch, see, settle, shed, shoot, slap, snap, soak, square, straighten, sweep, talk, tell, throw

The number of phraseological patterns associated with each verb on the list ranged from 83 (*break*) to 11 (*grasp*). As mentioned in Section 2, idioms are considered as separate patterns in PDEV and are explicitly labelled; this enabled us to automatically extract all idioms found in the 80 PDEV entries with relative ease. The list was then manually checked and validated; for the purpose of this study, we decided to focus on idioms that are relatively fixed in terms of their lexical and syntactic behaviour so as to facilitate computational processing. As a result, idioms exhibiting a high number of lexical alternations or word order configurations (e.g. non-projective structures, particles) were removed from the list. The procedure yielded a final list of 106 idiom candidates to be extracted from the BAC Corpus.

Idiom extraction. A cascaded set of regular expressions was used to process the corpus data and extract the idioms from BAC. The procedure can be broken down into the following steps:

1. Data pre-processing: this step involved cleaning the corpus by separating punctuation from words, normalizing spaces, and handling encoding issues in order to ensure that the corpus is encoded in UTF-8.
2. Morphological processing: for each selected idiom, a list of all potential word form combinations was generated. For example, the idiom *blow one's head* is conjugated in the following forms:
   a. *blow one's head, blew one's head, blown one's head, blowing one's head;*
   b. *blow his head, blow her head, blow their heads, blow my head, blow our heads, blow your head, blow your heads;*
   c. *blew his head, blew her head, blew their heads, blew my head, blew our heads, blew your head, blew your heads;*
   d. *blown his head, blown her head, blown their heads, blown my head, blown our heads, blown your head, blown your heads;*

e. *blowing his head, blowing her head, blowing their heads, blowing my head, blowing our heads, blowing your head, blowing your heads.*

The generation of these forms required writing rules and lists of word form alternations.

3. Statistical analysis: the generated list of word form combinations was used to extract all instances of the selected idioms in the BAC subcorpus, recording the number of occurrences in texts written by male and female authors, and these were then ranked to check their relative frequency per gender.

Manual validation. The extracted information was manually examined to ensure that the results were valid and could be meaningfully used in the analysis. Out of the initial list of idiom candidates, 101 were found in the corpus; one (*lose it)* was removed from the list due to ambiguity issues (most extracted sentences were instances of literal and not idiomatic use). Idiom candidates with a frequency lower than 5 in the subcorpus were also removed from the list in order to ensure that the results of the analysis were statistically significant. The process resulted in a finalized list of 85 idioms with 37 different verb bases, i.e. *abandon, act, answer, battle, beat, beg, bite, blow, break, breathe, burn, burst, call, cry, die, dig, eat, fly, follow, grasp, grind, hang, hit, laugh, live, lose, open, pack, point, pour, scratch, shoot, snap, soak, sweep,* and *throw.*

Data analysis. A qualitative analysis of the extracted idioms and the quantitative data was carried out in order to identify differences and similarities in the use of idioms between the two genders. Idioms were manually clustered into semantic classes based on semantic prosody, source and target domain (e.g. weapon and communication idioms), and other fine-grained semantic components.

3.2 Results

Two major conclusions can be drawn from the results obtained in the experiments: 1) some idioms appear to be predominantly used by one gender over the other, and 2) these tendencies do not necessarily correlate with the use of the verb lemma alone. Fig. 2 shows the gender ratio of all 85 idioms in the BAC subcorpus, sorted from left to right in descending order by male percentage, with the symmetrical curve in the middle demonstrating a healthy distribution between the two genders. Whilst a significant portion of idioms appearing in the middle of the graph appears to be equally associated with both male and female speakers, idioms appearing at the two extremities, which tend to be used predominantly by male (left) and female (right) speakers, constitute just as significant a portion of the examined idiomatic expressions.

Tables 2 and 3 list the top 20 idioms predominantly used by men and women respectively. Although some of the relative frequencies listed in the tables are still rela-

tively low,[1] the results do seem to point out to significant differences in the use of idioms between the two genders, with most idioms listed in the two tables being used by one of the genders in over two thirds of the corpus examples. A more detailed look into the two lists helped uncover further differences and similarities, enabling us to cluster idioms into semantically motivated groups and draw conclusions based on their gender profiles, as shown in the following subsections.

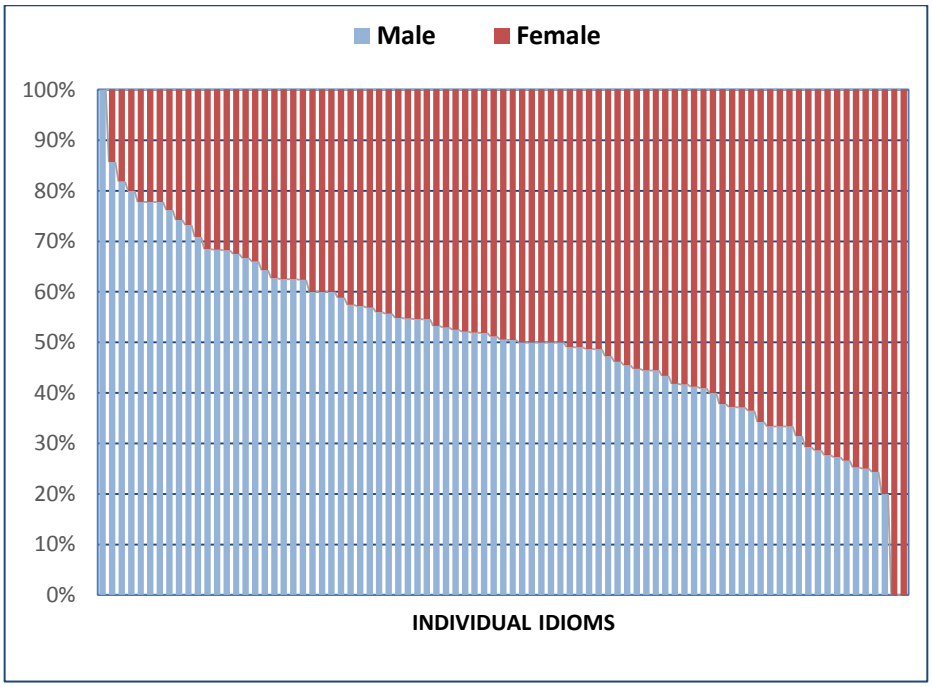

Fig. 2. Gender distribution of idioms in the subcorpus.

Table 2. The top 20 idioms predominantly used by male bloggers.

	Idiom	Frequency			Percentage	
		Male	Female	Total	Male	Female
1	*hit the headlines*	6	0	6	**100.00%**	0.00%
2	*throw the baby out with the bathwater*	6	1	7	**85.71%**	14.29%
3	*call the tune*	9	2	11	**81.82%**	18.18%
4	*answer the call of nature*	4	1	5	**80.00%**	20.00%
5	*grind to a halt*	28	8	36	**77.78%**	22.22%
6	*pour cold water on*	7	2	9	**77.78%**	22.22%
7	*pack a punch*	7	2	9	**77.78%**	22.22%
8	*bite the hand that feeds*	16	5	21	**76.19%**	23.81%

[1] This is not surprising; idioms are known to generally occur with very low frequencies in most corpora.

9	*hang in the balance*	23	8	31	**74.19%**	25.81%
10	*point the finger at sb*	30	11	41	**73.17%**	26.83%
11	*abandon ship*	17	7	24	**70.83%**	29.17%
12	*call a spade a spade*	13	6	19	**68.42%**	31.58%
13	*lose ground*	28	13	41	**68.29%**	31.71%
14	*call the shots*	30	14	44	**68.18%**	31.82%
15	*shoot oneself in the foot*	27	13	40	**67.50%**	32.50%
16	*breathe new life*	8	4	12	**66.67%**	33.33%
17	*follow suit*	128	66	194	**65.98%**	34.02%
18	*open the floodgates*	9	5	14	**64.29%**	35.71%
19	*hard to beat*	42	25	67	**62.69%**	37.31%
20	*plant the seed*	15	9	24	**62.50%**	37.50%

Table 3. The top 20 idioms predominantly used by female bloggers.

		Frequency			Percentage	
	Idiom	Male	Female	Total	Male	Female
1	*lose one's heart to*	0	7	7	0.00%	**100.00%**
2	*throw a wobbly*	0	5	5	0.00%	**100.00%**
3	*blow hot and cold*	2	8	10	20.00%	**80.00%**
4	*throw a fit*	28	87	115	24.35%	**75.65%**
5	*fly off the handle*	9	27	36	25.00%	**75.00%**
6	*burst into laughter*	55	163	218	25.23%	**74.77%**
7	*cry one's heart out*	47	130	177	26.55%	**73.45%**
8	*throw a tantrum*	12	32	44	27.27%	**72.73%**
9	*throw caution to the wind*	13	34	47	27.66%	**72.34%**
10	*breathe a word*	2	5	7	28.57%	**71.43%**
11	*act one's age*	19	46	65	29.23%	**70.77%**
12	*laugh one's head off*	39	85	124	31.45%	**68.55%**
13	*breathe a sigh of relief*	21	42	63	33.33%	**66.67%**
14	*not know whether to laugh or cry*	4	8	12	33.33%	**66.67%**
15	*live a double life*	3	6	9	33.33%	**66.67%**
16	*bite one's tongue*	50	96	146	34.25%	**65.75%**
17	*throw in the towel*	31	54	85	36.47%	**63.53%**
18	*pour one's heart out*	23	39	62	37.10%	**62.90%**
19	*soak up the sun*	16	27	43	37.21%	**62.79%**
20	*bite one's head off*	17	28	45	37.78%	**62.22%**

Verb base. The idioms were first grouped according to their verb base, and basic statistics were generated in order to explore whether any meaningful distinctions can be made in the use of idioms by male and female bloggers. The results seem to indicate that men tend to prefer verb bases that express physical actions (e.g. *pack, grind, abandon, call, follow, answer, open, plant, grasp, pour, beg*), some of which are particularly forceful or violent (e.g. *shoot, hit, battle, beat, break).* Conversely, women typically use idioms whose verb base denotes basic life actions and bodily functions, e.g. *breathe, laugh, cry, live,* and *act*. Only two verb bases associated with women

speakers encode some sort of force dynamics, i.e. *throw* and *blow*. However, *throw* typically appears in idiomatic light verb constructions (e.g. *throw a tantrum/fit/wobbly*) and is therefore delexicalised, hence the forceful aspect of its primary meaning is mostly not preserved.

In order to further explore the relationship between the analysed idioms and their verb bases, we extracted the total frequencies of occurrence of the 39 verb lemmas in the BAC subcorpus and compared them against their corresponding idioms. The new data showed that there are no statistically significant differences in the use of verb lemmas between the two genders, which clearly contrasts with their use in idiomatic expressions. Fig. 3 compares the gender profiles of idioms predominantly associated with male bloggers with their corresponding verb lemmas. The verbs *abandon*, *point*, *grant*, and *pack* are used more or less equally by men and women; however, the data shows a strong male bias in the use of their corresponding idioms *abandon ship*, *point a finger at somebody*, *grind to a halt*, and *pack a punch*. Furthermore, our analysis of *call* idioms uncovered significant differences in their gender profiles, with the percentage associated with men ranging from slightly over the 50% mark to as high as 81.82%. All of this seems to indicate that sociolexical profiling is indeed pattern-specific, which means that the different senses of a given word (physical, metaphorical, and domain-specific), as well as idioms, phrasal verbs, or other phraseological units, would theoretically require separate profiles.

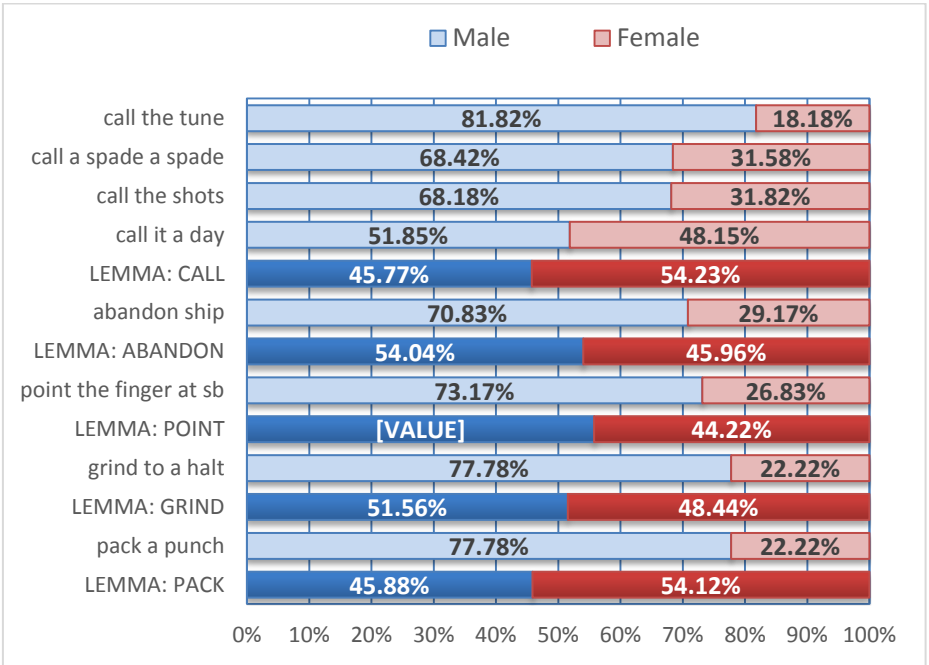

Fig. 3. Gender analysis of predominantly 'male' idioms and their corresponding verb lemmas.

Emotion idioms. Significant differences have been detected in the use of idioms that express intense emotional states and attitudes. In the analysed sample, 15 such idioms were identified, and based on overall and individual frequency data, we can conclude that this type of idioms are used significantly more often by women. More specifically, over two thirds of all occurrences (69.41% or 776 out of 1118) were attributed to female bloggers, with all of the individual idioms on the list showing a strong bias towards female usage, as shown in Fig. 4**.**

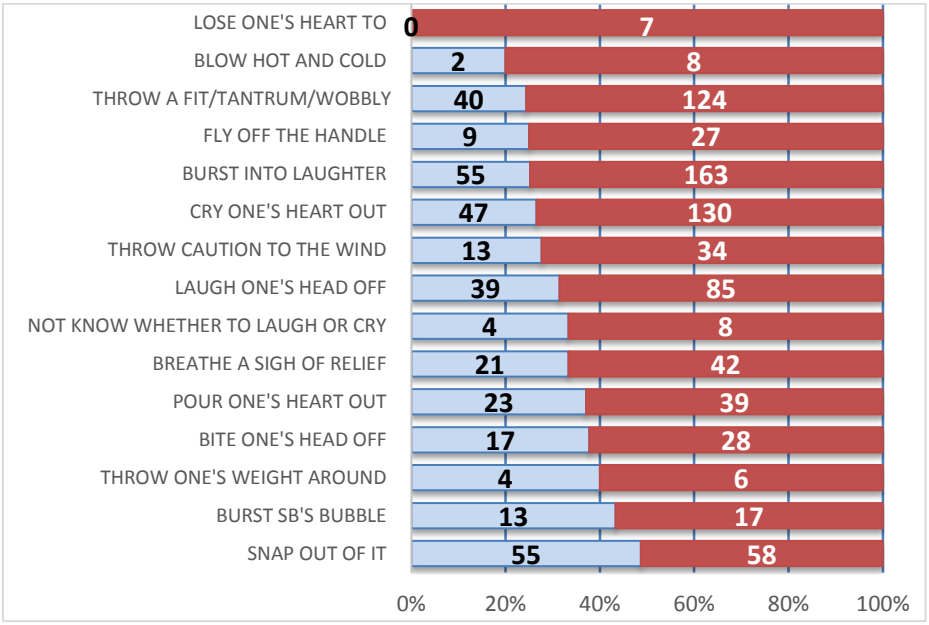

Fig. 4. Gender analysis of idioms incorporating a strong emotional component (blue – male; red – female).

Control and aggression. Idioms expressing aggression were found to be evenly distributed, with a minor bias towards male usage (54.96%). Nonetheless, a minor difference can be observed in the type of aggression that is expressed by the idioms in question: the only idiom expressing psychological aggression (bullying), i.e. *throw one's weight around,* was used slightly more frequently by women, although the frequencies extracted from the corpus are too low to make this observation conclusive. Conversely, male bloggers seem to have a slight preference for idioms that have a stronger physical component, with *pack a punch* and *battle it out* in particular being associated by male speakers (77.78% and 57.41% respectively), as shown in Fig. 5**.**

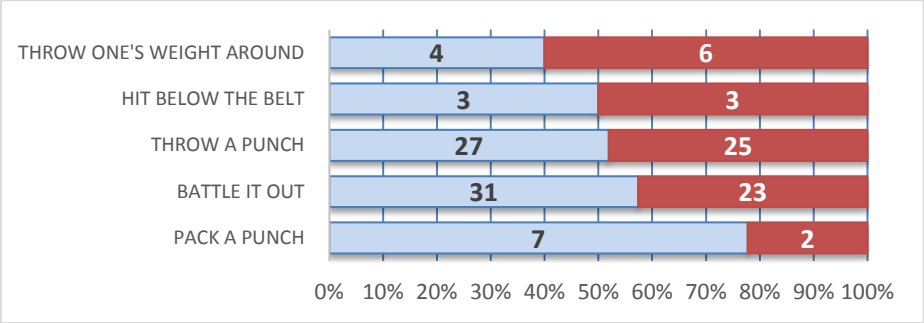

Fig. 5. Gender distribution of idioms expressing aggression (blue – male, red – female).

Control was another semantic component shared by a significant portion of the analysed idioms. Whilst the overall distribution of idioms expressing lack of control (e.g. *throw in the towel, hang by a thread, lose ground, bite off more than one can chew, abandon ship, eat humble pie*) was surprisingly even, this was not the case with idioms expressing highly assertive behaviour. More specifically, both *call the shots* and *call the tune* were found to be predominantly used by men, with 68.18% and 81.82% respectively. Both idioms have positive semantic prosody, indicating that speakers perceive the described act of taking control over a situation as proactiveness rather aggressive behaviour. This contrasts with the above mentioned *throw one's weight around*, which has decisively negative semantic prosody and is slightly more biased towards female usage.

Weapons, Military. Idioms can also be grouped into clusters based on their shared source domain(s). For instance, six idioms on the list were found to be associated with weapons and the military, hence a comparison was made to establish whether or not there might be any gender-specific tendencies in their usage. The results were somewhat inconclusive; overall, this group of idioms proved to be well-distributed between the two genders, with a minor bias towards male usage (55.13%), however, it appears that the picture is somewhat skewed by the two most frequently occurring idioms in the group, i.e. *hit the ground* and *bite the bullet*, which exhibited equal distribution between the two genders. It is worth noting that none of the six idioms in this group exhibited a positive bias towards female usage, whilst three other idioms, i.e. *abandon ship, call the shots,* and *shoot oneself in the foot,* were all found to be significantly more often used by male bloggers, with 70.83%, 68.18%, and 67.5% respectively, as shown in Fig. 6**.**

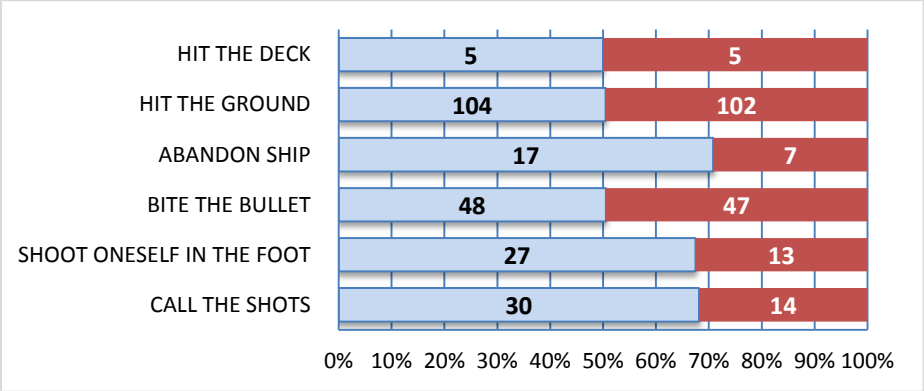

Fig. 6. Gender distribution of idioms originating from the Military/Weaponry domain (blue – male, red – female).

Communication. Several idioms loosely related to communication were identified and analysed in the study. Based on the results, these idioms constitute a relatively homogenous group in terms of their gender distribution (male bloggers: 47.45%, female bloggers: 52.55%). Nonetheless, it is worth noting that the only two idioms there were found to be predominantly used by women, i.e. *breathe a word* and *bite one's tongue*, are both associated with not speaking up or keeping silent. No such idioms were found among those that are evenly distributed or predominantly used by males.

An antonymous pair of idioms was also identified in the analysis, i.e. *beat around/about the bush – call a spade a spade*, which are used to describe one's style of communication (i.e. evasive versus direct). Whilst *beat around/about the bush* was found to be used equally by both genders, *call a spade a spade* was found to be used more frequently by men (see Fig. 7).

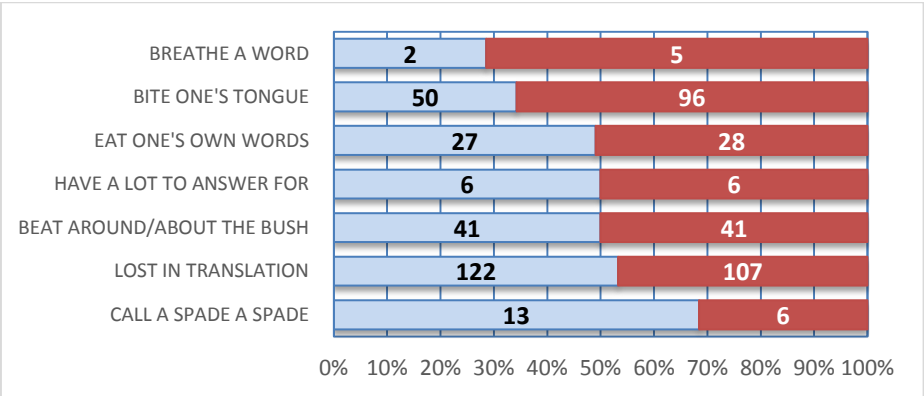

Fig. 7. Gender distribution of communication idioms (blue – male, red – female).

4 Conclusion

In this paper, we presented a novel approach to the study of lexical and pragmatic meaning called sociolexical profiling. The results of our case study on idiom usage clearly demonstrated that there are significant differences in the way speakers of different genders use idioms, and that verb lemmas and phraseological units in which they are found do not necessarily share the same gender profile. This led to the conclusion that sociolexical profiling is pattern-specific. In the future, we intend to extend our approach to all other types of patterned language, generating a wide range of demographic profiles (gender, age, profession, *inter alia*) for all patterns in PDEV, including those corresponding to core senses, conventional metaphors, and other types of phraseological units such as phrasal verbs and proverbs.

In order to achieve this long-term goal, we plan to extend our current research in the following directions: i) increasing the size of the corpus: large quantities of data are likely to reveal new undiscovered behavioural patterns and deepen our understanding of the correlation between language use and demographic features. We have already started collecting data using distant annotation techniques, which encode demographic information as metadata that can be exploited as annotation despite not being initially intended for this purpose; ii) using more advanced statistical tests: in this study, only lemma and idiom frequencies were used, and they proved to work surprisingly well as the data was evenly distributed. After increasing the size of our corpus, however, we intend to use more sophisticated statistical tests and methods to explore the data, i.e. mainly logistic regression and its associated odds ratio measure; iii) using machine learning for the prediction of social attributes: we will use interpretable machine learning algorithms (e.g. tree-based algorithms) and not-so-interpretable machine learning (e.g. neural network methods) to classify documents as belonging to one demographic class or another based on their linguistic features.

Several potential areas of application have been envisioned for our research; in addition to electronic lexicography (PDEV), we have identified creative writing, translation, and Natural Language Processing (NLP) as the main areas that could benefit from research in sociolexical profiling. For instance, when writing or translating dialogue between characters and trying to decide between near-synonyms, it might help the writer or translator to know which expressions (e.g. verb senses, idioms, or phrasal verbs) are typically used by specific groups of people. Knowing, for example, that the adjective *pretty* is more likely to be used by women and *beautiful* by men could help writers and translators make dialogues sound more authentic and believable. The potential impact of this research for various NLP tasks in language understanding and generation is just as significant. If performed on a large scale, computational sociolexical profiling could lead to the creation of a new resource akin to SentiWordNet (Esuli and Sebastiani, 2006). Whilst SentiWordNet provides users with information on the semantic prosody (also known as 'polarity') of word senses, defining them in terms of 'positive', 'negative', and 'neutral' polarity, a similar lexical resource (e.g. 'SocioWordNet') that profiles lexicogrammatical patterns in terms of various demographic features could be compiled using sociolexical profiling as working methodology. Such a resource could be used in various NLP tasks. For instance, sociolexical pro-

files could be used as an additional feature in authorship attribution and profiling to improve the accuracy of existing systems, which currently still rely on relatively simple features such as word and character n-grams.

References

1. Argamon, S. Koppel, M., Schler J., Pennebaker, J.: Automatically profiling the author of an anonymous text. Communications of the ACM 52 (2), 119–123 (2009).
2. Baisa, V., El Maarouf, I., Rychlý, P., Rambousek, A.: Software and Data for Corpus Pattern Analysis. In: Horák, A. et al. (eds.) RASLAN. pp. 75–86 Tribun EU (2015).
3. Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, citeseer.ist.psu.edu/esuli06sentiwordnet.html, (2006).
4. Grieve, J.: Quantitative authorship attribution: an evaluation of techniques. Literary and Linguistic Computing 22(3), 251–270 (2007).
5. Hanks, P.: Corpus Pattern Analysis. In: Williams, G., Vessier, S. (eds.), 11th Euralex International Congress. Proceedings, pp. 87-97. Lorient: Université de Bretagne-Sud (2004).
6. Hanks, P.: Lexical Analysis: Norms and Exploitations. Cambridge, MA, MIT Press (2013).
7. Ježek, E., Hanks, P.: What lexical sets tell us about conceptual categories. Lexis: E-journal in English lexicology. 4: Corpus Linguistics and the Lexicon, 7-22 (2010).
8. Ježek, E., Magnini, B., Feltracco, A., Bianchini, A., Popescu, O.: T-PAS: A resource of corpus-derived Typed Predicate Argument Structures for linguistic analysis and semantic processing. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 890–895. ELRA (2014).
9. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431 (2017).
10. Juola, P.: Author attribution. Foundations and Trends in Information Retrieval 1(3), 233–334 (2008).
11. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography 1(1), 7-36 (2014).
12. Nazar, R., Renau, I.: Ontology Population using Corpus Statistics. In: Papini, O. et al. (eds.) Proceedings of the Joint Ontology Workshops 2015 co-located with the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015) (2015).
13. Leech, G.: 100 million words of English: the British National Corpus (BNC). Language Research 28(1), 1–13 (1992).
14. Oakes, M.: Literary Detective Work on the Computer. John Benjamins, Amsterdam/Philadelphia (2014).
15. Renau, I., and Nazar, R.: Verbario, http://www.verbario.com, last accessed 2019/05/14.
16. Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., Scheffczyk, J.: FrameNet II: Extended Theory and Practice. Berkeley, CA, ICSI (2006).
17. Savoy J.: Comparative evaluation of term selection functions for authorship attribution. Literary and Linguistic Computing 30(2), 246–261 (2015).
18. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (2008).
19. Schler J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 199 –205. AAAI (2006).