Some Theoretical Considerations in Off-the-Shelf Text Analysis Software

Emma Franklin
University of Wolverhampton
emma.franklin@wlv.ac.uk

Abstract

This paper is concerned with theoretical considerations of commercial content analysis software, namely Linguistic Inquiry and Word Count (LIWC), developed by social psychologists at the University of Texas. LIWC is widely cited and forms the basis of many research papers from a range of disciplines. Here, LIWC is taken as an example of a context-independent, word-counting approach to text analysis, and the strengths and potential pitfalls of such a methodology are discussed. It is shown that text analysis software is constrained not only by its functions, but also by its underlying theoretical assumptions. The paper offers recommendations for good practice in software commercialisation and application, stressing the importance of transparency and acknowledgement of biases.

1 Introduction

Due to the ever-increasing availability of digital texts, automated text analysis methods are gaining popularity, and not only among linguists. Commercial text analysis programs can offer fast, inexpensive content analysis and are generally quite easy to use. To a social scientist faced with large amounts of discourse to analyse, such a product is almost irresistible. However, there are several caveats of which end-users should be made aware, and to which software developers could give more explicit attention.

There are a number of user-friendly and easily obtainable text analysis programs currently on the market. One which has been described as the "most widely used program for analysing text in clinical psychology" (Alpers et al., 2005: 363), and which continues to grow in popularity across a number of fields, is Linguistic Inquiry and Word Count[1] (LIWC, pronounced 'Luke'). This paper takes LIWC as an example of commercially available, easy-to-use text analysis software. It aims to illustrate the ways in which text analysis software is tightly bound to its theoretical basis, and the practical implications this can have on usage and performance. We find this a timely discussion due to the diverse range of researchers now turning to linguistic analysis software without necessarily understanding its construction.

The paper is organised as follows. Section 2 outlines the LIWC system and its development, followed by some of its perceived theoretical assumptions in Section 3. Section 4 describes some previous experiments using LIWC in Franklin (2015). Section 5 revisits LIWC's theoretical assumptions, and offers suggestions for good practice in software commercialisation and application. Section 6 concludes the paper.

2 LIWC

First developed by social psychologists at the University of Texas in the 1990s, and now in its second version, LIWC2007 is described as "a transparent text analysis program that counts words in psychologically meaningful categories" (Tausczik and Pennebaker, 2010: 24). It was originally employed in social psychology for investigating the connections between word use and mental health recovery (Pennebaker, 1997), later followed by relationship satisfaction (Agnew et al., 1998), university grades (Pennebaker and King, 1999), and testosterone levels (Pennebaker et al., 2004), amongst others. It has also been used to track collective responses to upheaval, such as the 9/11 attacks (Cohn et al., 2004).

The creators of LIWC have since used the program to analyse the essays, speeches, blog posts and Tweets of thousands of individuals across a variety of projects (Pennebaker, 2011). This has led to many others carrying out LIWC-based research in a range of applications: personality profiling (Mairesse and Walker, 2006), deception detection (Pérez-Rosas et al., 2014), sentiment analysis (Paltoglou et al., 2010), content analysis (Tumasjan et al., 2010) and review spam detection (Ott et al., 2013), to name a few.

[1] http://www.liwc.net

Proceedings of the Student Research Workshop associated with RANLP 2015, pages 8–15,
Hissar, Bulgaria, 7-9 September 2015.

2.1 Functions, Development and Validation

LIWC performs a simple word-count analysis by reading text files and matching each word against its inbuilt categories, or dictionaries. These categories – of which there are 68 – each constitute a 'dimension' of words, e.g. positive emotion words, prepositions, motion words, and so on.[2]

Some dimensions attempt to describe themes, or content (e.g. 'work'), while others count grammatical features (e.g. 'verbs'). For each text file, LIWC generates a score for each dimension, which reflects the percentage of the total words which match that category. So, a score of 3.9 for 'past' would indicate that 3.9% of that text consists of words which can be found in the LIWC 'past' dictionary. This allows the user to easily track changes in LIWC scores over time. A word can belong to more than one category; 'died', for example, appears in 'past', 'verbs' and 'death'.

The dictionaries are "the heart of the LIWC program" (Tausczik and Pennebaker, 2010: 27), and were finalised over the course of several years. The dictionaries were populated with the help of existing resources (e.g. Roget's Thesaurus) and "brain-storming sessions among 3-6 judges" (Pennebaker et al., 2007: 7). The word lists were then rated by three judges, who voted on whether each word should be used in that category, and whether new words should be added; two out of the three judges needed to be in agreement for a decision to be passed. The refined dictionaries were then rated once more by three different judges, with the same criteria for selection and deletion. Inter-judge agreement "ranged from 93% agreement for Insight to 100% agreement for Ingestion, Death, Religion, Friends, Relatives, and Humans" (Pennebaker et al., 2007: 7). LIWC was appraised again in 1997 and 2007 (Pennebaker et al., 2007).

External validation of the dictionaries was carried out in Pennebaker and Francis (1996) for a select number of categories, by asking four judges to rate student essays against LIWC-compatible dimensions. The judges' ratings were reasonably correlated with the corresponding LIWC categories, although different correlation scores are reported in Pennebaker et al. (2007), presumably due to the revision of the LIWC dictionaries. The (Pearson) correlation coefficient for these few categories vary from 0.07 for Sadness to 0.87 for Family. The authors conclude that these results give "support for LIWC's external validity" (Pennebaker et al. 2007: 9).

LIWC has been generally well received, and its lexicon has been described as "the standard for social psychological analysis of lexical features" (Jurafsky et al. 2009). Some scholars explicitly note the effectiveness of LIWC's simplistic approach (Mehl, 2006), while others find that LIWC performs well, but only in certain categories and for certain domains (Loughran and McDonald, 2011; Pérez-Rosas et al., 2014; Grimmer and Stewart, 2013). In some studies, only the dictionaries from LIWC are adopted (Skowron and Paltoglou 2011, Bae and Lee, 2012), and in others the authors create their own dictionaries to be used with the LIWC processor (Loughran and McDonald, 2011; Osherenko and André, 2007). To a social scientist looking at language, LIWC might seem an obvious choice of analysis tool.

3 Theoretical Assumptions

Text analysis programs are inherently philosophical. That is to say, all software is underpinned by theories and assumptions, and in the case of text analysis software, these are theories of language and assumptions on what constitutes meaning. This may seem obvious, but the implication is that the end user of a program also, perhaps unknowingly, applies these theories to their research, unless they are able to fully understand the way the program works and account for these effects later. Therefore, it is crucial that we identify and critically assess the theoretical underpinnings of a text analysis program before using it.

Below are some of LIWC's main features and its perceived theoretical assumptions (numbered in brackets).

- **LIWC counts words.** Based on the context in which LIWC was created, the underlying assumption (A1) is that the frequency of a word can tell us something about a person or about the content or tone of a text. A secondary assumption (A2) is that a computer program is ideal for carrying out this task.
- **It only considers single words.** In doing so, LIWC assumes that words have meaning in isolation (A3). There is also an implicit assumption (A4) that inaccuracies due to negation, word order, particles (e.g. in phrasal verbs), ambiguity of word senses, type of discourse and other context-dependent factors are negligible or unimportant.

[2] The full list of LIWC categories is available at http://www.liwc.net/descriptiontable1.php

- **It matches words against dictionaries.** The assumption here (A5) is that by creating dictionaries *a priori*, and by finding and counting only the words which match them, the interesting or relevant parts of a text will be identified. A secondary assumption (A6) is that a score corresponding to a dictionary label will also correspond to the intended semantic value of that dictionary. Thirdly, it is assumed that dictionary values are more useful than word frequencies (A7).

Of course, there are many text analysis approaches which share several of these assumptions, and LIWC is not an unusual case, although it is particularly insensitive to context.

Section 4 describes some experiments carried out by Franklin (2015) (based loosely on Cohn et al.'s (2004) LIWC analysis of American blogs written before and after the 9/11 attacks). Results from the study, which examined the transition experienced by new university students and the different outcomes of LIWC and keywords analyses, are selected so as to address the assumptions listed above (A1-7) as succinctly as possible.

4 Word-Count Software in Practice

Franklin (2015) sought to better understand the changes undergone by first-year university students following the move to university, with particular focus on student identity and preoccupation. The study was also an investigation into the efficacy of a word-count approach compared with more manual corpus analysis methods. Taking as data the blog posts of thirty new students in the two months preceding and following the move to university, language changes over this period were examined. A LIWC analysis was carried out, using all of the standard LIWC2007 dictionaries, followed by a log-likelihood keywords analysis. In both cases, Corpus B (blogs written after the move) was compared against Corpus A (blogs written before the move). Results were examined manually using concordancer AntConc (Anthony, 2011). Corpus details are given below.

Corpus	**Tokens** (types)	**Total**
A: Blogs written before moving	232,242 (14,248)	389,721
B: Blogs written after moving	157,479 (10,536)	

Table 1: Corpus details

LIWC scores (for all 68 dimensions) were generated for each student's 'before' and 'after' blog posts. 'Change scores' were then calculated for each student, in each LIWC dimension, by dividing the LIWC scores for all of their entries written after the move to university by the LIWC scores for their entries written beforehand, then subtracting 1. This produced a negative score (a drop in LIWC score), a neutral score (no change), or a positive score (an increase in LIWC score). Overall LIWC change scores were then calculated for each category by subtracting the number of people for whom the change was negative from the number of people for whom it was positive. This was carried out three times, with different thresholds[3], and then the scores averaged. This final score was used to rank the LIWC dimensions and determine the categories, or dimensions, whose scores changed the most overall.

A keywords analysis was then carried out on Corpus B, using Corpus A as the reference corpus. Finally, the results for the LIWC analysis and keywords analysis were compared.

4.1 Findings

Table 2, below, gives the fourteen LIWC categories with the greatest overall change across all students, be it positive (+) or negative (-). However, the problem with results such as these is that they do not illustrate actual changes in word use. For function-word categories such as 'we', whose dictionaries contain a small number of unambiguous words, the LIWC score can paint a reasonably clear picture of general language changes. For larger, vaguer categories such as 'leisure', 'health' and 'religion', however, the scores alone cannot realistically convey what is happening in the data.

Categories with the greatest change scores			
future	-16.00	***filler***	+11.33
we	+15.67	***humans***	+10.33
see	-15.33	***health***	+9.33
leisure	-14.33	***excl***	+9.00
assent	+13.67	***cogmech***	+9.00
number	+12.67	***relig***	+9.00
motion	-11.33	***preps***	-8.67

Table 2: Categories with greatest change scores

[3] First, taking 0 as the threshold, i.e. anything above 0 was considered a positive change and anything below 0 a negative change; then with a threshold of -/+0.5; then with a threshold of -/+1. This was done to account for both the strength and the breadth of the LIWC changes.

The AntConc concordancer was used to examine the LIWC words in context, which helped to explain the results. The drops in 'future' and 'motion' scores, for example, were corroborated by the concordance lines; before university, the students were anticipating the move and used words such as 'gonna', 'will', and 'leave', which decreased once the move date had passed. Increases in 'number' and 'humans' scores were also predictable; the students are now *first*-year students, meeting *people* and joining *societies*. LIWC was also correct to identify a greater 'health' preoccupation; the new students were reportedly tired, hung-over and suffering from 'freshers' flu'.

The 'see' category score, however, was highly skewed by mentions of the word 'looking', as used in 'looking forward' [to university], which dropped following the move. This was therefore a somewhat misleading score change, since the students did not appear to be 'seeing' less – at least, not in the literal sense. However, a concordance analysis revealed some interesting changes in *how* they saw things; the construction *LOOK + adj* tended to feature quite general, positive adjectives before the move (e.g. 'good', 'nice'), with slightly more specific, critical adjectives being used after the move (e.g. 'weird', 'edgy').

The drop in the 'leisure' score suggested that the students were now engaging in fewer leisure activities, which may have been true, given their busy university schedules. However, this drop in score was also masking some *increases* in leisure words. The word 'reading', for example, was found to be used more frequently after the move. Going on word frequency alone, this might lead the researcher to assume that academic reading had become a greater preoccupation. However, on examining the context, a main cause was found to be the students' mentions of 'reading' with relation to their own blogs, which increased by almost half. A concordance analysis found that the students became increasingly concerned with the impressions they gave to readers, something which could not be identified using LIWC alone.

Increases in 'assent' and 'filler' were interesting, as these categories were meant for transcripts of spoken language. The results were characterised by word sense errors, namely the adjective 'cool' in the 'assent' category, and the verb 'like' in 'filler', but investigations into these categories using the concordancer still yielded useful findings: students were using words such as "yeah" and "so yeah" to relate to the reader, and "feel like" and "it's like" to describe their new university experiences. From this, and other findings, it was discovered that the bloggers displayed a greater preoccupation with their readership after the move to university. In this case, LIWC played a pivotal role in prompting this line of inquiry.

The most effective LIWC category was 'we', which made it possible to reliably track all mentions of first-person plural pronouns (though the referents of the pronouns had to be manually identified). Despite not being able to tell us to whom these pronouns referred, this small, closed-class category proved useful in measuring a sense of inclusiveness and collective identity. The fact that this dictionary is unlikely to be affected by noise and ambiguity made it possible to plot each student's individual 'we' scores on line graphs, demonstrating the rises and falls in these 'we' words on a post-by-post basis, over time.

The increase in the 'religion' score was of particular interest in the context of this study, as the literature suggests that students who move away for university tend to become *less* religious (Bryant et al., 2003). On closer examination it was found that the increase was mostly due to noisy matches such as 'seminar' (due to the inclusion of *seminar**, intended to match 'seminary' and 'seminaries'). Further erroneous matches were 'demonstration' (from *demon**), 'scuba diving' (*divin**) and 'monkeys' (*monk**). There were also a number of 'religious' words which were actually not religious in the context of student blogs (e.g. 'Christmas' as an end-of-term marker as opposed to religious holiday). In fact, when all LIWC 'religion' hits were manually checked, it was found that there was not an *increase* in religious uses of these terms, but a *decrease*.

When compared against the findings yielded by a keywords analysis, there was high overlap; out of the 38 findings of the study, 25 were shared by both the LIWC and keywords analyses. However, significantly more time was spent on 'unravelling' the LIWC results than those generated by the keywords, as some of the LIWC words triggered misleading categories due to contextual or morphological inaccuracies. For both LIWC and keywords, however, a manual examination of the context was crucial; out of all 38 findings, 28 relied upon the consideration of context. See Table 3 in the Appendix for a list of all findings.

5 Discussion

5.1 Theoretical Assumptions Revisited

Taking some of the above findings as examples, and drawing on other examples where relevant, the validity and implications of assumptions A1-7 from Section 3 are now discussed.

A1: *the frequency of a word can tell us something about a person or about the content or tone of a text.*

Several psychological studies have used word frequencies to show correlations between word use and the mind, due to latent, albeit crude, associations with words (Rosenberg, 1990; Mehl, 2006). The bag-of-words approach has been taken by many researchers in other fields, too; Biber (1988), for example, has successfully used word frequencies to discriminate text type and genre. Word frequencies were certainly useful in the student study, but had to be examined in context.

A2: *a computer program is ideal for counting words.*

Computers are undoubtedly more efficient at counting than humans. In the context of psychoanalysis, it has also been argued that computers are better at seeing 'past' meaning and counting the less interesting but nonetheless relevant language patterns to which a human annotator might be desensitised (Spence, 1980).

A3: *words have meaning in isolation.*

This assumption is problematic – or, in the view of Hanks (2013), false. Words, he argues, do not have meaning, but *meaning potential*; their meanings can only be activated by context. This is not to say that single words cannot act as discriminating features of texts, but that semantic value cannot legitimately be ascribed to them.

Words which are less affected by this problem are closed-class, i.e. function words. This would explain why, out of all of the categories analysed in Section 4, the 'we' category was found to be the most accurate and reliable. It might also explain why there are many successful LIWC studies concerning pronoun use (Pennebaker, 2011).

A4: *inaccuracies due to context-dependent factors are negligible or unimportant.*

The justification for a context-independent system is that a word-count program is probabilistic, and therefore such inaccuracies are, statistically, so rare that they do not impact on results in a serious way. This is probably true, overall, when considering all LIWC features together, due to high accuracy rates in some categories. However, there are some categories and domains for which this effect is particularly strong and *does* affect the results in a serious way. Bond and Lee (2005), for example, found LIWC to be reasonably accurate, but not accurate enough to be used in "high-stakes" investigations; in their study of deceptive statements, 30% were classified incorrectly.

It has also been argued that a general-purpose dictionary such as LIWC's cannot be accurately applied to all domains and discourses. Loughran and McDonald (2011), for example, found that when using the Harvard IV dictionary (a lexicon similar to that of LIWC), three quarters of all words classified as 'negative' were in fact not negative in the *context* of the financial domain, just as many 'religious' words were not religious in the *context* of student blogs. Again, such levels of inaccuracy could not be considered negligible.

A5: *by creating dictionaries* a priori, *and by finding and counting only the words which match them, the interesting or relevant parts of a text are identified.*

It is worth mentioning that 'religion', the category which suffered the most from inaccuracies in the student study, was one of the few dictionaries reported as having "100%" inter-judge agreement. We know therefore that 100% inter-judge agreement (between two judges) does not guarantee a well-compiled dictionary. But even if a content-word dictionary were impeccably constructed, with high agreement among hundreds of judges, it would still have the problem of being subjective and culture-specific (Mehl, 2006). A dictionary-based approach to text analysis therefore suffers from two biases: first, the top-down, pre-defined nature of its word-matching process (as opposed to a bottom-up, data-driven, inductive approach); and secondly the bias that comes with domain-specific, culture-bound dictionaries.

A6: *a score which corresponds to a dictionary also corresponds to the intended semantic value of that dictionary.*

Due to context- and dictionary-related problems, some categories used in the student study provided misleading scores, reflecting instead an increase in the use of words which were indicative of some other topics or events. A cursory glance at the LIWC scores, without actually looking at the text (which is what many LIWC analyses consist of), might lead a researcher to falsely conclude that moving to university is associated with becoming more religious or seeing less, for example.

A7: *dictionary values are more useful than single word frequencies.*

A problem encountered with LIWC, and presumably other dictionary-based approaches, is that dictionary scores do not tell us the actual linguistic changes that have occurred. Instead, we are given a simple numerical output. Despite being described as 'transparent', LIWC is, in this sense, surprisingly opaque.

The main issue with this approach, however, is that a LIWC score can theoretically mean nothing. Two texts might have the same LIWC score in the same dimension, and yet be made up of completely different words. Secondly, as in the case of ‘leisure’ in Section 4, a LIWC category change score might show an overall change in one direction, while simultaneously masking the opposite change for particular words within that category. Such problems are, fortunately, easy to overcome with the use of complementary qualitative analysis tools, such as a concordancer.

On the other hand, there are dictionary values which are arguably more useful than individual word counts. The ‘we’ category, for example, made it possible to track mentions of collective identity over time, something which would have been far less convenient to do otherwise.

5.2 Recommendations for the Use and Development of Text Analysis Software

There are two parties involved in any software use: the developer, and the end user. We therefore propose two main courses of action in order to maximise the benefits and avoid the pitfalls of an off-the-shelf text analysis package such as LIWC.

The developer could:

1. Formulate a list of the main analytical functions of the program and their perceived theoretical assumptions, as is done in Section 3 of this paper. Our own assumptions can be hard to determine without the help of others, so this should be a collaborative, peer-reviewed effort. This will help the developer to identify any potentially problematic assumptions embedded in their software.
2. Publicise the above information as a clear and concise "readme" document, along with the usual user manual and validation papers. This would ensure that the end user, whose background may be in an unrelated discipline, is easily made aware of the potential philosophical biases and constraints of the software, rather than simply knowing how to install and run it.
3. Attempt to avoid dictionary-related problems by thoroughly checking their contents for morphological errors and likely ambiguity. Employ raters from a range of cultural and educational backgrounds and ensure that at least one linguist is involved in the creation and validation of such modules.
4. Try to use bottom-up, data-driven approaches to dictionary population, if applicable.

The end user could:

1. First assess their own research needs and their existing theoretical assumptions, and to make sure that the software they choose is in line with those. Of course, this is only possible if the program's theoretical and philosophical underpinnings have already been established.
2. Combine top-down, pre-defined, quantitative analytical approaches with more bottom-up, inductive, qualitative approaches. This will add depth to findings and avoid misleading dictionary scores being taken at face value.
3. Favour smaller dictionaries with closed-class words, i.e. pronouns and function words, which tend to be less ambiguous in meaning.
4. Prioritise context: if necessary, create a customised, domain-specific dictionary suited to the research area; and always examine results in context, e.g. by using a concordancer.

6 Conclusions

This paper used the program Linguistic Inquiry and Word Count (LIWC) to exemplify some of the main advantages and potential pitfalls of off-the-shelf text analysis software. Given the growing popularity of computerised text analysis, it is important that reductive, word-count programs such as LIWC are used with caution, particularly by researchers outside of linguistics and natural language processing. It should be made especially clear to users that, far from being ‘objective’ and philosophically neutral, all computer software is based on theoretical assumptions, some of which are the subjects of ongoing debate.

As regards the efficacy of a word-count-based program such as LIWC, it appears that this approach has several limitations for content analysis. However, if both the program developer and the end user are careful and reflexive in their consideration of theoretical assumptions, such limitations can be addressed. LIWC appears to perform better in conjunction with other, more qualitative analysis tools, and it has become clear from the experiments presented that context is paramount when measuring meaning in texts.

Acknowledgements

The author is grateful to the anonymous reviewers for their helpful comments on an earlier draft.

References

Georg W. Alpers, Andrew J. Winzelberg, Catherine Classen, Heidi Roberts, Parvati Dev, Cheryl Koopman and C. Barr Taylor. 2005. Evaluation of computerized text analysis in an Internet breast cancer support group. *Computers in Human Behavior*, 21(2):361-376.

Christopher R. Agnew, Paul A. M. Van Lange, Caryl E. Rusbult and Christopher A. Langston. 1998. Cognitive interdependence: Commitment and the mental representation of close relationships. Journal of personality and social psychology, 74(4):939-954.

Laurence Anthony. 2011. AntConc (Version 3.2.2) [Computer software]. Tokyo, Japan: Waseda University. http://www.laurenceanthony.net.

Younggue Bae and Hongchul Lee. 2012. Sentiment analysis of Twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and Technology*, 63(12):2521-2535.

Pedro P. Balage Filho, Thiago A. S. Pardo and Sandra M. Aluısio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *9th Brazilian Symposium in Information and Human Language Technology, Fortaleza, Ceara*.

Douglas Biber. 1988. *Variation across speech and writing.* Cambridge: CUP.

Alyssa N. Bryant, Jeung Yun Choi and Maiko Yasuno. 2003. Understanding the religious and spiritual dimensions of students' lives in the first year of college. *Journal of College Student Development*, 44(6):723-745.

Michael A. Cohn, Matthias R. Mehl and James W. Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science*, 15(10):687-693.

Emma Franklin. 2015. *Fresher perspectives: the transition to student life examined using LIWC and a keywords analysis*. M.Res. thesis, University of Birmingham.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

Patrick W. Hanks. 2013. *Lexical Analysis: Norms and Exploitations.* Cambridge, MA: MIT Press.

Dan Jurafsky, Rajesh Ranganath and Dan McFarland. 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 638-646. Association for Computational Linguistics.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35-65.

François Mairesse and Marilyn Walker. 2006. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 85-88. Association for Computational Linguistics.

Matthias R. Mehl. 2006. Quantitative text analysis. In Michael Eid and Ed Diener (eds.) *Handbook of multimethod measurement in psychology*, pages 141-156.

Alexander Osherenko and Elisabeth André. 2007. Lexical affect sensing: Are affect dictionaries necessary to analyze affect?. In *Affective Computing and Intelligent Interaction*, pages 230-241. Springer Berlin Heidelberg.

Myle Ott, Claire Cardie and Jeffrey T. Hancock. 2013. Negative Deceptive Opinion Spam. In *HLT-NAACL*, pages 497-501.

Georgios Paltoglou, Stéphane Gobron, Marcin Skowron, Mike Thelwall and Daniel Thalmann. 2010. Sentiment analysis of informal textual communication in cyberspace. In *Proceedings of Engage 2010, Springer LNCS State-of-the-Art Survey*, pages 13-25.

James W. Pennebaker. 2011. *The Secret Life of Pronouns: What Our Words Say About Us.* New York: Bloomsbury.

James W. Pennebaker. 1997. Writing about emotional experiences as a therapeutic process. *Psychological science*, 8(3):162-166.

James W. Pennebaker, Roger J. Booth and Martha E. Francis. 2007. Linguistic Inquiry and Word Count: LIWC [Computer software]. Austin, TX: LIWC.net [online]. http://liwc.net.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales and Roger J. Booth. 2007. The development and psychometric properties of LIWC2007. Austin, TX: LIWC.net.

James W. Pennebaker and Martha E. Francis. 1996. Cognitive, emotional, and language processes in disclosure. *Cognition & Emotion*, 10(6):601-626.

James W. Pennebaker, Carla J. Groom, Daniel Loew and James M. Dabbs. 2004. Testosterone as a social inhibitor: two case studies of the effect of testosterone treatment on language. *Journal of abnormal psychology*, 113(1):172.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296-1312.

Veronica Perez-Rosas, Cristian Bologa, Mihai Burzo and Rada Mihalcea. 2014. Deception Detection Within and Across Cultures. In *Text* Mining, pages 157-175. Springer.

Stanley D. Rosenberg, Paula P. Schnurr and Thomas E. Oxman. 1990. Content analysis: A comparison of manual and computerized systems. *Journal of personality assessment*, 54(1-2):298-310.

Marcin Skowron and Georgios Paltoglou. 2011. Affect bartender - affective cues and their application in a conversational agent. In *IEEE Symposium Series on Computational Intelligence 2011, Workshop on Affective Computational Intelligence*.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24-54.

Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner and Isabell M. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178-185.

Appendix

	Finding	**Found with LIWC**	**Found with keywords**
1	More focus on collective self after the move (“we”, “our”)	X	X
2	**Social picture changes dramatically**	X	X
3	**Transition effects begin before the move date**	X	X
4	More focus on individual self after the move (“I”, “my”)	X	X
5	**Less preoccupied with media, celebrities, current affairs**	X	X
6	**More concerned with abstract ideas**	X	X
7	**More attempts to engage with/appeal to reader**	X	X
8	**More considered writing style**	X	X
9	More mentions of ‘first’, e.g. ‘our first lecture’	X	X
10	**Less general/philosophical after the move**	X	X
11	Fewer mentions of the (distant) future	X	X
12	Less focus on people, esp. other people, e.g. he, she, them	X	X
13	Preoccupation with moving before the move	X	X
14	More mentions of living arrangements	X	X
15	**Less focus on family after the move**	X	X
16	**More preoccupation with excursions, nights out**	X	X
17	**More tentative after move**	X	X
18	Less time spent on leisure activities	X	
19	Students undergo more dramatic changes than non-students	X	X
20	**More focus on food and the kitchen**	X	X
21	**Adjectives less generic**	X	X
22	**More interest in blogs and readership after the move**	X	X
23	**Appear more self-aware after the move**	X	X
24	**More mention of feelings**	X	X
25	**More comparisons and similes – describing, defining**	X	X
26	**Less preoccupation with general groups/society**	X	X
27	Fewer (rhetorical) questions	X	
28	**More advice and predictions after the move**		X
29	**More wisdoms before the move**		X
30	**Dip in ‘we’ words immediately before moving**	X	
31	**Preoccupation with flu after the move**	X	
32	**Concerned with sleep and lack thereof**	X	
33	**More mentions of favourite music, films, etc.**	X	
34	**Poetry more common after the move**	X	
35	**Fewer mentions of religious words**	X	
36	**Emphasis on what they are capable of doing after move**		X
37	**More focus on recent past; less reminiscing**		X
38	**More obligations before the move**		X

Table 3: All findings from the Franklin (2015) study regarding student changes, listed in descending order of amount of evidence to corroborate the finding. All findings in **bold** relied on examination of context to be found.