Gender Stereotypes in Film Language: A Corpus-Assisted Analysis

Lucia Busso §
CoLingLab-Università di Pisa*
l.busso0@fileli.unipi.it

Gianmarco Vignozzi §
Università di Pisa**
gianmarco.vignozzi@fileli.unipi.it

§ The research and the writing were carried out by both authors equally. G. V. is responsible for sections 1, 2, 3.1 and 3.2., L.B. for sections 3.3., 4 and 5.
* on leave at the Department of English and Applied Linguistics, University of Birmingham
** on leave at the Department of Linguistics, University of Sydney

Abstract

English: The present study concentrates on the representation and the reception of gender stereotypes. The analysis was first carried out on an ad hoc corpus of cult *romantic comedies and dramedies* of Anglo-American pop contemporary culture and secondly with a perception test. Both the corpus-driven analysis and the test results provide useful insights into the representation, recognition and entrenchment of gender stereotypes in language and in western culture. The preliminary findings generally confirm and validate the scientific literature, although showing some notable new elements.

Italiano: *Il lavoro si incentra sulla rappresentazione e la percezione degli stereotipi di genere. La ricerca è stata prima condotta su un corpus costruito ad hoc di film cult della cultura pop contemporanea anglo-americana appartenenti ai generi* romantic comedy e dramedy, *ed in seguito con un test di percezione. Il duplice approccio utilizzato fa luce sulla rappresentazione, il riconoscimento e il radicamento degli stereotipi di genere nella lingua e nella cultura occidentale. I risultati si trovano in linea con la letteratura, sebbene mostrino alcuni nuovi elementi.*

1 Introduction

In the era of digital revolution and screen proliferation, movies have undoubtedly acquired, thanks to their significance, a pivotal role in shaping our worldviews. In fact, popular films have the power to sway our collective imagination and influence our attitudes on crucial issues related to race, class, gender, etc. Characters in films reflect and perpetuate the status and options of them in today's society and culture, and thus play an active part in creating symbolic role models (Kord 2005, Bednarek 2015). Accordingly, it is interesting to examine the ways in which both females and males are represented on celluloid to better understand the ideologies they bear, and how gender identities are idealized. There seems to be wide agreement on the fact that characterization in filmic discourse heavily relies on archetypes and simplification (Culpeper 2001; Bednarek 2010). This is especially true in gender representation, as stereotypical roles simplify characterization in a way that it is easier to be received by the viewing audience. This, however, often results in an extreme polarization of gender roles. Film dialogues are therefore an ideal ground on which to study gender stereotypes and their linguistic representation and reception. Hence, this paper aims to fathom the discursive representation and the perception of well-established gender stereotypes in the dialogues of a sample of cult British and American romantic comedies, by integrating the tools of discourse analysis, corpus linguistics and perception analysis.

2 Films, language and gender

The nature of film language is still an object of debate. Movie scripts can be classified as texts that are "written-to-be-spoken-as-if-not-written" (Gregory & Carroll 1978: 42). Dialogues, in fact, portray a sort of "prefabricated orality" in that they are carefully written to be performed and sound natural to the audience, who longs for authenticity (Chaume 2012: 81). Corpus-based studies have proved that spontaneous conversation and scripted dialogues are very similar in nature, sharing almost the same array of lexico-grammatical features (Quaglio 2009, Bednarek 2010, Forchini 2012, Baker 2014, amongst others), but due to the evident need for clarity and speed in audio-visual texts, there may be changes in terms of their frequency. In fact, film scripts, sometimes tend to over-use features of spontaneous conversation (e.g.: greetings and leave-takings, Bruti & Vignozzi (2016)) both for dramatic reasons and to render the speech of characters as natural-sounding as possible.

Starting from the premises that gender is socially constructed (Cameron 2010) and that a large part of its perception relies on the observation of pre-established models, television and films provide the perfect field for examining generalized western social representation of accepted human behaviour (Shrum 2008). In this vein, verbal language becomes one of the pivotal means to create, reinforce and most importantly perpetuate stereotypical representations. Canonical research on language and gender has shown that traits such as hedges, empty adjectives, excessively polite forms, intensifiers, troubles talk etc. are more typical of women (Lakoff, 1975; Tannen 1994; Coates 1993), whereas males are associated with substandard and diatopically marked registers (Trudgill 1972; Tannen 1991) and a use of language that is aimed at retaining status and attention. However, nowadays many of these ideas have been partially rejected and framed as stereotypical norms around feminity and masculinity, which do not leave space for diversity (Cameron 2010, Mullany 2007; Bednarek, 2015). In recent times, corpus linguistics and computational linguistics have shown interest in analysing differences in language between genders (Argamom et al, 2003, Baker 2006, Herring & Paolillo 2006, McEnery 2006, Monroe et al. 2008, amongst others). This body of literature represents the backbone structure of our work, which aims to put together "corpus linguistics and gender analysis: two strands of linguistic research that do not go together frequently" (Kreyer 2014: 570).

3 Data and corpus driven analysis

The corpus. We compiled a corpus out of the orthographic transcriptions of eight English and American romantic comedies, using the web software *SketchEngine* (Kilgarriff et al. 2004, 2014). The films were chosen not only for their themes, but also for chronological coherence, as they cover approximately the first decade of the 21st century (table 1).

Title	Year	Nation
Sliding Doors	1998	UK
Billy Elliot	2000	UK
Bridget Jones' Diary	2001	UK/USA
Bend It Like Beckham	2002	UK
The Devil Wears Prada	2006	USA
Juno	2007	USA
Eat, Pray, Love	2010	USA
Letters to Juliet	2010	USA

Table 1: corpus rationale

The resulting corpus is therefore a synchronic *ad hoc* corpus of 95,036 tokens. We further subdivided it into two subcorpora consisting of the turns of female and male characters – respectively 55,766 (58.7%) and 39,270 (41.3%) tokens (henceforth: *M* and *F*). We chose to gather a new corpus – instead of relying on existing ones – to obtain a higher control on the data. Moreover, popular romantic comedies are the perfect humus for a polarized representation of gender roles, because of their content and intrinsic structure. As will be seen, however, our results are comparable with the ones extracted from much the larger film corpus *Cornell Movie-Dialogs Corpus*.[1]

Keywords and semantic domains clouds analysis. We used the online text analysis software *WMatrix* (Rayson 2003, 2004) to compare *M* and *F* both against each other and a reference corpus – the *BNC-spoken*. *WMatrix* performs automatic semantic analysis (of English) texts. This semantic analysis is carried out by a first POS tagging phase; the output is then semantically tagged from a set of 21 predefined semantic fields, further subdivided into 232 category labels for more fine-grained classification. Thus, from the comparative analyses starting from males and females' subcorpora, keywords and semantic domains clouds (calculated with log-likelihood statistic). Statistically significant items are the ones with LL values near or over 7, since 6.63 is the cut-off for 99% confidence of significance. The automatically obtained clouds were manually analysed to filter possible errors and select the more significant semantic domains associated with our sub-corpora. From the comparisons of the two sub-corpora against each other and against the *BNC Spoken*, we selected the most relevant semantic domains and keywords (i.e. with the higher LL values) for more qualitative-like evaluation. Tables 2 and 3 report the domains and the keywords that we selected.

[1] The fact that *F* is bigger than *M* should not come as a surprise. The film genre of romantic comedy is generally addressed to women and has therefore more female leading characters.

Sem. domains *F*	Sem. domains *M*
Business: Selling	*Industry*
Evaluation: Authentic	*Evaluation_Inaccurate*
Clothes and Personal Belongings	*Sports*
Time: New and Young	*Money_Generally*
Judgments of Appearance	*Greedy*
People: Female	*People: Male*
Kin	*Foolish*
Informal/Friendly	*Able:Intelligent*
Anatomy and Physiology	***Anatomy and Physiology***
Intimacy and Sex	***Intimacy and Sex***

Keywords *F*	Keywords *M*
Feelings (in_love, love)	*Friendship (lads, man, mate)*
God, oh God, my God	*Swearing (fuck, fuck off, fucking)*
Swearing and Euphemisms (Shit, Shagging)	*Right, all_right*
Mom	*Dad*
Politeness (Thank You, Sorry)	*sorry*
People (Me, My, You)	

Table 2 and 3: WMatrix semantic domains and keywords used in the test

As it can be seen, in our corpus women tend to speak about shopping, cleaning, personal care, and family, whereas men appear to discuss money, sports, work and male friendship. In table 2 are also present semantic domains which were relevant for both *M* and *F* speech, i.e. “Anatomy and Physiology” and “Intimacy and Sex” (in bold). These last two domains may emerge as strongly relevant due to corpus-specific reasons.

Romantic comedies, in fact, are most often centred around romantic and quite physical relationships. However, what we think is of interest when analysing the overlapping between semantic domains between females and males is the different wording. Women and men refer to their bodies and their relationships in different ways, which are consistent with a polarization of gender roles (E.g.: *breasts* vs. *boobs*). Keywords are also worth mentioning. Their evaluation showed that women make larger use of politeness forms, while men resort to more swearwords and interjections, such as “right, all right”.

Interestingly, the tendencies that emerged from our small corpus are in line with Schofield and Mehr (2016)’s analysis of the *Cornell Movie-Dialogs Corpus* (Danescu-Niculescu-Mizil et al. 2012a), a vast corpus of more than 600 films of different genres. The similarity of the results gave us confidence in using the stereotypical representations of genders’ speech to investigate its reception by means of a test.

The test. With the aim of testing the reception and entrenchment of gender stereotypes in speakers, we developed a perception test based on the results of our corpus-driven analysis. We manually extracted 18 lines per subcorpus[2], each containing one or more of the stereotypical semantic domains and keywords that emerged from the previous *WMatrix* analysis. The resulting 36 extracted lines were used as stimuli in the perception test[3]. The choice of such limited number of sentences was determined by two reasons. The first, theoretically motivated, was not to repeat the same keywords and stereotypes too many times. Such repetition, in our opinion, could have influenced or biased the participants. The second reason, of a more practical nature, was to construct a reasonably-sized test to maintain participants’ attention and avoid fatigue, which could have influenced the responses. We extracted film lines containing a variable concentration of stereotypes, ranging from sentences referring to only one to several stereotypical domains. The selection was done manually, based on the rather obvious hypothesis that sentences more “stereotypically dense” would be recognised more easily. The stimuli-sentences were also chosen as deprived of context as possible, in order not to give any clue about the film of origin. Proper names were omitted, and when this was not possible, substituted with the string [XXX]. For example, in (1) the name of the male romantic partner was obscured so that the only clue to the gender of the speaker would be the linguistic stereotypes (shopping, mitigated swearwords, weaving).

1) *When [XXX] and I broke up for two weeks, I bought a loom, a frigging loom*

The test was presented to 22 native, bilingual or highly proficient speakers of English, 15 women and 7 men (mean age: 39.5). The task was to decide whether a given sentence had been uttered by

[2] The stimuli-sentences were chosen to be as representative as possible of the entire corpus: they are evenly distributed among all the films of the corpus, with two or three instances from each film for each subcorpus.

[3] For reasons of space we do not include the complete list of the sentences extracted and used for the test. Several examples are reported in the text and in following footnotes.

a man or a woman. In order not to force participants to a necessarily binary choice, the option "I don't know" was also included. We additionally asked speakers to specify words, expressions or general concepts that influenced their answers. This provided us with interesting insights into participants' process of thinking and categorizing.

4 Results

Several interesting considerations arise from the analysis of the data. Firstly, it appears that overall the stereotypes were correctly spotted and categorized.

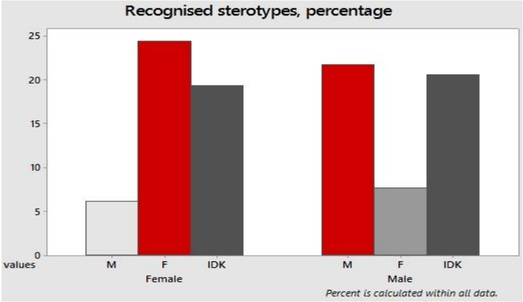

Chart. 1: Percentage of recognised stereotypes (in red)

However, it also emerges that female stereotypes were more unambiguously recognisable, with fewer answers assigned to the other gender or to the "I don't know" category (chart.1).

By examining more closely the results, a subdivision of the data can be made to account for the differences in it: recognised (more than 50% correct), ambiguous (between 25-50% correct) and completely misunderstood (less than 25% correct) stereotypes. Table 4 illustrates the distribution of answers in the three frequency slots.

	> 50%	**25-50%**	**< 25%**
***F* LINES**	**61,1 %**	27,8%	11,1%
***M* LINES**	33,3%	38,9 %	27,8%

Table 4: distribution of participants' answers

As was firstly hypothesized, sentences with a higher "density" of stereotypical keywords or semantic domains were usually the ones that speakers better recognised. Stimuli in the first group, therefore, consist of clear-cut and well recognisable clusters of linguistic and conceptual stereotypes[4]. The second group is instead formed by stereotypes that were recognised by a substantial part of the informants, but not by the majority. This, in our opinion, may be due to several factors: some concepts, for example, could be perceived as less prototypical than others. In addition, some linguistic features (e.g. discourse markers) were not fully recognised as stereotypical due to our limitation to the written dimension. Prosody, contextual information and multimodality are in fact fundamental aspects of language that were inevitably excluded from our experimental design[5]. Finally, the last group consists of stereotypes that were not perceived as such by speakers (e.g.: family as a typical argument of women's speech), and of what we called *reverse stereotypes*. That is, utterances that conceptually represented ambiguous events or anti-prototypical situations: a woman swearing, a man talking about his feelings.[6] As predicted, these stereotypes were not recognised at all by participants, who tended to assign them to the opposite gender. It is interesting to note that also some male-produced sentences were not recognized by our informants, perhaps due to the composition of our corpus. Several predominant keywords and domains in *M*, in fact, may be strictly related to the chosen film genre. For example, the massive presence of the *WMatrix* domain *Evaluation_inaccurate* -- i.e. apologies --reflects the archetypical situation in romantic comedies of men apologizing for their mistakes to women. Being so context-related, however, speakers were not able to correctly locate sentences containing expressions from this domain.[7]

Another aspect that was taken into consideration in our analysis was the gender of the informants, to see if a relation with the data could be recognised. There was a statistically significant difference between the gender of the participant and the answer to the test (H (2) = 9.2388, p-value = 0.0024, *Kruskal-Wallis* test with *Wilcoxon post-hoc, Bonferroni* p-value correction).

A chi-square test of independence was performed as well to examine the relation between gender of the speaker and responses given.

[4] E.g.: "*Give me the bag! I've got to get some proper shoes for the wedding now*" (71%) (f); *"What are you doing, eh? You're me best mate!*" (82%) (m).

[5] E.g.: "*God! My mum had a fit when she saw the boots!*" (47%) (f); "*He's a kid. He's just a fucking little kid*." (47%) (m).

[6] The reverse stereotypes utterances are the following.

I. *Oh, shit! I stubbed my foot on the side of the shagging bath! (f)*
II. *This is the first time in 18 years I'm going to be able to call the shots in my own life*! *(m)*

[7]*- I made a mistake, such a big, BIG mistake and I'm sorry. I'm truly, truly sorry.*
- We accept that we fight a lot, and we hardly have sex anymore, but we don't wanna live without each other.

The relation between these variables was significant. (χ^2 = 10.298, p-value= 0.0058).

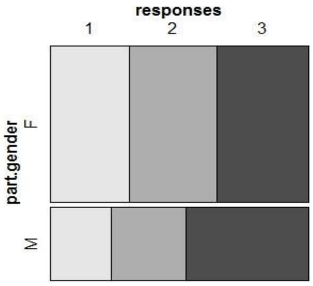

Chart. 2: mosaic plot of the results divided by gender.

Chart 2 shows the difference in male and female informants' answers. The numbers of the variable "responses" indicate the three possible answers of the test: "male" (1), "female" (2), "I don't know" (3). As it can be seen, men assigned overall more utterances to the "I don't know" option rather than to one of the two genders. Women, instead, show a fairly equal distribution of responses among the three conditions. Furthermore, both men and women assigned more utterances to female characters than to male ones (see table 5). This result is in line with the fact that women stereotypes were better recognised overall, in the sense that fewer answers were assigned to the other gender.

	MEN	WOMEN
m	23%	30%
f	29%	34%
idk	48%	36%

Table 5: distribution of informants' answers divided by gender of the speaker

Other useful insights into the data came from the words our informants identified as relevant to their decision. In fact, two tendencies emerged: speakers either indicated specific words, collocations or phrases, or answered with abstract concepts and pragmatic inferences based on the utterances. Interestingly, words and expressions exactly replicated keywords, while general and abstract concepts reflected the semantic domains that emerged in the corpus analysis. In addition, several speakers performed actual pragmatic inferences based on the stereotypical concepts contained in the sentences. For example, to (2) subjects reacted either with a specific word like in a) or with a more general consideration as in b).

2) *Ooh, you must feel like you're about to find your long-lost soul mate!*

a) "soul mate"
b) talking about feelings in general

5 Conclusions

The present paper proposes an original take on investigating gender stereotypes in language. The novelty in our approach lies in the hybrid methodology that falls neither in the tradition of the literature on "gendered discourse" nor in the more recent field of corpus linguistics, but combines the two and adds insights from psycholinguistics as well. This kind of integrated analysis provided us with preliminary results that help identify gender archetypical roles, behaviours and linguistic representations in modern western culture. What is interesting to note is that the gender representations coming to light from our corpus of pop-culture films are based on features that are now dismissed as clichéd and stereotypical by the literature (see Cameron 2005, 2010; Bexter 2006), but which seem to be nonetheless entrenched in our interpretation of reality.

The archetypical depiction of characters is particularly evident in popular comedies, which do not examine characters' psychology in depth. The test validated our assumption that film language stereotypically portrays the way in which men and women talk drawing on recognisable traits attached to femininity and masculinity in our culture. In fact, speakers were mostly able to correctly assign the utterances to the right gender.

In addition, all our informants showed metalinguistic –or second-level –awareness about stereotypical concepts and linguistic clues, and several of them also provided us with insightful and creative inferences based on the event described in the utterance. We interpret this as a sign of stereotypes being conceptual in nature, deeply entrenched in our representation of the world and accessed via linguistic clues. The "reverse stereotypes" also reinforce this idea.

Acknowledgements: We would like to thank professors Silvia Bruti, Alessandro Lenci, Belinda Crawford and Monika Bednarek for useful comments. We also gratefully acknowledge the reviewers for their valuable comments.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. Text & Talk, 23(3), 321–346.

Paul Baker. 2006. Using Corpora in Discourse Analysis. London: Continuum.

Paul Baker. 2014. Using Corpora to Analyze Gender. London; New York: Bloomsbury Academic.

Monika Bednarek. 2010. The Language of Fictional Television: Drama and Identity. London: Continuum.

Monika Bednarek. 2015. Corpus-Assisted Multimodal Discourse Analysis of Television and Film Narratives. In P. Baker, T. McEnery (Eds.), Corpora and Discourse Studies: Integrating Discourse and Corpora, 63-87. Basingstoke, UK: Palgrave Macmillan.

Silvia Bruti and Gianmarco Vignozzi. 2016. Routines as social pleasantries in period dramas: a corpus linguistic analysis. in R. Ferrari,S. Bruti (eds), A Language of One's Own: Idiolectal English, pp. 207-239, Bologna: I libri di Emil.

Deborah Cameron. 2010. Gender, Language and the New Biologism. Constellations, 17 (4), 526–39.

Frederic Chaume. 2012. Audiovisual Translation: Dubbing. Manchester, St Jerome.

Jennifer Coates. 1993. Women, Men and Language. London: Longman.

Jonathan Culpeper. 2001.Language and characterisation: people in plays and other texts. Harlow: Longman.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You had me at hello: How phrasing affects memorability. In Proceedings of ACL, 892–901.

Penelope Eckert and Sally McConnell-Ginet. 2003. Language and Gender. Cambridge: Cambridge University Press.

Pierfranca Forchini. 2012. Movie Language Revisited. Evidence from Multi-Dimensional Analysis and Corpora. Bern: Peter Lang.

Michael Gregory and Susanne Carroll. 1978. Language and Situation: TV Heroines: Contemporary Screen Images of Women. Lanham: Rowman & Littlefield.

Susan C Herring and John C Paolillo. 2006. Gender and genre variation in weblogs. Journal of Sociolinguistics, 10(4), 439–459.

Janet Holmes. 2006. Gendered Talk at Work: Constructing Gender Identity through Workplace Discourse. Oxford: Blackwell.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovvář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography,* 1, 7-36.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, David Tugwell. The Sketch Engine. 2004. Information Technology. Available online at: *www.sketchengine.co.uk*

Susanne Kord and Elisabeth Krimmer. 2005. Hollywood Divas, Indie Queens, and Language Varieties and their Social Contexts. London/New York: Routledge.

Rolf Kreyer. 2014. Review: P. Baker. 2014. Using Corpora to Analyze Gender. London/New York: Bloomsbury. International Journal of Corpus Linguistics 19, 570-575.

Robin Lakoff. 1975. Language and woman's place. New York, NY: Harper & Row.

Tom McArthur. 1981. Lexicon of Contemporary English. London: Longman.

Anthony M. McEnery, Richard Z. Xiao & Yukio Tono. 2006.Corpus-based Language Studies: An Advanced Resource Book. London/New York: Routledge.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372–403.

Louise Mullany. 2007. Gendered Discourse in the Professional Workplace. Basingstoke, NY: Palgrave Macmillan.

Paulo Quaglio. 2009. Television Dialogue: The Sitcom Friends vs. Natural Conversation. Philadelphia: John Benjamins.

Paul Rayson. 2009. Wmatrix: A Web-based Corpus Processing Environment. Computing Department, Lancaster University. Available online at: http://ucrel.lancs.ac.uk/wmatrix/

Paul Rayson, Dawn Archer, Scott Piao, & Anthony M. McEnery. 2004. The UCREL semantic analysis system. Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop, Lisbon, Portugal, 2004.

Alexandra Schofield and Leo Mehr. 2016. Gender distinguishing features in film dialogue. NAACL CLfL.

L.j. Shrum. 2008. Media consumption and perceptions of social reality. In J. Bryant &M.B. Oliver (eds.), Media Effects: Advances in Theory and Research, 3rd Edition. New York, NY: Routledge.

Mary M Talbot. 2003. "Gender Stereotypes: Reproduction and Challenge". In Holmes, J. & Meyerhoff, M. (eds.), The Handbook of Language and Gender. Oxford: Blackwell, 468-86.

Deborah Tannen. 1991. You just don't understand: Women and men in conversation. Virago London.

Deborah Tannen. 1994. Gender and Discourse. New York: Oxford University Press.

Peter Trudgill. 1972. Sex, covert prestige and linguistici change in the urban British English of Norwich. Language in Society 1, 179-195.