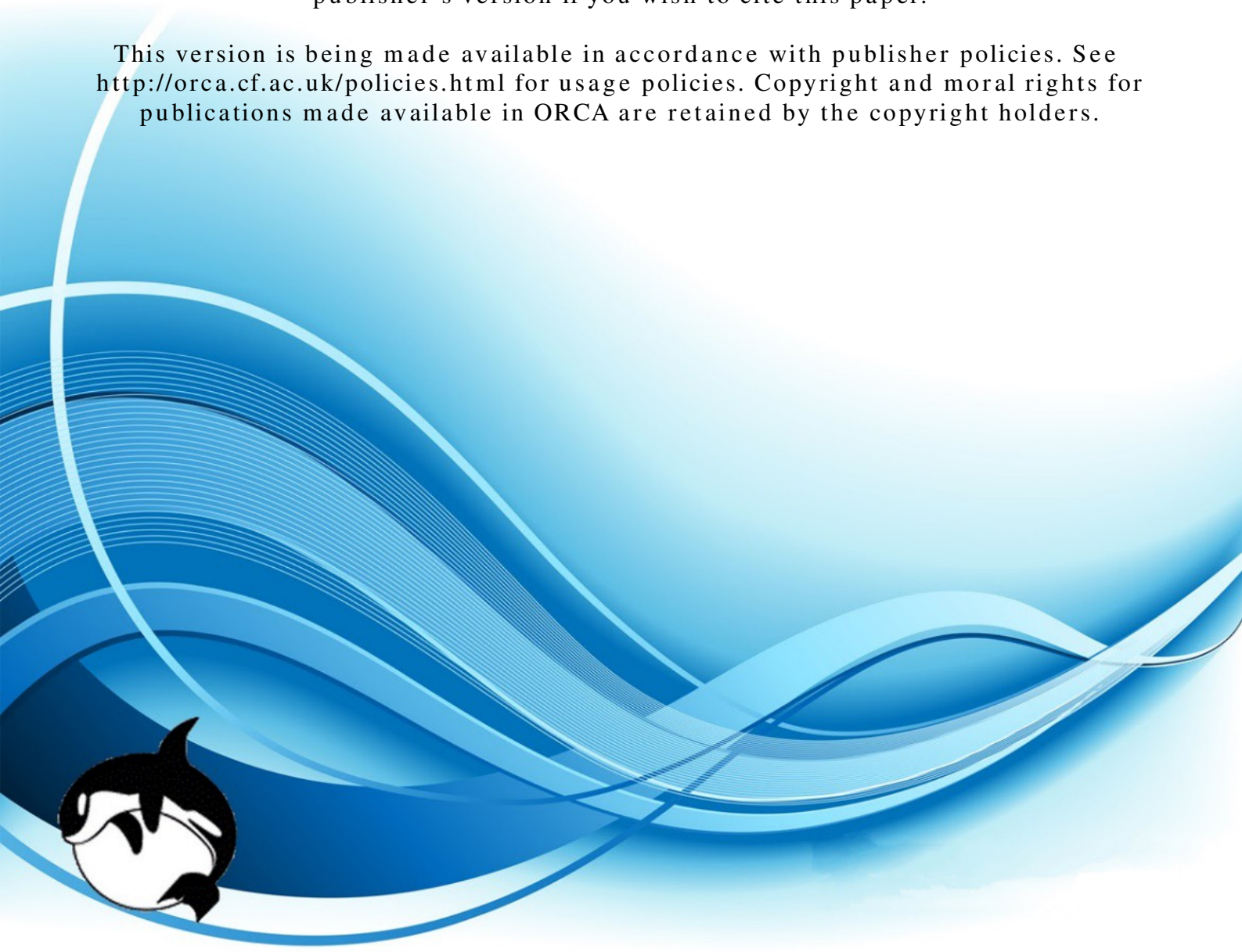

I'm having a spring clear out: A corpus-based analysis of e-transactional discourse

1. Introduction

In the contemporary digital age, one of the most common activities which people engage in is online buying and selling. Current estimates suggest that Amazon has around 244 million customers, while eBay has around 120 million (customers being defined as anyone who has made a purchase within the past 12 months). In light of this explosion of activity in the digital marketplace, applied linguistics has an important role to play in developing closer understandings of online transactions and in furthering research into the field of digital discourse.

This article outlines a number of ways in which the language of eBay descriptions is characterised according to whether transactions are listed by experienced or novice eBay sellers and whether items are new or used. A full description of these linguistic and discourse features is presented in the findings and discussion sections of the paper. In addition, the article suggests how online texts might be analysed by using a site's own categories to describe and analyse the discourse, focusing on sellers' definitions and descriptions of products, as advertised for sale on EBay. In addition, we demonstrate the ways in which identity plays a key role in categorising the features of online discourse by, for example, considering whether an individual has expertise in a particular field or not. We believe that the paper contributes to current debates on the precise nature and focus of discourse analysis and to the study of digital discourse more specifically.

When examining online discourse, traditional discourse differences, such as this distinctions between spoken and written texts, become rather blurred, particularly when we consider transactional and social online encounters, such as those taking place on eBay. In this context, based on findings from the study presented here, differences such as social distance and degree of formality – features which can be used to analyse and describe spoken and written texts – are less relevant to the encounter than whether sellers are experienced or not, and whether items are new or used. In other words, an analysis of online discourse might be more convincing if traditional contextual features are abandoned and those more immediately relevant to the genre are used instead. These features, we suggest, give transactional online discourse its textuality and provide greater insights into the ways in which language is used to facilitate buying and selling on social media sites.

Transactional discourse is typically characterised as utilising specific lexical items which help to provide an 'optimally efficient transmission of information' in its exchanges (Brown and Yule, 1983; Lakoff, 1989). Transactional discourse is commonly used in commercial contexts to convey factual information, often as a means of persuading third parties to 'buy-in' to what is being communicated and indeed, often, actually to purchase products. Interactional discourse, in contrast, 'has as its primary goal the establishment and maintenance of social relationships' (Kasper, 1995: 205). While language is never typically of one single variety (i.e. transactional *or* interactional, see Watzlawick et al., 1967; Kasper, 1990; Nattinger and DeCarrico, 1992 and Held, 1995), a dominance of one type of goal is often present.

In this paper, we look at the ways in which transactional and social interactions are accomplished through the use of specific lexical and discourse features. Our aim is to characterise patterns of language use in online auction sites (the 'e-marketplace'), with a specific focus on the language used in product descriptions listed on eBay. Since the birth of the 'e-marketplace' in the 1990s, the language of sites such as eBay has quickly emerged into

a very distinct genre of online discourse, a type of discourse with identifiable communicative events and clear purposes.

E-marketplaces aim to encourage/persuade the e-buyer to choose their item, and for listed items to be sold at the highest rate possible. Given this, the language used in the descriptions featured on eBay can be broadly categorised as a form of discourse which has a text-external goal of being 'e-transactional'. It has a number of features, which include the primary goal of selling all items through asynchronous activity, and communication which does not "*require that users be logged on at the same time in order to send and receive messages"* (Herring, 2007: 13). There is typically no direct form of interaction between buyer and seller (unless specific product-related questions are asked by the buyer) until the auction has ended, a feature which might mean, for some sellers, that the construction of personal identities in the product descriptions and this establishment and maintenance of relationships might also be important. These are features that are likely to emerge and/or be reflected in the language when the product descriptions are used. The extent to which this is true is investigated in the present article.

Through a corpus-based based analysis of a sub-corpus of a 6.3-million-word collection of eBay data, we aim to gain a better understanding of variance in descriptions, depending on two key factors: seller experience and item condition (a high number of 'new' listings for a particular seller is likely to be a sign of a more experienced seller). This paper describes the methods and approaches used to construct the eBay corpus and provides a detailed analysis of a sub-corpus of data as a means of identifying some of the key transactional and interactional features found in this genre of discourse.

2. Methodology

2.1. Aims and approach

eBay was chosen as the source website for the present study as this is the largest global e-marketplace in existence today. It is suggested that the eBay site is, therefore likely to provide data which a good representation of the typical patterns of language that is used in this specific genre of online discourse. The 10 five 'most popular' product categories (details in sections 2.2 and 2.3 below), defined in terms of the volume of individual listings included for each, were then selected for inclusion in this corpus. The current paper focuses specifically on the language used in one of these categories, shoes, which was selected at random from the initial five.

The research questions that are addressed in the study are:

1. How are different products (specifically 'shoes') described in e-marketplace contexts?
2. What specific linguistic features can be identified in the sales discourse of (a) experienced and (b) inexperienced sellers?
3. What do these descriptions tell us about the ways in which the identities of the sellers are constructed?

In this paper, our analysis utilises a multi-layered, corpus-based approach as a means of interrogating the data. From the data, we characterise some of the features of the ways in which 'new' and 'used' shoes are described. We also compare and contrast listings provided by 'experienced' and 'novice' sellers (i.e. with feedback scores of over 1000 (o1000) and under 100 (u100) respectively).

Corpus-based research focuses on defining and exploring recurring *patterns* in language use. If we want to understand the patterns of use of a particular word we first need to find out how common this word is in a language. The typical 'way-in' to the analysis of

corpora is through generating frequency lists, to map out and, as required, compare and contrast how frequent particular word forms are across either an entire corpus or across particular sub-sets (sub-corpora). As the present study is explorative, it seeks to 'describe and explain the observed phenomena' (Sinclair, 2008: 30). Consequently, the utility of basic frequency counts is viewed as an appropriate initial step for the analysis presented here.

However, to add a certain level of robustness to the analyses we complemented these frequency counts with log-likelihood scores. These provide a basic statistical measure of the relationship between the frequencies, indicating whether specific patterns of significant differences are likely to exist by chance or not. In the next section, where log-likelihood scores are presented, a '+' log-likelihood score (ll.) indicates that a particular rate of use is statistically higher in the first cited variable compared to the other variable defined (e.g. new vs. used products). The relative frequencies denote the number of times the specific search term (i.e. 'word') is used at a 'per word' rate in the entire sub-corpus. Typically speaking, a statistically 'significant' difference in the frequency of usage across two compared parameters/sub-corpora, as defined by the log-likelihood score, is signified when the *p* value is <0.01 (with a critical value range of >6.63). So a '+' indicates a statistically higher frequency of a word in one sub-corpus, thus a statistically lower frequency in the other sub-corpus it is compared against.

Beyond frequency counts and log-likelihood comparisons, explorations of key collocates of search terms, and some more semantically based analysis of specifically chosen concordance outputs, will be used as the basis for mapping patterns of the use of the words and clusters used across the sub-corpora to gain a better understanding of their roles and functions in the e-marketplace context.

Rayson's WMatrix software (2003) and Scott's Wordsmith Tools (1999) have been selected to carry out these analyses. Both of these concordance tools include utilities for carrying out word, cluster and parts of speech queries (centring on the production of key word lists and key-word-in-context, KWIC, outputs), allowing users to explore the patterned use of these features in a corpus. In addition to this, with the use of the WMatrix semantic tagger, common themes and semantic associations connected with corpora can also be queried using the software and comparisons between sub-corpora can easily be made (based either on raw frequencies or semantic associations).

2.2. Extracting the data

In terms of creating the corpus, eBay offers a rich set of APIs (Application Programming Interfaces) that can be used to access its platform and we used these as a basis for extracting the data. The APIs were accessed via writing custom software scripts that search and retrieve information from the online auction site. The scripts were written in PHP (PHP: Hypertext Pre-processor), a popular server-scripting language, while data downloaded was saved in a Structured Query Language (SQL) database, making it possible to search for and export data easily.

The methodology adopted revolved around 3 main steps. In the first step, a list of potential categories of products was compiled. In the first instance, 5 main thematically linked categories that are defined as being amongst the 'most popular'[1] were targeted. These included Cars, Clothing, Electronics, Furniture and Shoes, each of which is further divided into specific product sub-categories. So, for example, for electronics, data from the following product categories were included: consoles, games, laptops and netbooks, iPads, tablets and e-readers, iPad and tablet accessories, mobile and smart phones and mobile phone and PDA

[1] As listed on http://pulse.ebay.co.uk

accessories. A manual check on eBay's website returned the identifiers for each of the product categories, which were then inputted into our application.

Once the list was compiled, the application used eBay's API to retrieve listings of products for each of the categories. Although it was possible to impose filters on the list, we did not apply them as we wanted a random selection of products.

The only filter imposed was a maximum limit of products per category (typically 2000 products per category). In the second step, our software iterated around the product listings and downloaded all the available information about them. In compiling our corpus we were primarily interested in the description of the products. Additional secondary information such as their condition, the seller's feedback scores and product location were used for creating sub-datasets for comparison purposes. For example, we were able to create two sub-datasets of products offered by novice vs. experienced sellers. Secondary product information was not used in the linguistic analysis. In the third step the data was exported in the necessary format for it to be analysed further.

2.3. Corpus contents

In total we compiled a corpus of over 10 million words of product description data from the UK eBay site. The word count distribution of data across the 5 main data categories is presented in Table 1 (each with 2000 listed products). The sub-corpora include content from a wider range of different participants in a range of different contexts.

No.	Sub-category	Category	Total no. of words
1	CARS	Car accessories	486,032
2		Car manuals and	380,559
3		Car parts	324,862
4		XCar tuning and styling	400,026
5		Car wheels, tyres and trims	348,788
6		Cars	376,546
7		Classic cars	537,616
8	CLOTHING	Boy's clothing	123,444
9		Girl's clothing	117,524
10		Men's clothing	266,791
11		Women's clothing	301,789
12	FURNITURE	Bedroom furniture	258,145
13		Living room furniture	180,939
14		Storage	355,483
15	ELECTRONICS	Consoles	326,165
16		Games	421,148
17		Laptops & netbooks	671,457
18		Laptop accessories	1,618,312
19		iPads, tablets and	747,920
20		iPad and tablet	842,569
21		Mobile and smart phones	403,387
22		Mobile and PDA accessories	694,721
23	SHOES	Men's shoes	289,691
24		Girls' shoes	111,259
25		Boys' shoes	112,762
26		Women's shoes	134,611
			6,374,002

Table 1: Contents of the eBay corpus (taken from the UK eBay site).

As a preliminary study of this corpus, the current paper focuses on the 'shoes' data, a sub-corpus which was selected at random from the entire eBay corpus. This sub-corpus amounts to circa 650,000 words across 8000 product descriptions (2000 for each sub-category). We can see from Table 1 that the average length, in words, of the product descriptions was fairly consistent across the girls', boys' and women's listings (at 57, 56 and 67 words per description, respectively), but much higher for the men's shoes (with average rates of 145 words per description) in this sub-corpus.

As seen from Table 2, there is also a higher number of word 'types' used in the men's description (i.e. range of different word families used). However, when calculating the type-token ratio of the descriptions, which is calculated by dividing the number of types of words

by the net number of words used in a text and multiplying the result by 100, we see a much higher ratio for women's, boys' and girls' shoe descriptions (circa 6.5, 6.4 and 6 respectively) than the men's shoes (at a ratio of 4). This suggests that these product descriptions have a lower lexical density than the others as the lower the type-token ratio score, the less varied the text in terms of the types of lexis used. We can infer from this that there is more repetition in the descriptions of men's shoes and the text is likely to be less complex and/or dense.

	Shoes	Men's	Women's	Boys'	Girls'
Entries	8000	2000	2000	2000	2000
Total types	21962	11551	8749	7170	6664
Total words	648323	289691	134611	112762	111259
New	370689	211903	71621	44255	42910
Used	277634	77788	62990	68507	68349
u100	190149	27654	28814	101628	32053
o1000	254224	211505	21884	6252	14583
u100u	119539	17151	17293	66428	18667
u100n	70610	10503	11521	35200	13386
o1000u	48612	24514	14636	0	9462
o1000n	205612	186991	7248	6252	5121

Table 2: Contents of the UK shoes sub-corpora.

As a means of drilling even further into the data, we sub-divided it according to whether products were listed as new or used and the level of online selling experience of those listing the items. Once a product is sold on EBay, the buyer has an opportunity to provide feedback on the service provided by the seller. This can include the addition of comments relating to the quality of the product; the level and detail of the communication between buyer and seller; the speed of delivery of the product and so on. The buyer can also provide a score for the feedback, from 1 to 10, which is indicated by a coloured star on their profile. Over time the seller will accrue numerous feedback scores which indicate how many products that they have sold. So, for example, feedback scores of 95 (regardless of the individual rate/ coloured star), indicates that the seller has sold 95 products in the past. The higher the feedback score, therefore, the more experienced the seller. Based on this, we crudely defined any seller who has feedback scores of 100 or less as relative novices on eBay, while those who have feedback scores of 1000 or more are defined as being more experienced. As with the new vs. used descriptions, comparisons between the language used in listings made by experienced and novice sellers is presented in 3 of this article.

Details of the number of words used in descriptions of new and used products as well as numbers for sellers who have feedback scores of 100 or less (u100), over 1000 (o1000) and used/new items that have scores of u100/o1000 (i.e. u100n/u100u, o1000n/o1000u and so on) are also seen in Table 2.

From this table, we see that the number of words in the new, used, under 100, over 1000 (for feedback) categories, and varieties thereof, are highly variable from one sub-corpus to the next. Given that the sizes of the sub-corpora examined are unequal, the results are normalised using statistical measures to make frequencies from samples of markedly different sizes comparable by bringing them to a common base. This is explained in more detail in the following section.

3. Results

3.1. New vs. used listings

Results of simple keyword comparisons of the new and used data are presented in Table 3. Keywords are described by Scott (1999) as words which have a significantly higher or lower frequency in a target text or corpus compared to a reference corpus. A reference corpus is a corpus which purportedly provide a representative account of 'general' language use (thus providing patterns of usage which are as close to a 'norm' as possible). The reference corpus utilised here is the 100 million word British National Corpus[2] (BNC). Positive keywords are those which are significantly more frequently in the target text/corpus while negative keywords are significantly less frequent in the target. Keywords can be calculated automatically using the Keyword tool in Wordsmith Tools (Scott, 1999). Similarly, keyword clusters are phrases (collocating words) which appear at a statistically higher or lower rate in a target corpus compared to the reference corpus.

	New	Used		New	Used		New	Used		New	Used
1	we	good	16	shipping	them	31	checkout	pet	46	womens	pink
2	your	worn	17	may	plenty	32	the	pair	47	mail	girls
3	you	used	18	customer	they	33	shop	me	48	fee	leather
4	our	Boots	19	if	still	34	number	few_times	49	goods	shoes
5	us	I	20	exchange	hardly	35	purchase	soles	50	direct	life
6	brand_new	but	21	refund	trainers	36	footwear	once	51	store	marks
7	return	my	22	must	Mobile	37	after	great	52	information	a
8	boots	posted	23	order	nike	38	service	boys	53	details	these
9	address	wear	24	that	home	39	day	happy	54	deliver	lots
10	to	Velcro	25	bank	other	40	not	left	55	policy	smoke_free
11	item	excellent	26	will	thanks	41	tags	free	56	an	clean
12	business	clarks	27	never	with	42	working	lovely	57	shipped	scuffs
13	days	in	28	be	looking	43	email	black	58	form	times
14	delivery	very	29	it	bidding	44	is	few	59	contact	an
15	or	only	30	returned	smoke	45	can	twice	60	pay	check_out

Table 3: Keywords and key clusters used in the new and used shoe descriptions.

On first inspection we see that practical descriptions of how products may be paid for, packaged and transported are listed as being frequently used in the new product descriptions. Words such as *return*, *delivery*, *mail*, *address*, *shipping*, *returned*, *deliver* and *shipped* all feature in the top 60 most common words in this sub-corpus and are all used at a significantly higher rate than in the used descriptions (to ll.>6.63). The use of words with a semantic relation to the sense of obligation, quality, time and speed and possession as well as to the notion of help and service (including contact details and information provision) are also found to be used at a statistically higher rate in the new shoe descriptions compared to the used descriptions (to ll.>6.63), with words and clusters such as *brand new*, *exchange(d)*, *order*, *refund*, *bank*, *checkout*, *shop*, *purchase*, *pay*, *fee*, *store, customer*, *service, information, details, day(s), direct, email* and *contact* being seen in the most frequent terms listed in Table 3. Shoes are often referred to as *items*, *goods* and *footwear* in this sub-corpus, which, when accompanied by the aforementioned terms and frequent references to other nouns such as *policy*, *forms* and *business*, invokes a strong sense of professionalism and rigour: clear references to the notion of the 'business' of online selling.

[2] The British National Corpus, BNC, is a 100 million word corpus of written and spoken discourse in English. For more information see: http://www.natcorp.ox.ac.uk/

On closer inspection, we see that many of the new listings include such terms in what appears to be almost e-signatures at the bottom of their listings: standardised information about fee payment and delivery information. An example of such follows (**emphasis added**):

> *We never wait for buyers to leave feedback first. As a **buyer** your obligation is only to **pay** nothing more. As a **seller** my obligation is to make sure that you are 100% satisfied. Please contact us before leaving negative or neutral feedback. There is no problem that we can't solve. We are committed 100% to providing outstanding **customer service** and want you to have the very best ebay experience when shopping in our **store**. 100% Satisfaction Guarantee: We guarantee that all of our items are 100% authentic and of the highest quality. If you are unhappy with your item for any reason you can **return** the item and we will reimburse you the **purchase price** minus **shipping cost** and less a 25% restocking **fee**. All returns must be unworn and in the original packaging with tags attached.*

This functions to offer the buyer some reassurance on the quality of the product on sale (with reference to it being *100% authentic* and offering a *100% satisfaction guarantee*, for example) and to the *outstanding customer service* offered by the buyer. It also states the obligations and roles of buyer and seller, acting almost as a written contract between the buyer and seller, one to which the buyer agrees should they bid on and win the listed item.

In comparison, we see that keywords utilised in the descriptions for the used items are generally more evaluative (in a positive way), with a significant use of adjectives referring to the general appearance and physical properties of these items. These include terms such as *good*, *excellent*, *very*, *great* and *lovely*. In fact, adjectives are used at a statistically higher rate in the used items data at a difference of log-likelihood -1057.73 compared to the new item data (with relative frequencies of 10.75 and 8.24 respectively). This is a very significant rate of difference. An example of this is seen in the following listing:

> *Here I am listing a **lovely** pair of sandals from Mini Boden Size European 30 or uk size 12. 100% leather Synthetic soles. Pastel shaded polka dot design. Three buckles that can be easily adjusted to size. These **delightful** sandals have been worn by my daughter but sadly she has outgrown them... hence this sale! Ideal for the spring and summer months ahead. From a smoke free home. Thankyou for looking.*

Here we see references to the appearance of the sandals, and the affective associations of such, with words including *lovely* and *delightful* featuring in the description. In addition to such terms, words relating to colour such as *black* and *pink*, along with references to the materials *Velcro*, *soles* and *leather* are words that are all cited as keywords in the used shoe descriptions in Table 3. So too are evaluative adjectives, words and clusters referring to the frequency of usage, with *few (times), used, worn, good (condition), wear, excellent (condition), plenty, great (condition), hardly, only* and *twice* all featuring as keywords in Table 3. This is complemented by direct references to the condition of the product and the place in which it has been stored, so with references to *smoke free* and *clean* as well as the physical condition of the shoes, with the *condition*, *worn*, *good*, *wear*, *used*, *excellent*, *hardly*, *great*, *plenty* and *few times* featuring in the top 50 expressions that are used at a statistically higher rate than in the used sub-corpus.

3.2. Experienced vs. novice sellers

In addition to the traditional facilities for corpus analyses such as searching, generating wordlists and KWIC (Key Word in Context) outputs, as offered by standard concordancing tools, WMatrix is equipped with a semantic tagger which automatically classifies individual words in a corpus according to thematic categories according to the semantic tags assigned to them (see Rayson, 2003). The tagset used in this tool is loosely based on the Longman Lexicon of Contemporary English (McArthur, 1981). For example, words such as *vote*, *political*, *Tory* and *election*, would be classified under the general thematic grouping of 'politics'. If we compare the most frequent thematic groupings of words (rather than individual lexical items, as seen in table 3) used in descriptions listed by novices (with 100 or fewer on their feedback score, denoted by 'u100') vs. experienced sellers (with 1000 or more on their feedback score, denoted by 'o1000') and experienced vs. novice sellers we see a similar language profile to the new vs. used descriptions outlined above (results taken from WMatrix):

	U100 vs. O1000	O1000 vs. U100
1	Clothes and personal belongings (e.g. worn, shoes, boots, sandals, footwear)	Putting, pulling, pushing, transporting (delivery, shipping, sent, dispatched)
2	Measurement: size (e.g. size, size 6, size 5, size 7, size 4, fit)	Time: period (days, hours, period, weekends, Monday, bank holidays)
3	Colour and colour patterns (e.g. blue, black, colour, pink, white, red)	Business: generally (business, office, ltd, agents, compant, companies)
4	Evaluation: good (e.g. good, great, well, fantastic, super, high quality, look great)	Pronouns (you, we, your, our, us, my)
5	General appearance and physical properties (e.g. condition, padded, bow)	In power (order, leading, control, boss, leading, manager, management)
6	People: Male (boys, boy, mens, men, man)	Telecommunications (email, spam, telephone, phone, helpline, callers)
7	Happy (happy, looking happy, happily, lol, delighted, joy)	Information technology and computing (email, online, messaging, website, pc)
8	Other proper names (Nike, Timberland, Adidas, Marks, Post Office)	Getting and giving; possession (exhange, exhanges, exchanged)
9	Seem (looking, look, show, seem, seems, looks)	Likely (can, may, make sure, would, guarantee, secure, sure, clearly)
10	Kin (son, daughter, sons, mum, daughters, husband, brother)	Entire; maximum (all, any, full, 100%, entire, every, complete, intact)

Table 4: Comparing the semantic categories of content listed by novice and experienced sellers.

As with the new product descriptions, there is an increase in words linked to professionalism and retail, with words relating to themes such as transportation and exchanging items, telecommunications and business frequently being used in the descriptions listed by more experienced sellers (refer to the right hand column of this Table). Again, this suggests that there is an increase in factual information in this sub-corpus, something that is perhaps characteristic of this type of transactional discourse (given the frequency of such). In comparison, listings from more novice sellers focus more on descriptions of the products themselves, in terms of their size, colour and general appearance. Correspondingly, we again

see a significant difference in the number and type of adjectives used in the descriptions by experienced and novice sellers.

As with the used data, references to colour, dimensions (i.e. size and measurement) and appearance (refer to 'seem' and 'general appearance and physical properties' in the left hand column) are more common with the novice sellers when compared with experienced sellers. This includes a frequent use of the clusters *smoke free*, *looking happy*, *look great*, *looked after*, *pet free*, *look fab*, *a little (worn)* and *wear and tear* in this data. Such terms also underline the importance that novice sellers attach to the description of the condition of products (namely because, on average, they are more likely than the experienced sellers to list used rather than new items).

Check out my other items having a	clear out!.	Payment via paypal only and with…
ning form Size 6 Having a massive	clear out	so please see my other items . Great..
Having a massive family	clear out	so please check out my other auctions.
ear left in them . Am having a huge	clear out!.	Stored in non-smoking household..
s message ill reply asap Having a	clear out	so take a look at my other items a..
her items as i am having a little	clear out.	Pay pal only please and i always…
See other listings. Having a huge	clear out.	All great items. Please don't hesitate to...
r items having a big spring clean	clear out.	These cica trainers are a size 4
best shoes worn only once having a	clear out	I paid 70 please look at other items…
real bargain. I am having a major	clear out	and have beautiful items to sell p..

Figure 1: Concordance output for *clear out*.

Many of these clusters create an impression that listed items are highly desirable but no longer needed, have been grown out of, or are simply being sold as part of a *clear out* or *spring clean* (clusters which are used 47 and 5 times in the novice sub-corpus and only 11 and 3 in the experienced sub-corpus). Examples of clear out are seen in Figure 1. Here we see that there is an emphasis on the scale of the *clear out*, with adjectives relating to size, such as *major*, *huge*, *big* and *massive* appearing as common collocates to this cluster, words which perhaps entice the buyer to browse other items listed by the seller, or indeed to bid on particularly items in the knowledge that such sales are infrequent, on-chance-only sales so something that the buyer should invest in before someone else beats them to it.

Based on this analysis, it seems that experienced users sell more than novice users and use a discourse which is more oriented to professionalism and less personal. Novice users, on the other hand, tend to sell less than experienced users and adopt a discourse which is closely related to personal identity and relationships. The frequent use of such informal multi-word expressions may be an attempt to create a sense of closeness, a mutual understanding and sense of rapport between the seller and the buyer, a human identity behind the listing, in the sense that the seller is expressing opinions about and affective associations with the products they are listing. The use of this language can, therefore, be described as less clinical and professional, but more interactional than the language of more experienced counterparts.

This notion of rapport and interactional, human-centred discourse is further supported by an increase in the use of words related to 'kin' in the novice data (refer to Table 4). Words referring to personal family relationships such as *son*, *daughter*, *husband*, *mum* and so on are shown to be far more frequent in the descriptions listed by these users. This is also true of those listing used items, as opposed to those listing new items. Words related to this notion of kin and their (relative) frequencies are tabulated in Table 5:

	u100n	u100u	o1000n	o1000u		u100n	u100u	o1000n	o1000u
bridal	0	0	1	0	momma	0	0	1	0
bridesmaid(s)	4	2	5	4	mother	0	1	0	1
brother(s)	4	2	1	0	mother-in-law	0	0	0	1
cousin	1	1	0	0	mum	0	13	0	10
dad(s)	1	1	0	0	mummy	1	2	0	0
daughter(s)	14	56	7	39	nanna	0	1	0	0
family	1	3	1	1	niece	1	1	0	0
father	0	0	1	1	nephew	1	0	0	0
gran	0	0	0	2	offspring	0	1	0	0
granddaughter	0	3	0	0	parent(s)	0	2	0	0
grandkids	0	0	1	0	sister(s)	1	2	0	1
grandson	0	2	0	0	son(s)	40	194	0	1
household	3	14	2	0	triplet	0	0	0	1
hubbies	0	0	1	0	twin	1	2	0	0
husband(s)	6	4	0	1	wedding(s)	14	44	17	5
ma	0	2	0	0	wife	0	3	0	0
						93	356	38	68

Table 5: References to kin in the shoe sub-corpus.

In this table we see that the least experienced sellers listing used items utilise a wider range of terms relating to kin, at a significantly higher frequency than the other sellers examined (with *son(s)* being the most frequently used term at a rate of 194 occurrences). Those who are most experienced and listing new items use these expressions far less frequently than the other sellers. This use of terms relating to kin may reflect an attempt on the part of the seller to relate to the buyer, with the use of mutually understood and personally applicable referents.

3.3. Patterns across new and used listings for experienced and novice sellers

3.3.1. Pronoun usage

It is interesting to note that Tables 3 and 4 also revealed a disparity in pronoun use across descriptions of new and used shoes. This is something that is also seen in those descriptions listed by experienced and novice sellers. Regarding Table 3, the top five frequently-used words in the new data, *we*, *your*, *you*, *our* and *us*, are all examples of this word class. These pronouns were found to be more frequent in new listings targeted at all ages and genders, i.e. products aimed at men, women, boys and girls.

On closer inspection it is interesting to note that, in contrast, while these particular pronouns are not frequently utilised in the used listings, other forms of pronouns *are* frequent, although these are used at a less frequent rate, and are of a more limited variety to those seen in the new shoes data. These forms are listed in Table 6.

Item	Used freq.	Rel. freq.	New freq.	Rel. freq.	LL
I	3322	1.2	1592	0.43	(+) 1224.70
my	2494	0.9	1026	0.28	(+) 1128.97
me	7429	0.41	555	0.15	(+) 411.41

Table 6: The use of first person personal pronouns in the 'used' and 'new' product descriptions.

Table 6 indicates that while the new descriptions were shown to use third person pronouns at a higher rate than the used data, the latter shows frequent use of first person personal pronouns, with *I*, *my* and *me* being the most common forms used in this sub-corpus.

Similarly, Table 4 also indicated that there was, again, a higher number of pronouns in the more experienced sellers' descriptions, compared to the novices and again these tended to be third person pronouns, rather than first person personal pronouns, which were conversely more frequently used by the novices. Experienced sellers who were listing new items utilised pronouns in their product descriptions at a more significant rate of difference when compared to novice sellers listing used items than any other type of seller (i.e. an experienced seller listing used items or a novice seller listing new items).

The use of personal pronouns is typically seen as being a characteristic of less formal, spoken, discourse (see Chafe and Danielewicz, 1987; Biber, 1992; Biber et al., 1999; Heylighen and Dewaele, 2003; Carter and McCarthy, 2006; Atkins, 2011 and Knight et al., 2013, 2014). Modes of discourse that frequently use personal pronouns are considered to be characteristic of a form "*of communication that allows for an immediate or near-immediate information exchange, a forum for communicating reports of events and incidents in near real-time, as the understanding of the temporal referent is shared*" (Knight et al., 2014). They "*mark a relatively low informational load, lesser precision in referential identification or a less formal style"* (Biber, 1988) and indicate an ego involvement in the discourse process (Ko, 1996). So, although online discourse is asynchronous, the frequent use of personal pronouns often functions "*as a means of establishing and reconfirming a shared 'digital space' between senders and recipients*" (Knight et al., 2013 - for more detailed discussion of synchronicity see Condon and Cech 1996; Ko 1996 and Herring, 2007), in a similar way to what we have already seen with the use of references to kin discussed above.

Although those listing new items use pronouns, on the whole, more frequently than do those selling used items, the fact that these are more frequently third rather than first person personal pronouns serves to objectify the content rather than personalise it. This creates a certain level of distance between the seller and buyer, as the descriptions appear to be directed at a more general readership, rather than an individual buyer. This is particularly true of instances where more than one unit of the described product is available for purchase (something particularly indicative of sellers with particularly high feedback scores listing multiple items).

3.3.2. Politeness

In addition to pronouns, the analysis also revealed a marked difference in the rate at which 'politeness markers', as classified by the WMatrix semantic tagger (which is based on the UCREL[3] semantic analysis system - see Wilson and Rayson, 1993), are used in the data, as presented in Table 7:

	u100n	u100u	o1000n	o1000u
thank(s)	114	412	67	119
thankyou	4	35	12	9
compliment(s)	1	2	1	1
complimentary	0	1	0	0
tact	0	0	1	0
	119	**450**	**81**	**129**

[3] The University Centre for Computer Corpus Research on Language, Lancaster University.

Table 7: The use of politeness markers in the shoes data.

This table suggests that common markers of politeness such as *thanks*, *thank you* and *compliments* and so on are also markedly more frequent in the used shoe data rather than the listings for new shoes, and are particularly frequent in used item listings posted by novice sellers. *Thank(s)*, the most common of these, is used 412 times in u100u compared to 67 times in o1000n (with a ll. of +492.17). A comparison of 1 million words of spoken and written data taken from the BNC reveals that the rate of use of *thank(s)* in the u100u, o1000u, u100n, data is significantly higher than in both spoken and written discourse (all with ll. significantly higher than +6.63), with the most significant difference seen with u100u, then o1000u and so on (i.e. with used items listed by novices utilising the highest amount of politeness markers and new items listed by experienced sellers utilising the lowest amount).

While a study by Herring suggested that "*public CMD [computer mediated discourse] tends to be less polite than private CMD*" (Herring 2003: 19), recent work by Knight et al. (2013, 2014) found that levels of politeness in online discourse (CMD) often increased in specific forms of public CMD, especially when a large readership is likely. This is particularly true for forms of CMD where the maintenance of face and positive politeness are critical ingredients for maximising the number of people that will follow your online profile, for example or indeed, as in this case, where buyers are persuaded to invest in particular descriptions and favour them over others, so as to purchase the items listed. While the terms of service of the site may also influence the degree of politeness adopted, this was not the focus of the current study and so has largely been ignored.

3.3.3. Modality

When exploring the use of modal verbs in the corpus we see that the frequency patterns observed in the use of pronouns are completely reversed. *Must*, *can*, *may* and *will* were all identified as keywords for the descriptions of new products in Table 3 (these terms are highlighted for ease of reference) and when we drill deeper into the data it is apparent that they are indeed most frequently used by experienced sellers listing new items and least frequently used by novice sellers listing used items. The rates of use of the most common modal verbs are tabulated in Table 8 (with 'freq.' denoting the raw frequency of use and 'Rel.' denoting the relative frequency of use).

	u100n		u100u		o1000n		o1000u	
	Freq.	Rel.	Freq	Rel.	Freq.	Rel.	Freq.	Rel.
may	85	0.12	44	0.04	701	0.34	40	0.08
might	5	0.01	8	0.01	15	0.01	5	0.01
can	353	0.5	335	0.28	1047	0.51	171	0.35
could	20	0.03	32	0.03	21	0.01	12	0.02
must	83	0.12	45	0.04	548	0.27	41	0.08
should	87	0.12	48	0.04	110	0.05	36	0.07
will	670	0.95	632	0.53	2180	1.06	421	0.87
would	43	0.06	153	0.13	125	0.06	40	0.08
shall	0	0	5	0	5	0	2	0
	1346	**1.91**	**1302**	**1.1**	**4752**	**2.31**	**768**	**1.56**

***Table 8*:** Frequency of modal verb usage across the corpus.

In a study by Leech et al. (2001) it was suggested that modal verbs are, on the whole, more frequent in spoken than written English. In their own study of data from the BNC corpus they

discovered that modals accounted for 19,543 words per million of the spoken data and only 13,635 per million words of the written data. They also identified individual differences in the types of modal verbs used across the modalities, with, for example, *can* being more frequent in the spoken data and *may* being more frequent in the written.

Modal verbs are used to express a variety of different meanings, including possibility/permission, necessity/obligation and prediction/violation (Friginal, 2009: 150) and their rate of use, accordingly, is somewhat context bound and highly dependent on the aims and objectives of the discourse. For example, a study of the language of call centres carried out by Friginal (2009: 154) revealed that while *must* was a modal verb typically used at a statistically lower rate in the conversations analysed, when talk shifted to topics concerning account policies or service contracts, the rate of use increased significantly. Similarly, Biber (2006) found that in a university context, such necessity/obligation modal verbs are typically infrequent in spoken discourse, but commonly used in written documents, especially when the topic of the writing shifts to focus on rules and policies, in official documents, rules and regulations, for example.

It is interesting to note that the most experienced sellers listing new items use a considerably higher number of modal verbs in their product descriptions than with any other grouping (with a relative frequency of 2.31 per 100 words), this is followed by novice sellers listing new products. The fewest modal verbs are used by novice sellers listing used products (relative frequency of 1.1), which suggests, in short, that modal verbs are more frequently used in new product listings on eBay. In this table we can see that the most common modal used in the corpus is *will,* which marks a prediction/violation, followed by *can*, which typically marks possibility, permission and ability.

If we explore the most frequent collocates (i.e. words that regularly co-occur in discourse) of the most frequent modal verbs in this corpus, *can* (Table 9) and *may* (Table 10), within the o1000n data specifically, we see that a similar pattern in language use emerges. In these tables the search terms are depicted in the middle column of the table and the most frequent collocates are listed in the specific position in the span of the word (i.e. its immediate co-textual environment) at which they most frequently occur. For example, in Table 9 we see that *we* is most commonly positioned to the left of *can*, at L1, so one word prior to *can*. While *be* is the most frequent collocate at one position to the right of *can* (R1), suggesting that the most frequent cluster in this data is *we can be*. In both of these Tables the top 20 most frequent words in each position to the left and right of the search term are presented.

N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	THIS	THIS	SO	THAT	WE	CAN	BE	YOUR	TO	THIS	SATISFACTION
2	REFER	OUR	RETURN	POLICY	YOU		BID	FLEXIBLE	ON	BECAUSE	WE
3	WENT	METHOD	CREDIT	CARDS	I		ACCEPT	WITH	CONFIDENCE	ONLY	THROUGH
4	WITHIN	WRONG	AND	SO	PAYMENT		RESOLVE	IT	CARDS	YOUR	A
5	BID	ON	YOU	ITEMS	THEY		ONLY	CREDIT	FOR	IN	WILL
6	ADOPTED	AS	POLICY	AS	2011		DISPATCH	YOU	ITEMS	YOU	PROVIDING
7	LEAVING	EUROPEAN	QUICKLY	ALTHOUGH	OR		PURCHASE	US	ORDER	A	NOTE
8	PAYPAL	FEEDBACK	METHOD	OF	DELAYS		SATISFACTION	GUARANTEED	A	ON	LISTED
9	GUARANTEED	7	MULTIPLE	AUCTION	ITEMS		DELAY	SHIP	BID	THE	ARRANGEMENT
10	RESPOND	PURCHASING	THIS	CHARGE	AND		I	RESOLVED	NEGATIVE	PLEASE	50
11	CUSTOMERS	ABOUT	DAYS	HOW	HOW		RETURN	OR	100	FEEDBACK	REFUND
12	COLLECT	AND	DELIVERY	WITH	STYLES		GET	THE	IF	COUNTRIES	THE
13	ARE	ITEMS	HAPPY	HOPE	THINGS		MONITOR	AND	IMPROVE	BUYER	NEED
14	MY	PURPOSES	IT	DECEMBER	WHAT		INSTRUCT	RETURNED	IT	TO	PREFER
15	ANALYSIS	NOT	23RD	IMMEDIATELY	DELIVERY		OFFER	A	WHAT	IMMEDIATELY	SERVICES
16	OCTOBER	TO	YOUR	NEED	FIT		ALSO	THEM	UP	ANY	31ST
17	YOU	FOR	THE	IF	PROCEDURE		NOT	FOR	BY	UP	OR
18	PAY	FROM	OR	PUMA	REFUNDS		CONTACT	OFF	THE	OUR	TO
19	HAVE	WON	ALL	CARD	Q		HOLD	PAY	OR	WE	OF
20	DAYS	WHAT	THAT	SOMETIMES	REFUND		SEE	FIND	MY	VERY	FROM

Table 9: Common collocates of *can* in the shoes sub-corpus.

N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	OUR	OR	ANY	YOUR	YOU	MAY	BE	THE	TO	OF	YOUR
2	AND	OF	OF	QUESTIONS	THAT		HAVE	LONGER	YOU	3	HAS
3	YOU	FOR	DAYS	PROBLEMS	WE		ALSO	BEFORE	ONE	IT	ITEMS
4	OUT	ASK	NATIONAL	HOLIDAYS	CARRIER		TAKES	CONTINUE	PURCHASING	ADDITIONAL	WORKING
5	PAY	WORKING	TO	SHIPPING	IT		TAKE	1	SINCE	PLACE	FOLLOWING
6	TWO	ALL	YOUR	PURCHASE	PURCHASES		DELAY	CONSOLIDATED	SHIPPING	ONE	INVOICE
7	WEEKENDS	WITHIN	ALL	HRS	CHEQUES		SEND	BECAUSE	UNDER	TO	AS
8	REPLIED	RESOLVE	24	QUERIES	THERE		PROVIDE	A	WILL	RESPOND	AND
9	TO	IT	SERVICES	YOU	CONTRACTORS		CONTACT	WE	RETURN	FEE	FOR
10	RESEND	SERVICES	AND	SHOES	DETAILS		USE	YOUR	INTEREST	FOR	OUR
11	WITH	AGENTS	THESE	SUB	FEEL		APPLY	REQUIRED	BY	THE	ADVISERS
12	FROM	ANOTHER	BUSINESS	WE	CHECKS		WIN	YOU	SHARE	POST	YOU
13	PRODUCTS	TIME	IN	TIME	SURCHARGES		BECOME	DISCLOSED	INFORMATION	OUR	INFORMATION
14	QUALITY	CUSTOMER	THAT	ADDITION	BOXES		OFFER	AND	BACK	SIGN	E
15	CUSTOM	ON	OTHER	DEPARTMENT	LOOKING		NOT	OF	YOUR	IN	REFUND
16	SHIPPING	CHARGES	UK	DELIVERY				THEM	THIS	TIME	COUNTRY
17	WE	APPRECIATE	5	ITEMS				TO	POSTAGE	DELIVERY	PERIOD
18	AN	BOX	THANKS	SHOE				WITHIN	DAMAGED	ON	TRANSIT
19	5KG	EXTRA	THE	0				SLIGHTLY	FOR	CUSTOMS	SOME
20	WITHOUT	ANY		DELIVERIES				FREE	ODD	MARK	FEES

Table 10: Common collocates of *may* in the shoes sub-corpus.

In Tables 9 and 10 we see the frequent use of terms relating to future events, confirmation or clarification of products, services and procedures relating to the process of purchasing and receiving the products. These include references to *bids*, *procedures*, *policy*, *purchasing* and *payment*, *credit cards*, *cheques*, *surcharges*, *shipping*, *business*, *agents*, *contractors*, *services* and *customs* (italicised terms are listed as frequent collocates in Tables 9 and 10). The utility of these modal verbs functions to provide buyers with a reassurance of the procedural aspects of investing in certain products, some form of confirmation, peace of mind and assurance that they know specifics about the product and how the order/purchased item will be handled by the seller. The *buyer/ customer* is encouraged to ask *questions* and *queries* about products, they are offered a *guarantee* of a *response, details* or some form of *information* and are provided with guidelines on how to *return* items if they are not *happy* or *satisfied* or if something is *wrong* with the *quality* of what they receive. Terms such as these again strengthen the notion of professionalism and help to create a profile of the business side of online selling.

4. Discussion

The analysis has outlined various ways in which the language of eBay descriptions (as evidenced by the UK site specifically) is characterised depending on whether they are listed by experienced or novice eBay sellers and whether items are new or used. A summary of these patterns follows:

- Words relating to the business of online selling, professionalism and retail, transportation and item exchange and telecommunications are used at significantly higher rates in the new product descriptions vs. used descriptions and by experienced sellers rather than novices.
- Words relating to frequency of usage, colour and appearance are more frequently used in used product descriptions vs. new product descriptions and by novice rather than experienced sellers.
- The novices utilise a larger number of terms relating to kin in their description (specifically those listing used items) than more experienced sellers.
- New items listed by experienced sellers feature a larger number of pronouns in their descriptions, although the majority of these are third person pronouns.
- Used items listed by novices feature a significantly greater number of first person personal pronouns in their descriptions than new items listed by experienced sellers.

- A greater number of politeness markers are included in descriptions for used vs. new items (both with novice and experienced sellers), with the highest amount seen in the used items listed by novices.
- Modal verbs are used at a significantly higher rate in the new product listings vs. used products. This is particularly true for those listed by experienced sellers.

The identities of the experienced and novice sellers appear to be significantly different in the data. This sense of identity is formulated, to a large extent, by the specific types of linguistic features used in their listings. According to Benwell and Stokoe (2006), identities are constructed in specific contexts and vary within that context according to the goals of the participants. In our data, the context of the online marketplace highlights the ways in which identities vary according to whether a seller is experienced or not and whether the product is new or old. Other factors such as age, social class, social distance and mood appear less relevant, partly because of the absence of an audio-visual context, which means that 'normal' discourse features such as accent, sound of voice, pausing, hesitation, formality, coherence and so on are absent, whereas information about the product and the seller are key to the interaction. In the eBay data, identities are co-constructed 'in the moment' through transactions, according to what is being sold and who is doing the selling. The notion of a 'virtual identity', so often used to describe identities created in cyberspace, is less important than the fact that people's identities are closely related to what is being sold. These 'transactional identities' are established through the specific linguistic and politeness features highlighted in the preceding discussion.

5. Conclusion

Transactional discourse is typically characterised as utilising specific lexical items which help to provide an "*optimally efficient transmission of information*" in its exchanges (Brown and Yule, 1983; Lakoff, 1989). Transactional discourse is commonly used in commercial contexts to convey factual information often as a means of persuading third parties to 'buy-in' to what is being communicated and indeed, often, actually to purchase products. Interactional discourse, in contrast, "*has as its primary goal the establishment and maintenance of social relationships"* (Kasper, 1995: 205). While language is never solely one variety or the other, it is typical for one type of goal to be present.

The study reported in this article set out to characterise specific patterns of language used in digitally based online auction sites, the 'e-marketplace'. Using a corpus-based approach to data collection and analysis, we characterized the language used in product descriptions listed on eBay and highlighted the ways in which the goal of selling a product through this medium creates its own discourse and constructs particular identities. Two key factors – seller experience and product condition – were found to exhibit important differences in both language use and seller identity. The paper has relevance for both the understandings of the key defining features of online discourse and for enhancing methodologies for studying this text type.

Future research might extend the current study by, for example, examining the levels of 'success' of sellers (i.e. rated by the number of buyers bidding for particular items and the prices at which they are sold), or by comparing patterns of language used by experienced and novice sellers listing other types of new and used items (such as clothes, cars, IT equipment and so on). There is also scope for research which compares patterns of language used in listings targeted at male and female buyers and/or children and adults. Other studies might look at the differences in language used across different varieties of English eBay sites across the world. For example, how do sellers who list items on eBay in the USA or Australia describe items in similar and/or different ways to the UK based site examined in the present

article? In short, the potential for research focused on an online marketplace, such as that of eBay, is enormous and likely to result in considerable advances in the study of online discourse, one of the fastest growing areas of inquiry in contemporary applied linguistics.

6. References

Atkins, S. 2011. *A Cognitive Linguistic Perspective on Social Space in Online Health Communities.* Unpublished PhD Thesis. The University of Nottingham.

Knight, D., Adolphs, S. and Carter, R. (2013). Formality in digital discourse – A study of hedging in CANELC. In Romero-Trillo, J. (Ed.) *Yearbook of Corpus Linguistics and Pragmatics.* London: Springer Verlag.

Knight, D., Adolphs, S. and Carter, R. (2014). CANELC – constructing an e-language corpus. *Corpora journal 9*(1): 29-56.

Benwell, B. and Stokoe, E. (2006). *Discourse and Identity.* Edinburgh, Edinburgh University Press.

Biber, D. 1988. *Variation across speech and writing. Cambridge: Cambridge University Press.*

Biber, D. 1992. "On the complexity of discourse complexity: A multidimensional analysis". *Discourse Processes 15:*133-163.

Biber**,** D**.** 2006. University language: A corpus-based study of spoken and written registers. Amsterdam: John Benjamins.

Biber, D. Conrad, S. Leech, G. Svartvik, J. & Finegan, E. 1999. *The Longman Grammar of Spoken and Written English.* Harlow: Longman.

Brown, G. & Yule, G. 1983. *Discourse Analysis*. Cambridge: Cambridge University Press.

Carter, R.A. & McCarthy, M.J. 2006. *Cambridge Grammar of English.* Cambridge: Cambridge University Press.

Chafe, W.L. & Danielewicz, J. 1987. "Properties of spoken and written language". In Horowitz, R. & Samuels, S.J. (Eds.) *Comprehending oral and written language.* New York: Academic Press. pp.83-113.

Condon, S.L. & Cech, C.G. 2001. "Profiling turns in interaction". In *proceedings of the 34th Annual Conference of the Hawaii International Conference on System Sciences*. Los Alamitos, CA: IEEE Computer Society Press.

Friginal, E. 2009. *The language of outsourced call centres: A corpus-based study of cross-cultural interaction*. London: John Benjamins Publishing Company.

Held, D. 1995. *Introduction to Critical Theory: Horkheimer to Habermas.* Cambridge: Polity Press.

Herring, S.C. 2007. "A faceted classification scheme for computer-mediated discourse". *Language@Internet 4*(1): 1-37.

Heylighen, F. & Dewaele, J. -M. 2003. "Variation in the contextuality of language: an empirical measure". *Foundations of Science* 7: 293–340.

Kasper, G. 1995. Wessen Pragmatik? Für eine Neubestimmung sprachlicher Handlungskompetenz. *Zeitschrift für Fremdsprachenforschung* 6: pp. 1–25.

Ko, K. 1996. Structural characteristics of computer-mediated language: A comparative analysis of InterChange discourse. *Electronic Journal of Communication 6*(3).

Lakoff, R. 1989. The limits of politeness. *Multilingua* 8 (2/3): 101–129.

Leech, G., Rayson, P. & Wilson, A. 2001. *Word Frequencies in Written and Spoken English: based on the British National Corpus.* London: Longman.

McArthur, T. 1981. *Longman Lexicon of Contemporary English*. London: Longman.

Nattinger, J. & DeCarrico, J. 1992. *Lexical phrases and language teaching.* Oxford: Oxford University Press.

Rayson, P. 2003. *Matrix: A Statistical Method and Software Tool for Linguistic Analysis Through Corpus Comparison* . Unpublished PhD thesis. Lancaster University.

Scott, M. 1999. *Wordsmith Tools* [Computer program]. Oxford: Oxford University Press.

Sinclair, J. 2008. Borrowed ideas. In Gerbig, A. and Mason, O. (Eds.) *Language, People, Numbers- Corpus Linguistics and Society*. Amsterdam: Rodopi BV, pp. 21–42.

Watzlawick, P., Beavin, J. & Jackson, D. 1967. *Pragmatics of Human Communication*. W. W. Norton: New York.

Wilson, A. and P. Rayson. 1993. 'Automatic content analysis of spoken discourse' in C. Souter and E. Atwell (eds) Corpus-based Computational Linguistics, pp. 215–26. Amsterdam: Rodopi.