International Journal of English Linguistics; Vol. 2, No. 4; 2012
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Noun Phrase Cohesion in English Discourse: a Corpus-based Analysis of Patterns and Influences

Clarence Green[1]

[1] School of Languages and Linguistics, University of Melbourne, Melbourne, Australia

Correspondence: Clarence Green, Bld. 139 Parkville, Melbourne 3010, VIC, Australia. Tel: 61-3-8344-8032. E-mail: c.green4@pgrad.unimelb.edu.au

Received: May 28, 2012 Accepted: June 8, 2012 Online Published: July 10, 2012

doi:10.5539/ijel.v2n4p44 URL: http://dx.doi.org/10.5539/ijel.v2n4p44

Abstract

This study investigated the patterns of noun phrase cohesion in English discourse, with a specific interest in the role of the different clause types. Using corpus methodology, the study synthesized into a single framework for analysis the central features of previous research regarding cohesive devices, preferred argument structure, genre, information packaging and clause structure. A corpus of 1206 noun phrases was coded for factors drawn from previous research, starting with whether or not the NP contained old/cohesive information. Results of frequency cross-tabulations and a factor analysis indicated that preferred argument structure, noun phrase form, and discourse genre were more significant influences on patterns of noun phrase cohesion in English discourse than clause structure. However, patterns of cohesive noun phrases according to the distance to their antecedents revealed that the more grammatically integrated clauses, such embedded infinitival clauses, the fewer cohesive noun phrases with antecedents in the immediate context they tended to have. This indicated that at the local inter-clausal level noun phrase cohesion and the level of grammatical integration of a combined clause existed in somewhat complementary distribution. Conclusions drawn included that clause grammar codes cohesion locally, displacing the need for noun phrases to mark cohesion in the immediate discourse. The study therefore quantitatively supports previous theories that discourse cohesion and the different types of combined clauses in English exist along a continuum from grammar to discourse.

Keywords: cohesion, coherence, corpus linguistics, English grammar and discourse

1. Introduction

A feature of fundamental importance in the successful use of language is maintaining coherence amongst elements in discourse during communication (Beauchamp & Dressler, 1984). Research has shown that the different ways that this is accomplished marks different stages of development in both L1 and L2 language development (Crossley & McNamara, 2009). Consequently, developing a better understanding of the patterns in English of discourse cohesion has significant value to English linguistics. Discourse cohesion, being a complex phenomenon, has been approached in previous research from a wide range of different perspectives including cognitive, applied, theoretical and computational linguistics. This has led to independent research traditions, each with a specific focus on a particular aspect of cohesion. These include the discourse analysis research into preferred argument structure (Dubois, 2003), those which use NLP and computational tools for coherence studies (McNamara et al., 2006), genre analyses studies (Biber, 1988), investigations into the role of clause structure in coherence (Givon, 1998; 2001), and the text linguistics approach of Halliday and Hasan (1976) in their seminal descriptive work Cohesion in English (1976). These valuable research traditions have developed so independently that there exists an unfortunate disconnection between them. The purpose of the current study was therefore to use a corpus methodology to bridge this disconnection and simultaneously analyze a range of features of importance in the patterns of English noun phrase cohesion at the discourse level. A central concern was to determine the role of clause structure in the patterns of English discourse cohesion. Previous research has suggested that the range of clause types differently relate to cohesion, but this hypothesis has not yet been quantitatively tested.

2. Cohesion in English

The current study uses a framework based primarily on the categories of cohesion described by Halliday and Hasan (1976). Their work founded the concept of cohesive devices, and remains the most in-depth description of English cohesion within text linguistics. They described two overarching categories that all noun phrase cohesion can be placed into depending on the form of the NP: grammatical and lexical. Grammatical and lexical NP forms are cohesive devices when they reference another grammatical or lexical unit in the discourse. This cohesion may be achieved by a pronoun, a repeated lexical item, a synonym, a hyponym, a collocation and so on. In Example 1 the underlined words illustrate grammatical cohesion, specifically a pronominal reference chain (Arnold, 2007). The writer in this example, rather than repeating the antecedent lexical item "branches" when developing the discourse has substituted it for the pronoun "they" in the later relative clause, and again in the subject of the following sentence. The words in italics in Example 1 illustrate a chain of lexical cohesion, with the lexeme "members" used (within this discourse context) as a synonym for the "candidates" in the first sentence.

Example 1. Grammatical and lexical noun phrase cohesion

> ACE sample> B19 <source> Northern Territory News/Daily News
>
> Branches consider primarily *candidates* that they believe will win. They also take into account qualities that might enable *members* to achieve high office.

By using the classification of noun phrases into lexical and grammatical, Halliday and Hasan (1976) provide a means of quantifying cohesion in a discourse, one that is amenable to a corpus methodology. However, their research tradition has not been overly interested in determining influences on the distribution of the patterns of cohesive devices. This study, however, was interested in why cohesion across the NP constituents in Example 1 was not maintained as follows: "Branches consider primarily *candidates* that branches believe will win. Branches also take into account qualities that might enable *candidates* to achieve high office". The fact that the data did not pattern like this through simple repetition to create cohesion, as noted by Arnold and Griffin (2007), suggests that there must be a range of discourse factors operating on noun phrases in English that affect the pattern of discourse cohesion. To explore these factors, this study drew on other research traditions and considered them within the Halliday and Hasan (1976) framework.

2.1 Preferred Argument Structure

One of the factors that very likely interacts with Halliday and Hasan's (1976) description of cohesive devices in English is known as preferred argument structure in discourse. This has been shown to exist in the discourse patterns of a range of the world's languages, and may quite possibly be a universal of discourse organization in human language (DuBois, 2003). It is preference in discourse organization that consists of two related aspects: 1. Reduced forms, such as pronouns, are generally preferred in the agent role (e.g. transitive subjects); 2. This syntactic role also favours old information rather than the introduction of new information. Clearly this can be drawn together with Halliday and Hasan's (1976) formal categories of cohesion. Grammatical NP forms are reduced forms, and an old information noun phrase means it must create discourse cohesion with an earlier element in the discourse. Indeed, the previous data of Example 1 which illustrated Halliday and Hasan's (1976) cohesive devices, also illustrates preferred argument structure:

Example 1 (b). Preferred argument structure and grammatical and lexical cohesion

> ACE sample> B19 <source> Northern Territory News/Daily News
>
> Branches consider primarily candidates that they believe will win. They also take into account qualities that might enable members to achieve high office.

The subject of the second sentence "they" is in the argument position of transitive subject. Following the tendency identified in preferred argument structure research, it is reduced to a pronoun. It is also old information as this subject is a referential NP with the lexical antecedent "branches" in the previous sentence. This makes it part of the grammatical category of cohesion in the Halliday and Hasan framework (1976). The two features of reduction and information salience are not always collinear however, as "they" had the possibility of being written as "branches" in the second sentence. Nor is the relationship between reduction, cohesion and subject categorical. This is demonstrated by "branches" in the first sentence, which is lexical, a subject and new information. Dubois (2003) has shown that the tendencies are cross-linguistically "soft constraints" on discourse, not categorical rules. This study aimed to quantitatively explore how these soft constraints pattern together with Halliday and Hasan's (1976) lexical and grammatical categories of cohesion in English discourse.

2.2 Clause Structure and Cohesion

Previous research has described a relationship between clause combination and cohesion in which the different clause types code different coherence relations (Givon, 2001). For example, subordination syntax codes more integration between clauses, i.e. tighter inter-clausal cohesion, than coordination syntax. However, the relationship between coherence and the combined clause grammar of English has not been explored quantitatively, nor has it been considered in relation to other factors that shape discourse cohesion. The hypothesis of a relationship between cohesion and clause structure is actually quite significant and controversial for theoretical linguistics. One strand of grammaticalization research, in contrast to UG, argues that the nature and origin of grammar is to provide a coding system for coherence relations amongst items that were at previous stages of the language related only at the discourse level (Givon, 1979; 2001). These theoretical questions need not concern the current research too much. What is important is that this research has proposes that combined clauses in English exhibit different levels of integration which reflect how tightly cohesive they are with surrounding clauses and discourse. If so, this has descriptive and applied value in understanding English grammar and discourse, regardless of theoretical implications. For example, Mathessien (2003) concluded that the types of English combined clauses and the system of cohesion in English discourse exist along a hierarchy as different manifestations of the same underlying phenomenon, as shown in Figure 1.

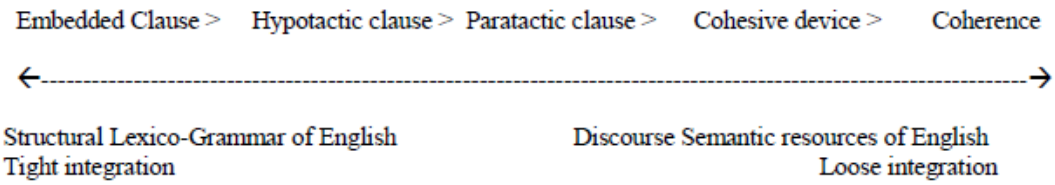

Figure 1. Hierarchy of clause combination to coherence (Mathessien, 2003)

Each of the clause types of English are more or less grammatically integrated, according to Mathessien (2003). Embedded clauses such as infinitival and complement clauses are the most tightly integrated clauses, while hypotactic clauses such as adverbial clauses are slightly less integrated. Paratactic clauses, which are essentially coordinate clauses, are the loosest type of English clause as they are easily separated into individual sentences. In the more integrated clauses, one of the major functions of grammar is to code the relationships between words. However, as one moves down the clause type hierarchy, English begins to use the discourse semantic resources of the language to indicate relationships between words. This is where Halliday and Hasan's (1976) cohesive devices begin to be used, such as lexical repetition, ellipsis, pronouns and so on, in order to explicitly mark the relationships between English words. Beyond cohesive devices, coherence based on pragmatic inference or illocutionary force is used to create meaning amongst discourse elements.

Cohesive devices such as repeated items or pronouns are not mutually exclusive with clause structure (Halliday & Hasan, 1976). For example, a repeated word or a pronoun in a subordinate clause can certainly have its antecedent within the same combined clause. However, if cohesive devices and the grammar of English clauses are part of the same hierarchical system as indicated in Mathessien (2003), the current researcher believed one should be able to expect some quantifiable evidence of such, and that evidence could be derived from a corpus analysis. Clause structure should affect the distribution and patterns of cohesive noun phrases in English discourse. Specifically, one would expect evidence of a pattern in English discourse cohesion supporting that complementary distribution exists between the level of grammatical integration of the different clause types of English and the number of cohesive devices they tended to contain. This is because if the hierarchy of clause type to coherence is valid, cohesion would be carried by grammar in the more integrated combined clause types of English, thereby displacing the need to mark cohesion through cohesive devices in noun phrases.

2.3 Genre and Other Influences on English Noun Phrase Cohesion

Previous research has established a range of other influences on English noun phrase cohesion that can be examined simultaneously as factors in a corpus study. Chafe (1984) and Givon (1990) have shown that distance from a cohesive noun phrase to its antecedent, as measured by the number of intervening clauses, impacts the discourse patterns of lexical and grammatical NP forms. Their work has indicated that pronouns tend to be favoured rather than lexical repetition of an NP when the antecedent of the NP is within 2.5 clauses of the cohesive noun phrase. Recently, Mizapour and Ahmer (2011) have shown that with respect to lexical cohesion, a high use of lexical NP repetition is a marked feature of the academic writing genre, as measured by data from journal articles. Similarly, Biber (1988) and Collins and Hollo (2009) have reviewed features of stylistic

variation and concluded that genre is an important influence on the discourse patterns of lexical NP's versus pronominal NP's. For example, writing styles associated with the legal genre heavily favour lexical repetition, to the point of redundancy, compared to grammatical cohesion. This is in order to avoid any possible ambiguity that might undermine legal agreements or decisions.

3. Scope of the Study

The current research had two specific goals, one was to bring together disconnected research traditions on discourse cohesion and consider them in a single corpus study; the other was to explore any relationship between clause structure and the patterns of noun phrase cohesion in English discourse. A corpus of noun phrases was therefore developed and coded based on factors such as clause type, preferred argument structure, information salience, NP form, genre, and Halliday and Hasan's (1976) description of lexical and grammatical cohesion in English. Patterns in noun phrase discourse cohesion were established through a statistical cross-tabulation of frequencies in the corpus, and a factor analysis which simultaneously considered relative influences on noun phrase patterns in English discourse. Further analysis was conducted between the specific clause types in which cohesive noun phrases occurred and the patterns of their antecedent distance. This corpus-based approach made it possible to determine whether the level of grammatical integration of an English clause affected the discourse patterns of English noun phrase cohesion in a quantifiable way.

4. Research Questions

Two research questions were addressed by the current study:

1) What are the patterns and influences on noun phrase cohesion in English discourse?

2) Do English noun phrases pattern differently in discourse depending on the clause type within which they occur?

5. Method

5.1 Data

The research data was taken from the Australian Corpus of English (ACE), a corpus of 17 genres and 500 samples of unedited text, modelled after the Brown and LOB corpora set in design and balance. Four genres were chosen on the basis of ostensibly being the most distinct genres available in the sample design: press reports, fiction prose, non-fiction prose, and letters. Discourse was extracted from the beginning of each of the genre samples up to the 300th Noun Phrase. To maintain the integrity of the discourse and not break up clauses unnecessarily, after the 300th NP the next adjunct constituted the cut off point for extraction of a data set. This resulted in genre sets that were 1-2 noun phrases above 300. This method produced a corpus of 1206 noun phrases, contextualized in that they occurred within unedited discourse and represented four genres, as shown in Table 1.

Table 1. Sample design, noun phrase across genres

Press Reports (Feature Articles)		Fiction Prose (Literature)		Non-fiction prose		Letters (to the Editor)		Total NPs	
302	25%	301	25%	301	25%	302	25%	1206	100%

5.2 Coding and Analysis

To explore patterns in the 1206 noun phrases, every noun phrase was coded for six factor groups (full coding schema provided in Appendix 1). Each factor group was based on a feature identified in previous research as important to noun phrase discourse patterns. These were: 1. NP form (pronoun, single lexical item, multiple lexical items etc); 2. Information salience (old/cohesive or new information); 3. NP syntactic role (Subject transitive, object, copula, adjunct etc); 4. Genre (press, literature, letters or prose); 5. Referential distance (if the NP was old/cohesive, then the number of clauses from the NP to its antecedent was coded 1, 2 3 to 9 or more); 6. Clause type (NP occurred in a to-infinitival, relative, coordinate clause etc). The data were coded directly into the statistical program Goldvarb (Sankoff et al, 2012) for analysis, as in Example 2.

Example 2. Data coding (press report genre)

Original ACE data	Coded ACE Data (codes in Appendix)
The Premier of Queensland, Sir Joh Bjelke Peterson, didn't disappoint the crowd at the opening of the $20 million extension to the State Government Computer Centre in Brisbane last week...	(!0AQM- The Premier of Queensland, (#0JQM- Sir Joh Bjelke-Peterson, didn't disappoint ($0OQM- the crowd at ($0JQM- the opening of the $20 million extension to (!0JQM- the State Government Computer Centre in (#1JQM1 Brisbane last week...

The procedure for analysis consisted of cross-tabulating frequencies across the factor groups to establish patterns of NP discourse cohesion. When required, Factor Group 1 (NP form) was collapsed into Halliday and Hasan's (1976) two categories of grammatical and lexical noun phrases. Further, a factor analysis was conducted to determine the relative weight of influence the factors from different research traditions had on noun phrase cohesion when considered together on the same stretch of discourse. The factor analysis used Factor Group 3, information salience (old or new information), as the dependent variable. Old information was for the purposes of analysis equated to a cohesive noun phrase.

Some difficulties in analysis and coding need to be mentioned. One was deciding in some of the more complex noun phrases when the data presented two NPs which should be independently coded or a single NP. It was decided that nouns functioning as complements in a larger NP constituent should not be considered independent but that an adjunct within a larger constituent should. For example, in Example 2 "the crowd / at the opening..." was coded as two independent NPs. However, "the Premier of Queensland" was not because "Queensland" functions as complement to the head "Premier". A principled decision was also made with regard to relative pronouns. Relative pronouns are technically NPs, but are not an independent factor in the statistical sense because they are obligatorily grammatical forms and obligatory cohesive/old information. As they occurred only in a specific clause type, the relative clause, and clause types were independently coded, all relative pronouns were coded as separate NPs but excluded from the later factor analyses.

6. Results

6.1 The Patterns of Cohesive Information across English Noun Phrases

A noun phrase contained old information if it had some cohesive device that explicitly referenced earlier discourse, such as repeated lexical item from earlier in the discourse, or a pronoun with an antecedent earlier in the discourse. The following series of results reports the frequencies and distribution patterns of old/cohesive information in the discourse data according to each of the factor groups.

6.1.1 The Patterns of Discourse Cohesion according to NP Form

The frequency of new and old/cohesive information is reported in Table 2, cross-tabulated with the different noun phrase forms of English.

Table 2. Patterns of noun phrase forms with old/new information

Noun Phrase Form	**Old Information N**	**%**	**New information N**	**%**	**Total N=100 %**	**% of all NPs**
personal pronoun	177	95.2%	9	4.3%	186	15.4%
poss. pronoun, + lexeme	62	92.5%	5	7.5%	67	5.6%
poss. pronoun, - lexeme	2	100%	0	0%	2	0.2%
demonstrative pronoun	44	93.6%	3	6.4%	47	3.9%
relative pronoun	34	100%	0	0%	34	2.8%
elided relative pronoun	9	100%	0	0%	9	0.7%
impersonal pronoun	9	47.4%	10	52.6%	19	1.6%
interrogative pronoun	1	16.7%	5	83.3%	6	0.5%
There-existential	0	0%	11	100%	11	0.9%
It-cleft, extraposition	9	100%	0	0%	9	0.7%
Multi Lexical + def art.	59	38.8%	93	61.2%	152	12.6%
Multi Lexical + indef art	19	25.7%	55	74.3%	74	6.1%
Multi Lexical no article	83	37.1%	141	62.9%	224	18.6%
Single Lexical+ def art	70	46.4%	81	53.6%	151	12.5%
Single Lexical+ indef art	5	12.5%	35	87.5%	40	3.3%
Single Lexical no article	73	44.8%	90	55.2%	163	13.5%
NP is a numeral	0	0%	12	100%	12	1%
Totals	656	54.4%	550	45.6%	1206	100%

The first thing to note in Table 2 is that the total number of cohesive NPs was 54.4%, while 45.6% of NPs introduced new information. So, the result indicates that in English discourse slightly over half of all noun phrases are cohesive. The main interest of Table 2, however, lies in the patterns of old/new information across the different lexical forms of NPs. In these forms, the highest proportion of old information occurred in noun phrases made up of a single lexical item marked by a definite article; these were cohesive 46.4% of the time. Unmarked single lexical NPs were cohesive 44.8% of the time, constituting the second highest proportion of old information amongst the forms. Multiple lexical NPs marked by the definite article were proportionally the next highest old information carriers at 38%, and unmarked multiple lexical NP's only slightly lower at 37.1%. Finally, the data showed that indefinite multiple lexical NPs and indefinite single NPs were cohesive only 25.7% and 12.5% of the time respectively.

What this means is that the proportion of cohesion declines for both single and multiple lexical NPs along the same pattern: definite > unmarked > indefinite. Yet, single word NPs more often carried discourse cohesion than multiple NPs. The figures therefore reflect an English discourse tendency to reduce forms when multiple mentions are made of the same information. For example, "*an English grammar book* was closed on her desk" - which is a multi-lexical NP - when mentioned later in the discourse may become "*the book* was opened"- a single lexical definite NP. This is probably why multi-word NPs with indefinite articles, like the one in the example just given, were proportionally the most common introducers of new information in English discourse at 74.3%. The pattern that definite article NPs were most frequently cohesive NPs and indefinite ones the least cohesive, reflects the function in English grammar for indefinite articles to mark non-specific, discourse new, information and definite articles to mark known information. However, what is interesting about the figures behind this relatively commonplace knowledge is that while the indefinite article favoured new information, the definite article did as well, although to a much lesser extent. Definite single NPs and definite multiple lexical NPs were new information carriers 61.2% and 53.6% of the time respectively. This means that the definite article in English grammar and discourse, while it does mark cohesive NPs more frequently than any other lexical NP form, actually most of the time introduces new information. It is therefore wrong to say the definite article is a-priori a cohesive device in English which marks some noun phrase as having prior discourse relevance.

NP forms were collapsed into to Halliday and Hasan's (1976) categories of grammatical (i.e. all pronominal forms) and lexical (i.e. all other forms). Cross-tabulation with old information revealed that neither grammatical cohesion nor lexical cohesion carried markedly more of the total quantity of discourse cohesion. This indicates that in English although a typical function of grammatical forms, such as pronouns for example, is to mark discourse cohesion (see Table 2), this does not translate at the discourse level to grammatical cohesion having a larger role than lexical cohesion in English discourse, which is the result shown in Table 3.

Table 3. Lexical and grammatical cohesion in English discourse

	Grammatical NPs		Lexical NPs		Total	
Old Information	347	52.9%	309	47.1%	656	100%

6.1.2 The Patterns of Discourse Cohesion according to Genre

The separate genres had a maximum 10% difference in their proportions of cohesive noun phrases, as reported in Table 4. Overall, there were relatively small differences amongst the genres. Press reports were the only genre to have a comparatively higher frequency of noun phrases cohesion. This might reflect that this genre generally focuses on a single issue or event, which would be conducive to noun phrase cohesion as repeated reference to the central issue/event would have to be made.

Table 4. Old/New information noun phrases across genre

	Old information NPs		New information NPs		Total NPs	
Press reports	183	60.6%	119	39.4%	302	25%
Fiction Prose	163	54%	139	46%	301	25%
Letters (to Editor)	158	52.5%	143	47.5%	302	25%
Non-fiction Prose	152	50.5%	149	49.5%	301	25%
				Total:	1206	100%

However, as reported in Table 5, when patterns of grammatical and lexical forms were considered along with the distribution of discourse cohesion more distinct genre patterns emerged. Press reports, which had the highest number of cohesive noun phrases, also had the highest number lexical form NPs. Together the results indicate this genre maintains a higher level of cohesion than other genres and furthermore does so mostly through lexical noun phrases. It is likely a conscious stylistic choice in press reports to minimize possible ambiguity in their texts by avoiding grammatical pronouns, which can sometimes have uncertain reference when there are multiple discourse participants. Conversely, fiction prose stands out in its use of grammatical NPs, with over half of its noun phrases in grammatical form at 55.3%. This comparatively high frequency means that grammatical cohesion plays a larger role in the discourse cohesion of fiction prose than in other genres.

Table 5. Lexical and grammatical NPs across genres

	Grammatical NPs		Lexical NPs		Total NPs	
Press reports	62	20.5%	240	79.5%	302	25%
Fiction Prose	158	52.3%	144	47.7%	302	25%
Letters (to Editor)	69	22.9%	232	77.1%	301	25%
Non-fiction Prose	101	33.6%	200	66.4%	301	25%

6.1.3 The Patterns of Discourse Cohesion according to Noun Phrase Argument Role

A discourse pattern predicted by cross-linguistic studies is preferred argument structure. This predicts that agent role noun phrases avoid the introduction of new information. In English grammar and discourse the A-role can be considered equivalent to the subject of transitive clauses for corpus manipulation (Ashby & Bentivligio, 2003). The distribution of old/new information was therefore cross-tabulated with the syntactic roles of the noun phrases. The results, shown in Table 6, indicated that noun phrases in the transitive subject position clearly avoided introducing new information, being cohesive 83.9% of the time. Discourse cohesion in the English data evidently fulfilled cross-linguistic tendencies of preferred argument structure.

Table 6. Old/New information according to syntactic role

	Old information NPs		New information NPs		Total NP forms	
Transitive Subject	162	83.9%	31	16.1%	193	16%
Intransitive Subject	101	71.1%	41	28.9%	142	11.8%
Copula subject	62	63.3%	36	36.7%	98	8.1%
Object	131	55%	107	45%	238	19.7%
Predicate copula	31	32.6%	64	67.4%	95	7.9%
Adjunct	164	38.3%	264	61.7%	428	35.5%
Passive agent (by-)	5	41.7%	7	58.3%	12	1%
				Total	1206	100%

However, the results also revealed that the other subject roles of English similarly favoured being referential, though not as strongly as the transitive subject. Intransitive subject NPs were cohesive 71.1% of the time, and copula subjects 63.3%. A discussion of agency by Payne (2011) proposed that one might view the English subject roles as typically having declining grades of agency. For example “she sang a song” is a transitive, highly agentive clause, while “she sang at the concert” is intransitive and moderately agentive, and finally “she was singer” is copulative and only slightly agentive. The current results seem to confirm such a gradation exists in English agency and that it is reflected quantitatively in the preferred argument structure of the language. For English discourse therefore, preferred argument structure might be broadly defined as a pattern for noun phrases to be cohesive across all subject positions.

Noun phrases in predicates of copula sentences, and in adjuncts, both preferred introducing new information at 67.4% and 61.7% respectively. Two patterns are suggested by these figures. One is that when new information is introduced into English discourse, it tends be done outside the core constituents of the clause, explaining the high number of adjuncts with new information. The other is that since predicate copula NPs tend to contain new information, but copula subjects tend to contain old information, the typical pattern for a copula sentence must be to start with a cohesive noun phrase and then predicate discourse-new information about it. According to Huddleston and Pullum (2002), the copula sentence grammatically ascribes or identifies a quality of its subject; the current results suggest adding to this that in English discourse it tends to ascribe or identify a discourse new quality about a discourse old, i.e. a cohesive, subject.

6.1.4 Clause Type and Noun Phrase Cohesion in English Discourse

The clause type in which a noun phrase appeared was cross-tabulated with old/new information status. As Table 7 shows, about half of all noun phrases (50.8%) in the data occurred in subordinate or coordinate clauses.

Table 7. Clause type and the distribution of NP discourse cohesion

	Old information NPs		New information NPs		Total NPs	
Main clause	314	53%	279	47%	593	49.2%
to-infinitival	39	55.7%	31	44.3%	70	5.8%
past participle	12	46.2%	14	53.8%	62	5.1%
present participle	25	40.3%	37	59.7%	62	5.1%
content clause	67	62.6%	40	37.4%	107	8.9%
relative clause	66	70.2%	28	29.8%	94	7.8%
comparative clause	3	60%	2	40%	5	0.4%
adverbial clause	76	55.5%	61	44.5%	137	11.4%
coordinate clause	39	55.7%	31	44.3%	70	5.8%
non-clausal	15	35.7%	27	64.3%	42	3.5%
				Total:	1206	100%

Relative clauses had the highest proportion of cohesive NPs, but 34 of the cohesive NPs were relative pronouns (see Table 2). Once disregarded, the comparative proportions of cohesive noun phrases were much the same across all clause types. Participle clauses, however, did contain slightly lower proportions of cohesive NPs.

6.1.5 Factors Analysis of the Patterns of Noun Phrase Cohesion in English Discourse

To determine which factors were or were not statistically significant in influencing cohesion across English noun phrases at the discourse level, a factor analysis considered all of the discourse factors simultaneously regarding their influence on old information patterns in the 1206 NP dataset. Three factor groups were statistically significant: noun phrase form, syntactic role and the discourse genre.

Table 8. A Factor analysis of influences on noun phrase discourse cohesion

Noun Phrase Form	Factor Weight	Syntactic role	Factor Weight	Genre	Factor Weight
Lexical	0.303	Trans. Subject	0.730	Press reports	0.642
Grammatical	0.887	Intrans. Subject	0.644	Fiction Prose	0.337
		Copula subj.	0.560	Letters	0.565
		Object	0.471	Non-fiction	0.462
		Predicate comp	0.292		
		Adjunct	0.398		
		Passive agent	0.380		
				Log likelihood = -587.012, $p < 0.05$	

As reported in Table 8, the factor most determinate of whether an NP was discourse cohesive was whether the NP form was grammatical. This is not surprising as grammatical NPs are made up of a range of pronominal forms which are typically anaphoric in reference, and so most NPs with a pronoun in them would fall into Halliday and Hasan's (1976) category of grammatical cohesion. More interestingly, the syntactic role of the noun phrase was shown to influence whether it would be discourse cohesive. As shown in Table 8, transitive subjects and intransitive subjects strongly favoured selecting cohesive noun phrases, with weights at 0.730 and 0.644 respectively. Copula subjects slightly favoured old information at 0.560. This indicates that preferred argument structure is a determining influence on the patterns of cohesion in discourse. Genre was also significant in the discourse patterns of noun phrase cohesion, with press reports in particular favouring cohesive NPs at 0.642.

Considered amongst the other factors at work on English discourse, the clause type in which a noun phrase occurred was not a significant influence on whether that NP contributed to discourse cohesion. The relationship between cohesion and clause structure does not translate to a significant influence at the discourse level on the quantity of cohesive noun phrases in specific clause types.

6.2 A Final Look at Clause Patterns and Noun Phrase Cohesion in Discourse

Frequency counts revealed a near lockstep decline existed in English discourse between the number of cohesive NPs and the clausal distance to their antecedent. As shown in Table 9, most cohesive noun phrases had antecedents which occurred close to them. As antecedent distance increased, measured by the number of intervening clauses, the quantity of cohesive NPs decreased.

Table 9. Cohesive noun phrases and antecedent distance

Antecedent Distance	Number of old information NPs
1 clause	275
2 clauses	135
3 clauses	77
4 clauses	39
5 clauses	28
6 clauses	22
7 clauses	10
8 clauses	2
Above 9 clauses	63

A final analysis was therefore conducted on the range of combined clauses (main clause and non-clausal NPs excluded) to see if the more integrated clause types had a lower quantity of cohesive NPs with close antecedents when compared to the less integrated clause types. Fewer NPs with nearby cohesive antecedents in tightly integrated clauses might be a result of cohesion being carried through their syntactic relations with their immediate context rather than needing to use cohesive NPs. Table 10 reports the cross-tabulation of clause type, old/cohesive information NPs and the distance to their antecedent.

Table 10. Cohesive noun phrases, clause type and antecedent distance

Antecedent distance	Clause type								Total NPs
	to-inf.	pres. part	past part	content clause	rel. clause	comp. clause	Adv. clause	Coord. Clause	
1 clause	13	10	3	23	48	1	29	17	144
2 clauses	4	6	4	16	6	2	13	11	59
3 clauses	7	3	3	9	4	0	12	4	42
4 clauses	7	2	0	3	4	0	9	3	28
5 clauses	1	0	0	3	3	0	2	1	10
6 clauses	3	1	0	1	0	0	3	1	9
7 clauses	0	0	0	0	0	1	0	1	2
8 clauses	0	0	0	2	0	0	0	1	3
> 9 clauses	3	2	4	10	3	0	5	1	28
Total									571

In nearly all combined clause types the majority of cohesive noun phrases had their antecedent only one clause away, which would mean the other clause with which it combined. This is the main clause for subordinates and the other coordinate for coordinated clauses. At this inter-clausal level it seems to be the case that the more integrated combined clause types, e.g. the to-infinitival, the present and past participle clauses, had the least number of cohesive noun phrases with an antecedent in their immediate context (i.e. within the previous 1-2 clauses). So, while the factor analysis revealed clause structure was not a significant determiner of the overall quantity of discourse cohesion a clause will contain, there nonetheless seems to be a pattern in which the more integrated an English clause type is grammatically, the less it needs to create cohesion through a noun phrase at the local level.

7. Discussion

The results of the corpus analysis of noun phrases in English discourse revealed some clear patterns for discussion. The lexical noun phrase form with the clearest role in English discourse cohesion was single word lexical NPs. These forms seem to be favoured to carry discourse cohesion because information tends to be reduced upon multiple mentions, as noted by Christensen (2011). Relatedly, the highest amount of new information in English discourse is introduced in multiple word NPs. The syntactic role of subject in English grammar has a tendency to be a cohesive noun phrase in English discourse. This is particularly so with transitive subjects (Ashby & Bentivligio, 2003), but is also a strong tendency across all subject types due to the different gradations of agency in English (Payne, 2011). Conversely, non-core constituents such as adjuncts are fundamental to the introduction of new information. This result indicates therefore that the core grammatical roles of the English clause are more associated with old information than are non-integral adjunct constituents, which favoured new information. One might propose from this result that the syntactic coding of NP constituents essential to the grammaticality of an English clause are more implicated in discourse cohesion than constituents which do not fulfil an essential argument role.

Different aspects of English grammar and discourse, however, have different influences on noun phrase cohesion. This study, by bringing together the central aspects of separate research traditions within a framework based on Halliday and Hasan's Cohesion in English (1976) has been able to show that when considered simultaneously on a corpus of NPs in their discourse context, the different focus areas of each tradition have different levels of influence on noun phrase patterns. NP form turned out to have the strongest influence on the patterns of English

discourse, followed by syntactic role, then by genre. Clause structure, however, had a relatively insignificant influence on the general pattern of discourse cohesion when considered amongst these other features.

Nonetheless, there emerged one important pattern with regard to clause structure, which has both applied and theoretical interest. Based on Givon (2001) and Mathessien (2003) who proposed that English combined clause types exist on a spectrum from highly integrated at the grammatical level to loosely integrated via cohesive devices at the discourse level, a hypothesis was developed that the more that cohesion is carried by the grammar of an English clause type (i.e. the more integrated it is), the fewer cohesive NPs it would need to create cohesion with local relations. This study showed quantitative evidence for such a complementary distribution between the level of integration of a clause and the number of cohesive devices in its immediate discourse context. At this local inter-clausal level, looser clauses such as coordinate and adverbial clauses had a higher number of cohesive NPs with close antecedents, while tighter clauses such as to-infinitival and participle clauses had fewer. Results may not have indicated any specific clause type carried typically more or less cohesion in the overall discourse, which was ruled out by the factor analysis, but the study has indicated that English clause integration at the local level displaces nearby English discourse cohesion. Previous studies of which this author is aware have not yet shown these quantitative patterns in English discourse cohesion that result from the different integration levels of the English clause types. The results of this study therefore support theories that clause combination and cohesion exist as part of a single phenomenon- a continuum of integration from discourse to grammar (Givon, 1979; Mathessien, 2003).

8. Limitations and Suggestions for Future Research

Future research should continue to try to draw together different research strands beyond those used in this study. For example, included in analyses might be features not considered in this study but which have significant prior literature, such as the effects of gender and multiple discourse participants on noun phrase patterns (Arnold, 2007). Larger data sets than the limited number of NPs coded for this study should also be employed. Further corpus-based studies have the potential to significantly contribute to a both an applied and theoretical understanding of the precise roles and influences of all relevant discourse features on noun phrase cohesion in English.

9. Conclusion

This study has attempted to show that separate previous research trends on discourse cohesion can better our understanding of patterns in English when they are productively bought together in a framework for simultaneous analysis. The current research has addressed two research questions. One established a range of patterns and influences on noun phrase cohesion in English discourse. The patterns revealed by this study included that noun phrase form has a strong influence on English discourse cohesion, followed by argument structure, and by genre. The second research question concerned whether English noun phrases patterned differently with respect to discourse cohesion depending on the clause type within which they occurred. When considered simultaneously amongst the influences of genre, syntactic role, and NP, form, it was shown that clause type was not so significant in determining the quantity or quality of NP cohesion. However, further investigation revealed clause grammar does seem to code cohesion locally. This study was able to show that the more integrated a clause was the more it displaced the need for noun phrases to mark cohesion in the immediate discourse through cohesive NPs with antecedents within 1-2 clauses distance. The study therefore quantitatively supports previous theories that discourse cohesion and the different types of combined clauses in English exist along a continuum from grammar to discourse.

References

Arnold, J. (2007). Reference production: production internal and addressee-orientated processes. *Language and Cognitive Processes*, *23*(4), 495-527. http://dx.doi.org/10.1080/01690960801920099

Arnold, J., & Griffin, Z. (2007). The effect of additional characters on choice of referring expression: everyone counts. *Journal of Memory and Language*, *56*, 521-556. http://dx.doi.org/10.1016/j.jml.2006.09.007

Ashby, W., & Bentivligio, P. (2003). Preferred argument structure across time and space: a comparative analysis of French and Spanish. In J. Du Bois, L. Kumph, & W. Ashby (Eds.), *Preferred argument structure: grammar as architecture for function.* Amsterdam: John Benjamins.

Beugrande, R., & Dressler, W. (1984). *Introduction to text linguistics*. London: Longman.

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511621024

Chafe, W. (1985). Linguistic differences produced by differences between speaking and writing. In D. Olson, N. Torrance, & A. Hildyard (Eds.), *Literacy, language, and learning: the nature and consequences of reading and writing*. Cambridge: Cambridge University Press.

Christiansen, T. (2011). *Cohesion: a discourse perspective.* Berlin: Peter Lang.

Collins, P., & Hollo, C. (2009). *English grammar: an introduction.* Chippenham: Palgrave McMillian.

Crossley, S., & McNamara, D. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, *18*, 119-135. http://dx.doi.org/10.1016/j.jslw.2009.02.002

Dahl, O. (2008). Two pathways of grammatical evolution. *Unpublished proceedings of the 12th biennial Rice symposium on language.* Houston: Texas.

Du Bois, J. (2003). Argument structure: grammar in use. In J. Du Bois, L. Kumph, & W. Ashby (Eds.), *Preferred argument structure: grammar as architecture for function.* Amsterdam: John Benjamins.

Givón, T. (1979). From discourse to syntax: grammar as a processing strategy. In T. Givón (Ed.), *Discourse and syntax.* New York: Academic Press.

Givón, T. (1990). *Syntax: a functional typological introduction.* Philadelphia: John Benjamins.

Givón, T. (2001). *Syntax: volume I.* Amsterdam/Philadelphia: John Benjamins.

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Huddelston, R., & Pullum, G. (2002) *Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Matthiessen, C. (2002). Combining clauses into clause complexes: a multi-faceted view. In Joan Bybee & Michael Noonan (Eds.), *Complex sentences in grammar and discourse: essays in honor of Sandra A. Thompson.* Amsterdam: John Benjamins.

McNamara, D., Ozuru, Y., Graesser, A., & Louwerse, M. (2006). Validating coh-metrix. In R. Sun, & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society.* Mahwah: Erlbaum.

Mirzapour, F., & Ahmadi, M. (2011). Study on lexical cohesion in English and Persian research articles. *English Language Teaching*, *4*(4), 243-255. http://dx.doi.org/10.5539/elt.v4n4p245

Payne, T. (2011). *Understanding English grammar: a linguistic introduction.* New York: Cambridge University Press.

Peters, P. (1986). *Austalian corpus of English (ACE).* Sydney: Department of Linguistics, Macquarie University.

Sankoff, D., Tagliomonte, S., & Smith, E. (2012). *Goldvarb: A variable rule application for Macintosh.* Department of Linguistics, University of Toronto.

Appendix:

Appendix 1. Summary of coding schema

FACTOR GROUP	CODE	FACTOR
1. NOUN PHRASE FORM	p	personal pronoun
	s	possessive pronoun with lexeme
	v	possessive pronoun without lexeme
	d	demonstrative pronoun
	r	relative pronoun
	e	elided relative pronoun
	i	impersonal pronoun
	?	interrogative pronoun
	t	There-existential
	c	It-cleft, extraposition
	!	Lexical NP, multiple lexical words, def article
	@	Lexical NP, multiple lexical words, indef article
	#	Lexical NP, multiple lexical words, no article
	$	Lexical NP, single word with definite article
	%	Lexical NP, single word with indefinite article
	^	Lexical NP single word, no article
	n	NP is a numeral
2. INFORMATION SALIENCE	1	old/cohesive information NP
	0	new information NP
3. NP SYNTACTIC ROLE	A	Transitive Subject
	S	Intransitive Subject
	C	Copulative Subject
	O	Object
	P	Complement, copula predicate
	J	Adjunct (no argument role)
	B	Passive Agent Adjunct (by phrase)
4. GENRE	Q	Press
	W	Letters
	E	Fiction prose
	R	Non-fiction prose
5. CLAUSE TYPE NP OCCURS WITHIN	M	Main clause
	X	non-clausal material
	T	to-infinitival
	N	past participle
	R	present participle
	C	content clause
	V	relative clause
	P	comparative clause
	H	adverbial clause
	A	coordinate clause

6. ANTECEDENT DISTANCE	1	1 clause
	2	2 clauses
	3	3 clauses
	4	4 clauses
	5	5 clauses
	6	6 clauses
	7	7 clauses
	8	8 clauses
	9	9 or more clauses
	-	No antecedent (non cohesive NP)