Modelling the flow of discourse in a corpus of written academic English

Nick Moore
Khalifa University, United Arab Emirates

Abstract

Discourse studies attempt to describe how context affects text, and how text progresses from one sentence to the next. Systemic Functional Linguistics (SFL) offers a model of language to describe how information flow varies according to context and co-text through the Textual metafunction, especially using the functions of Participant Identification and Tracking, Theme and Information Structure. These systems were evaluated by assembling a corpus of academic texts and assessing their information flow. Results of the analysis of the three grammatical systems in the Textual Metafunction demonstrate significant patterns, or unmarked choices, where the participant, thematic and information systems combine to powerful effect. Where the systems are not aligned, there is a recognisable effect on the flow of information.

Discourse

Various approaches to the study of discourse, while demonstrating a range of perspectives, seem to agree on two major points: discourse is derived from context and discourse derives meaning from the combination of sentences. The challenge for linguistics is to combine these perspectives into a coherent model of language, and the challenge for corpus linguistics is to incorporate a coherent model of discourse into corpus methods and analyses.

Discourse is a key element of any comprehensive description of language. The ability to assert that a sentence is well-formed, or to identify *hapax legomena* in a given corpus, neglects major aspects of the meaning, role and function of language. Hymes' notion of Communicative Competence, from a sociolinguistic and anthropological perspective, and his oft-quoted insight that "There are rules of use without which the rules of grammar would be useless" (1972, p.278) motivated more linguists to consider the social context in linguistic analysis. This has resulted in discourse analysis becoming a recognisable field of study with a focus on the influence of context on language:

> The analysis of **discourse** is, necessarily, the analysis of language in use. As such it cannot be restricted to the description of linguistic forms independent of the functions or purposes which those forms are designed to serve in human affairs. (Brown & Yule. 1983, p.1)

One response to the challenge of bringing context into language is to narrow down the possible social contexts and to focus on contextually-defined language use in genre-based discourse analyses, as typified by Paltridge who defines discourse analysis as "an approach to the analysis of language that looks at patterns of language across texts as well as the social

and cultural contexts in which the texts occur" (2006, p.1) and by Bhatia who uses the term discourse "in a general sense to refer to language use in institutional, professional or more general social contexts." (2004, p.3) An alternative approach to the challenge of combining language and context is to ignore the arbitrary form-meaning relationship of language and focus on the social factors that enable the construal of meaning and the exercising of power through discourse, as expressed in the work of Foucault and others:

> Of course, discourses are composed of signs; but what they do is more than use these signs to designate things. It is this **more** that renders them irreducible to the language (langue) and to speech. It is this 'more' that we must reveal and describe. (Foucault, 1972. p.54)

One could argue that neither of these approaches is entirely satisfactory since the genre-based studies limit context and the Foucault-inspired studies disregard meaning in language variation.

The second aspect of language that is commonly emphasised in discourse studies is that of the dynamic inter-relation of sentences that transform a random sequence into a recognisably coherent text. Stubbs views this as a central challenge to discourse analysis:

> The basic problem is to account for the recognizable unity or connectedness of stretches of language, whether this unit is structural, or semantic or functional. (1983, p.9)

While this challenge has been elucidated, it is not clear that it has been met, despite confident assessments of progress:

> Discourse analysis has grown into a wide-ranging and heterogeneous discipline which finds its unity in the description of language above the sentences and an interest in the contexts and cultural influences which affect language in use. (McCarthy. 1991, p.7)

Although I have suggested that the two approaches to discourse – emphasizing context and accounting for cohesion – are distinct, they often merge: "... language in use, for communication – is called discourse; and the search for what gives discourse coherence is discourse analysis." (Cook, 1989, p.6). Gee (2008) proposes that cohesion is not defined in purely linguistic terms, but must consider the social context:

> By "discourse" I mean stretches of language which "hang together" so as to make sense to some community of people, such as a contribution to a conversation or a story. ... Making sense is always a social and variable matter: what makes sense to one community may not make sense to another. Thus, to understand sense making in language it is necessary to understand the ways in which language is embedded in society and social institutions. (p.115)

The combination of both perspectives must necessarily take into account the way that language continually construes and re-construes context – language not only represents aspects of the reality that it attempts to refer to, but in doing so it simultaneously transforms that reality. This implies that discourse analysis must treat context as the dynamic response to co-textual and con-textual influences:

> a particular grammatical structure, in many cases, becomes inappropriate only in the context of preceding and subsequent discourse... It is in this somewhat negative and loose sense (i.e. not sentence level) that we are taking the term discourse context at present. (Hughes, Carter & McCarthy, 1994 p.49)

Thus, in order to incorporate a discourse-based perspective, Corpus Linguistics must take up the challenge of combining con-textual and co-textual meaning to account for language variation. I believe that the most applicable model of language that can respond to this challenge is offered by Systemic Functional Linguistics (SFL).

Context, Co-Text & the Textual Metafunction

SFL analyses context at two levels: the Context of Culture is realized by the Context of Situation (Martin and Rose, 2008). Instances of the Context of Situation can then be aggregated to typify recognisable patterns of behavior within the context of a culture (Halliday and Matthiessen, 2004). The Context of Situation is characterized as a configuration of Field, Tenor and Mode (Halliday and Hasan, 1985) and these are realized through the corresponding metafunctions of Ideational, Interpersonal and Textual meaning. Language associated with Ideational functions construes experience, language associated with Interpersonal functions enacts relationships, while language associated with Textual functions organizes and instantiates discourse (Martin and Rose, 2008). This paper focuses on aspects of the Textual metafunction that operate within the clause to instantiate a text within a context.

It is through the textual metafunction that SFL models the resources that locate an instance of text within the meaning potential of the systems of language, and so it is through the resources of the textual metafunction that we are able to trace the co-text and context through discourse. From the textual metafunction, the resources of Participants, Theme and Information enable us to analyse the cohesive, dynamic unfolding of context through discourse.

Cohesion & Reference

Linguistically modeling the development of ideas through discourse presents major challenges for formal, computational and corpus-based approaches to language description. While computational studies have shown that anaphora play a part in the flow of information, they cannot tell the whole story (Beaver 2004; Botley and McEnery 2000). Descriptions of reference are typically motivated by computer studies of anaphora and attempts to automate anaphora resolution (Mitkov 2000; Mitkov, Lappin and Boguraev 2001). One of the most successful among these approaches is that of Centering (Grosz, Joshi and Weinstein, 1995; Grosz and Sidner 1998) and its related theories (e.g. Strube and Hahn 1999; Karamanis *et al.* 2008, Taboada and Zabala, 2008), although centering studies typically suffer the same drawback as other computational studies by focusing on a narrow range of pronouns. An important insight offered by Centering studies is that as a referent is introduced to the text it is added to the 'stack' of forward-looking Centers that become candidates for backward-looking Centers in following discourse. Centering rules dictate which forward-looking Center is most likely to be taken up by the next backward-looking Center. Although this study does not aim for automatic resolution of anaphora, the system chosen should assimilate this

insight. The system of reference used in this study is based on Martin's (1992) description of Participant Identification and Tracking, itself a discourse semantics response to Halliday and Hasan's (1976) description of cohesion relations, as I believe it demonstrates a compatibility with a Centering approach by not specifying the psychological ordering within referencing. Emmott (1997) also demonstrates that an SFL-based system of reference is compatible with a carrying-forward approach to anaphoric relations.

A participant is any element in the clause that can act as (obligatory) Agent or (optional) Medium in a transitivity analysis (Martin 1992, p.129). A participant is typically a nominal group, but nominal groups such as the 'empty' *it* in "it's raining" are not participants. In addition, nominal groups within adverbial or prepositional phrases have the potential to become participants in discourse. Both participants and 'potential' participants can be analysed for their role in two separate systems. A Participant must first be identified as such. Participant Identification analysis (Martin, 1992) is primarily concerned with Presenting and Presuming reference, but also analyses categories of Comparison, and Generic or Specified reference (see Fig. 1). When a Participant has been identified it can enter a Tracking analysis which identifies the location of the identity of the participant (Martin, 1992). Participant Tracking may extend only so far as recognising the participant as an addition to the discourse, or it may be necessary to specify the location of the referent implied in the text. Participant Tracking classifies the different types of phoricity, and this paper follows a scheme which allows for semantic as well as grammatical relations to be tracked between referents. The location of the identity of a participant can be tracked to the context of culture, the context of situation or another participant in the co-text through a variety of semantic relations (Fig. 2), bringing into the clause the context and co-text of the discourse. This allows a tracking analysis to dispense with the concept of bridging which contributes little to our understanding of discourse (Caselli and Prodanof, 2006; Moore, 2008).

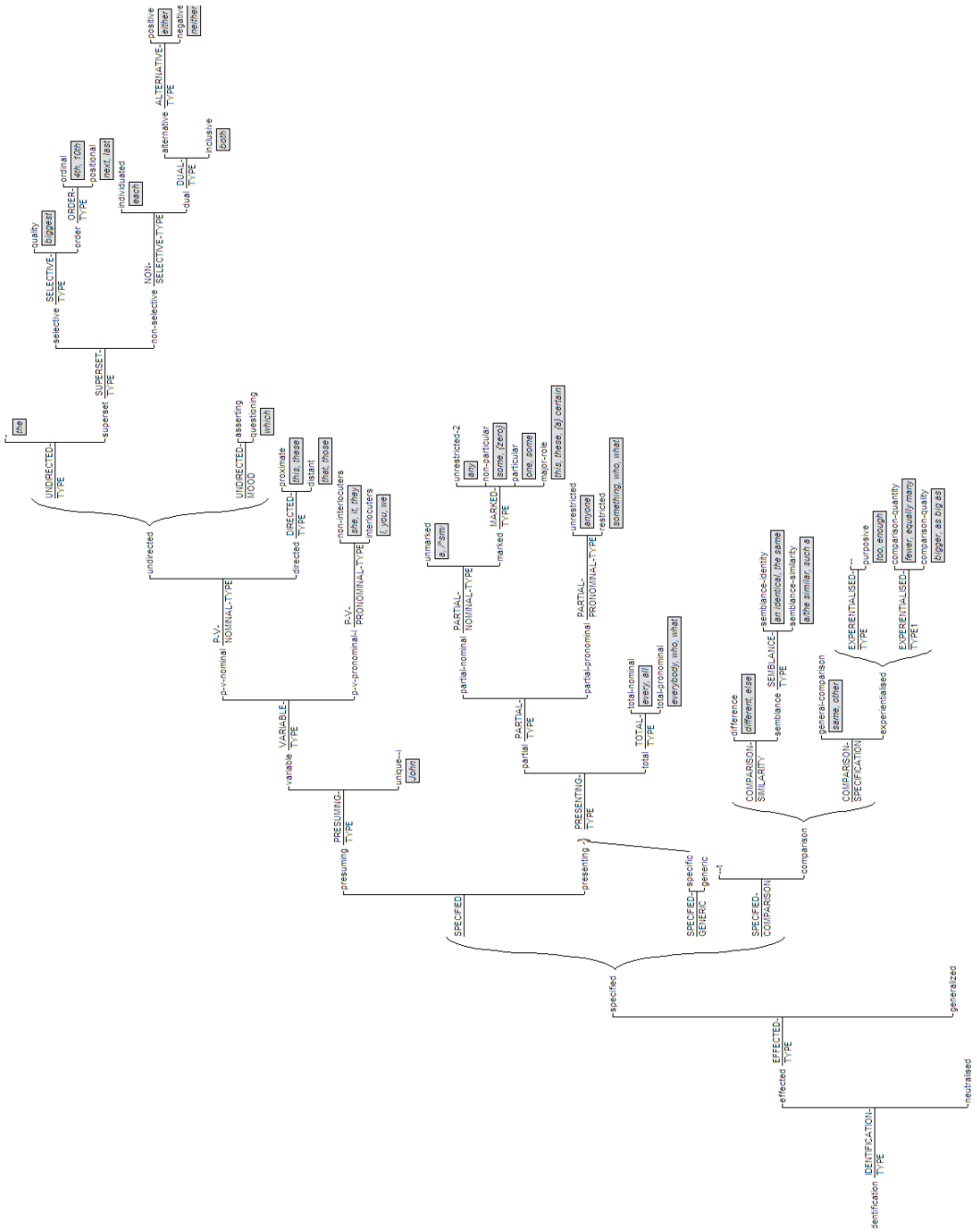

Fig.1 Participant Identification Network

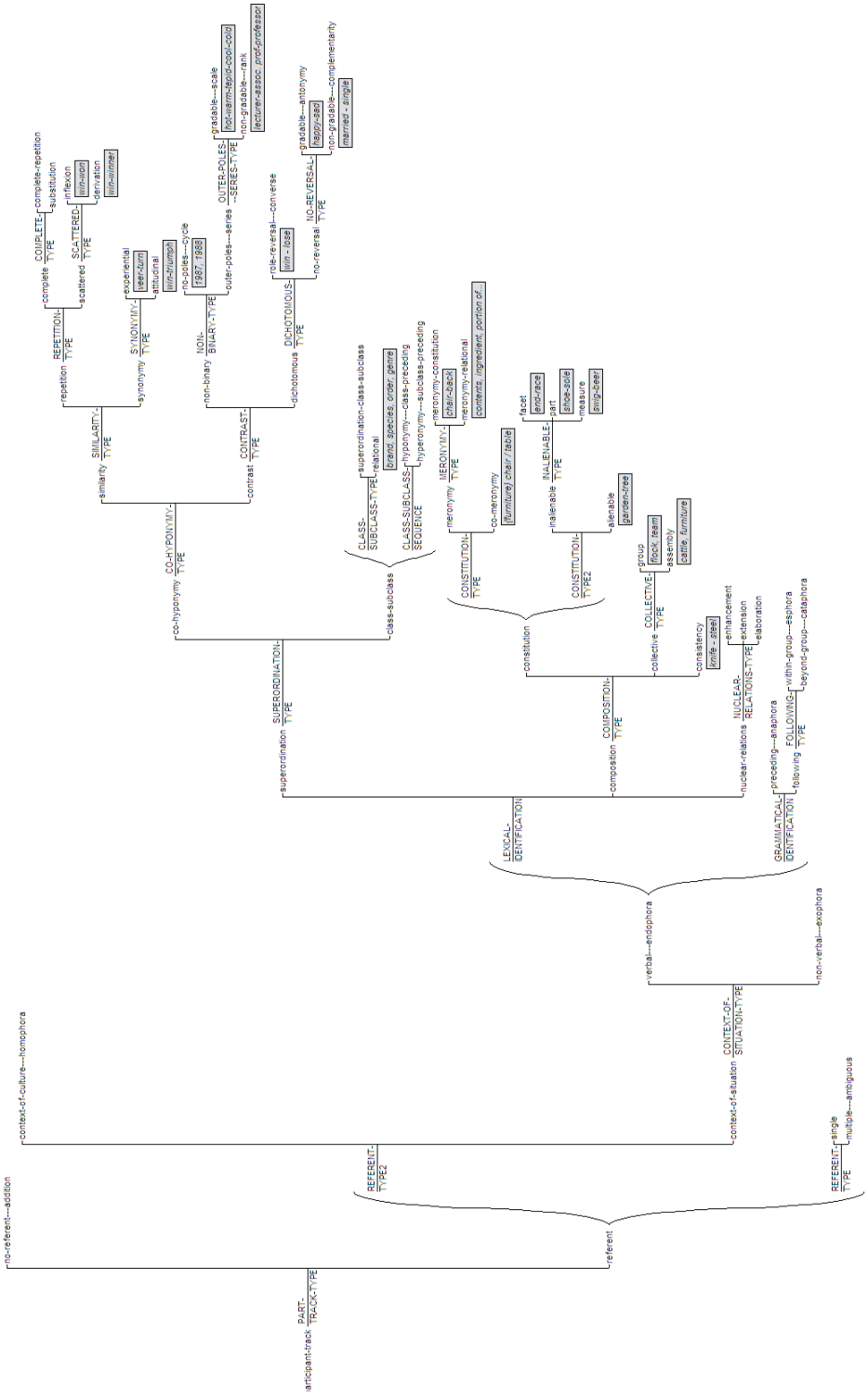

Fig. 2 Participant Tracking Network

Theme & Common Ground

The next area of the clause is Theme, which could be described as setting the context of a clause while it happens. The Theme is “what is being talked about, the point of departure for the clause as a message” (Halliday, 1967a, p.212). Before equating the first part of the quote

with a traditional definition of subject or Topic, it must be remembered that Themes can play a range of grammatical roles: only one type of Theme – the Topical Theme – could be compared to a subject. Textual and Interpersonal Themes do not 'talk about' agents in a transitive clause but may operate as a conjunction or as a form of address, respectively. That is, the role of the clause in discourse may be to 'talk about' relationships between interlocutors, rhetorically relate the current clause to previous discourse, or continue discussion of a topic. Themes are where the clause-as-a-message starts from. In English, the Theme is realized by initial position in a clause or tone unit (Halliday and Matthiessen, 2004).

While Themes are realised in a clause, it is their behaviour across clauses that contributes to an understanding of the dynamics of discourse. As Themes build through the text, they reveal a pathway through the discourse – where the text is heading at each clausal juncture – in terms of the three metafunctions (ideational, interpersonal and textual meanings). A string of Themes generally contribute to recognisable patterns, known as Method of Development (Fries, 1981, 2009; Crompton, 2004) or Thematic Progression. Thematic Progression patterns include Linear, Constant, Derived (Daneš, 1972), contiguous or interrupted (Dubois, 1987).

Information & Focus

There is a final, crucial part of discourse structuring realised within the clause, which is most recognisable when we return to spoken language. Speech is a continuous stream of sound, with very few spaces between sounds. However, speakers must draw breath, and language is better understood when divided into units which will necessarily be limited by the respiratory system. In English these units are the tone unit. Furthermore, a speaker is able to direct the attention of the listener through the non-arbitrary realisation of the tonic foot; the part of the message that is easiest for the ears to distinguish is the part of the message that the speaker wants to the listener to focus on. This function is what Halliday (1967b) referred to as Information Structure, with the tonic foot realising the obligatory New information, and the remaining part of the intonation unit being referred to as the function of Given in a spoken message. It is important to remember that the functions of New and not-New (Given) Information are realised and operate independently of reference (including Presenting and Presuming) and Theme in spoken English. The same functions can be presumed to operate in written English (Fries, 2000). While the realisations of Participant Identification, Participant Tracking and Theme are identical in written and spoken English, intonation contours and the tonic foot are not realised in written English. In SFL, the commonly-accepted realisation of New Information in written English is clause-final position (Matthiessen, 1995; Martin, 1992; Fries, 2002).

We can model the systemic choices in a written clause by combining the systems of Theme and Information as shown in Fig. 3. That is, Theme and New Information retain their position in the clause, but the elements placed in those positions will vary according to the requirements of the message.

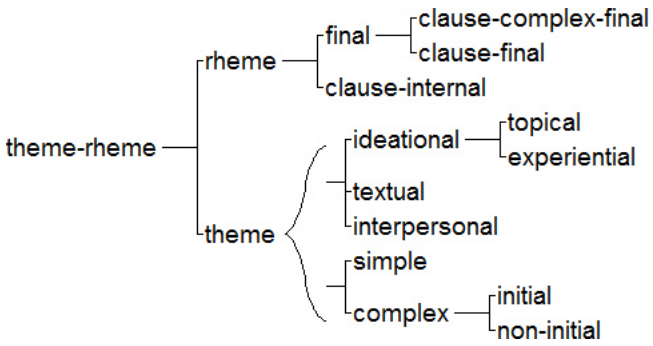

Fig. 3: Theme and Information Systems in the clause

When we recombine the three systems described above – Participants, Theme and Information – and examine their role in discourse, we find enormous potential for making and manipulating context and co-text-specific meanings in text. It is through the variety of combinations in and between the systems that SFL is able to account for a rich variety of meaning-making in discourse, rather than depending on accounts of Given and New which rely heavily on a folk psychology understanding of new information in text (e.g. Clark and Haviland, 1977). Despite the various proposals for this approach to modeling information flow in discourse (Martin, 1992; Fries, 2000; Martin and Rose, 2007) and various applications of SFL theory to corpus studies (e.g. Halliday and James, 1993; Halliday and Matthiessen, 2004; Matthiessen, 2005), however, few studies have examined the reliability of the SFL model of information flow against corpus data.

Case Study

The proposed model of information flow – the independent and combined systems of Participant Identification and Tracking, Theme and Information – was evaluated against a corpus of academic texts from engineering disciplines.

Corpus

The corpus used to evaluate the proposed model for the flow of discourse was a pedagogic corpus designed to sample typical texts that were set for students following undergraduate engineering degrees at my place of work (Khalifa University, formerly Etisalat University College, U.A.E.). These texts are summarized in Table 1.

Source	Total Tokens	Total Types	Type-Token Ratio	Average Word Length
Black, B. J. (1997). *Workshop Practices and Materials*. London: Butterworth-Heinemann.	953	281	0.294858	4.552990
Buchla, D. & McLachlan, W. (1992). *Applied Electronic Instrumentation and Measurement*. Englewood Cliffs: Prentice Hall.	1722	486	0.282230	5.024971
Coope, S., Cowley, J. & Willis, N. (2002). *Computer Systems: Architecture, Networks and Communications.* London: McGraw Hill.	407	218	0.535627	5.058968
Coulouris, G., Dollimore, J. & Kindberg, T. (2001). *Distributed Systems: Concepts and Design. 3rd Edition.* Harlow: Pearson.	445	211	0.474157	5.395505
Coulouris, G., Dollimore, J. & Kindberg, T. (2001). *Distributed Systems: Concepts and Design. 3rd Edition.* Harlow: Pearson.	654	270	0.412844	5.298165
Horowitz, P. & Hill, W. (1989). *The Art of Electronics*. Cambridge: CUP.	331	168	0.507553	5.141994
RAF Training Manual ref: RAF PTC CN 3787 1-1-6 06-528a/01/B50 1-1-7	2656	711	0.267696	4.729669
Rappaport, T.S. (2002). *Wireless Communications – Principles and Practice*. Upper Saddle River: Prentice Hall.	1214	386	0.317957	5.261944
Tannenbaum, A.S. (1995). *Distributed Operating Systems*. Upper Saddle River: Prentice Hall.	775	286	0.369032	4.703226
Total	**9157**	**1912**	**0.208802**	**4.937644**

Table 1: Outline of Corpus

Tools

Data in Table 1 were derived from CorpusTools (O'Donnell, 2010). CorpusTools is designed to accommodate SFL system diagrams (see figs 1-2). The analyst then divides a text into units of analysis, and manually assigns each unit a terminal description from the system diagram. CorpusTools also provides concordance lines and descriptive and comparative statistics. Data reported below were derived from CorpusTools 2.4.2.

Procedure

Texts from table 1 were stripped of visual **aids** and incorporated as text only as a single project into CorpusTools. System networks matching the grammatical categories in Figs 1-3 were specified, and the texts were divided into grammatical groups (verbal, nominal, prepositional, adverbial). Where possible, groups were assigned a terminal category in each network. The example in Table 2 reveals that although the unmarked pattern is for Participants to coincide with Theme or New Information, where Participants cannot be identified, Theme or New Information may be identified, or vice versa.

Text	When	data	is	purely descriptive	it	is said	qualitative data

						to be	
Theme - Rheme	Textual Theme	Topical Theme	Rheme		Topical Theme	Rheme	
Information Structure	← Given			New	← Given		New
Participant Identification		Presenting, Unmarked; Specified; No comparison			Presuming, Non-interlocuter; Specified; No comparison		Presenting, Unmarked; Specified; No comparison
Participant Tracking		Superordination, Complete repetition; Endophora: Preceding-anaphora; single			Superordination, substitution; Endophora: Preceding-anaphora; single		No referent – Addition

Table 2: Sample analysis of a Clause Complex (Buchla and McLachlan, 1992, p.36)

The clause analysed in table 2 reveals the unmarked pattern of placing an Additional (non-phoric) Participant with Presenting reference in the Rheme in New position in the clause complex (*qualitative data*). However, the analysis also reveals that the Theme of a clause (*When*) need not be a participant, that 'Given' positions may be occupied by Presented (*data*) or Presumed (*it*) Participants, and that the position for New information in the first clause has been taken by a Non-participant (*purely descriptive*) and so cannot be marked for reference. This sample analysis reveals differences to models of New information that rely on the realization of grammatical reference in order to assign information status (*e.g.* Prince, 1981; Gundel, 2010).

Results

The results presented here are a small sample of the many possible combinations by looking at the three systems (Moore, 2010). Typically, statistically significant results are selected so that their effect can be demonstrated in sample texts in the discussion section.

In Participant Identification, the unmarked choices in this corpus, in order of selection (Martin, 1992) are Effected, Specified, Variable, Nominal, Asserting and Undirected (see Fig. 1). According to the results obtained in this corpus, the probability of selecting their respective alternatives is very low: Neutralised (0.026%), Generalised (0.006%), Unique (0.0255%), Pronominal (0.127%), Directed (0.0468%), Question (0.1%) and Superset (0.06%). Halliday (1991; 1993) notes that the choices within systems may be skewed, so that one choice is far more likely than the other (it is, in linguistic terms, the unmarked option) and carries greater redundancy than the marked choice. That is, the probability score of 1.0 makes an utterance completely redundant – it is so likely that it did not need to be said.

Alternatively, grammatical choices may be equiprobable (Halliday 1991; 1993). For example, in the Theme-Rheme and New-Given systems, most results are approximately as likely as each other (except for Theme type). This is partly because Rheme and Given are defined as residues of Theme and New, respectively, although it is possible to have New without Given, and for Theme and Rheme to occur independently (such as in elliptical constructions). What matters in Theme and Rheme is not that they are chosen – in most clauses, at least one of the

pair is obligatory – what matters is the item that occupies that position; in the Theme and Information systems, choice, and therefore meaning, lies not in selecting Theme or New position but in selecting what is placed in that position.

Comparison across the different systems produces a range of significant results and selected results are provided here for illustration. Combining the two systems of Theme and Participant Identification, we find χ^2 (1) = 43.92, $p<0.001$ for Presuming reference in Theme position and Presenting reference in the Rheme. We also find χ^2 (1) =32.371, $p<0.001$ for Presuming Pronominals in Theme position compared to Presuming Nominals in New position in the clause. That is, even within Presuming reference, there is a significantly higher chance that a presumed participant will appear as a pronominal in Theme than a nominal in New position. This event does happen, but it appears marked when it does so because it is less likely. In contrast, a referent that cannot be tracked because it is Additional is significantly more likely to appear in New position than a Referent that can be tracked is likely to appear in Theme (χ^2 (1) =63.768, $p<0.001$).

Discussion

When comparing results across systems, we find examples of significant relationships that appear to have a correlation in the perception of ease of reading in a text. That is, the quantitative results of the systems proposed in the study have qualitative implications for readers. Using the sample results discussed above, among others, we can propose a typical pattern in written discourse where the reader would expect a Theme to contain Presumed, probably pronominal, participants, while the New position of a clause or clause complex would be taken up with a Presented participant. Where this does not happen it is likely to affect the reader's ability to follow the flow of discourse efficiently. For instance Version 1 of the text (the original) in Fig. 4 conforms to this pattern in a number of places: Presenting reference is placed in New position in the clause for *different properties* , *a particular record* , and *some field*. Similarly, Presumed participants occupy some of the thematic positions, such as *The record* and *In the latter case*. However, the New Information in the final clause is *quickly* while the Presented participants *hash table* and *a B-tree* are 'hidden' inside the clause. I would suggest that Version 2 of the text (the rewrite) provided below, with Presented Participants appearing in New position, conforms to an order where the clause focuses on presented information.

Version 1

On mainframes, however, **many types of files** *exist, each with* ***different properties.*** **A file** can be structured as ***a sequence of records,*** *for example,* ***with operating system calls to*** *read or write* ***a particular record.*** The record can usually be *specified* by giving *either its record number (i.e., position within the file) or the value of* ***some field.*** In the latter case, the operating system either maintains *the file as* **a B-tree** *or other suitable data structure,* or uses ***hash tables*** to locate *records quickly.*

Key

underlined	Theme
bold	Presented participants
lighter colour	Presumed participants
italics	New Information
font one size larger	clause final
font two sizes larger.	clause complex final

Version 2

On mainframes, however, **many types of files** *exist, each with* ***different properties.*** **A file** can be structured as ***a sequence of records,*** *for example,* ***with operating system calls to*** *read or write* ***a particular record.*** The record can usually be *specified* by giving *either its record number (i.e., position within the file) or the value of* ***some field.*** In the latter case, the operating system either locates *records quickly using* ***hash tables*** or maintains *the file as a suitable data structure such as* ***a B-tree***

Fig. 4: Sample text in original and modified form to show information flow (with key)

Conclusion

The flow of information from context through co-text to clause is managed in English through the simultaneous systems of Participant Identification and Tracking, Theme and Information. A corpus-based study which included the manual coding of approximately 10,000 words of academic text, revealed statistically-significant unmarked patterns that allow for a smooth progression within discourse. These patterns conform to a reader's perception of a smooth flow of information through discourse.

References

Beaver, D.I. 2004. The optimization of discourse anaphora. *Linguistics and Philosophy* 27/1, pp.3-56

Bhatia, V.K. 2004. *Worlds of Written Discourse*. London: Continuum

Botley, S. and McEnery, A.M. (Eds.). 2000. *Corpus-Based and Computational Approaches to Discourse Anaphora*. Amsterdam: John Benjamins

Caselli, T. and Prodanof, I. 2006. Annotating bridging anaphors in Italian: in search of reliability. LREC 2006. Retrived from http://pages.cs.brandeis.edu/~marc/misc/proceedings/lrec-2006/pdf/80_pdf.pdf on 20 Sep 2010

Clark, H.H. and Haviland, S. 1977. Comprehension and the Given-New contract. In R. Freedle (ed.), *Discourse Production and Comprehension*. New Jersey: Ablex

Cook, G. 1989. *Discourse*. Oxford: Oxford University Press

Crompton, P. 2004. Theme in discourse: 'Thematic progression' and 'method of development' re-evaluated. *Functions of Language* 11/2 p.213-49

Daneš, F. 1974. Functional sentence perspective and the organization of the text. In F. Daneš (ed.), *Papers on Functional Sentence Perspective*. Prague: Academia, pp.106-128

Emmott, C. 1997. *Narrative Comprehension - A Discourse Perspective*. Oxford: Clarendon Press

Foucault, M. 1972. *The Archaeology of Knowledge.* Abingdon: Routledge

Fries, P.H. 1981. On the status of Theme in English: Arguments from Discourse. *Forum Linguisticum 6/1* p.1-38

Fries, P.H. 2000. Issues in modelling the textual metafunction. In M. Scott and G. Thompson (eds.), *Patterns of Text: In honour of Michael Hoey*. Amsterdam: John Benjamins

Fries, P.H. 2002. The flow of information in a written text. In Fries, P., Cummings, M., Lockwood, D. and Spruiell, W. (eds.) *Relations and Functions Within and Around Language*. London: Continuum, pp.117-155

Fries, P.H. 2009. The textual metafunction as a site for a discussion of the goals of linguistics and techniques of linguistic analysis. In G. Forey and G. Thompson (eds.) *Text Type and Texture*. London: Equinox

Gee, J.P. 2008. *Social Linguistics and Literacies*. Abingdon: Routledge

Gundel, J.K. 2010. Reference and Accessibility from a Givenness Hierarchy Perspective. *International Review of Pragmatics*, 2/2, pp. 148-168

Halliday, M.A.K. 1967a. Notes on transitivity and theme in English. Part 2. *Journal of Linguistics* 3/2 pp.177-274

Halliday, M.A.K. 1967b. *Intonation and Grammar in British English*. The Hague: Mouton

Halliday, M.A.K. 1991. Corpus Studies and probabilistic grammar. In K. Aijmer and B. Altenberg (eds.) *English Corpus Linguistics*. Harlow: Longman, pp.30-43

Halliday, M.A.K. 1993. Quantitative studies and probabilities in grammar. In M. Hoey (ed.) *Data, Description, Discourse*. London: HarperCollins, pp.1-25

Halliday, M.A.K. and Hasan, R. 1976. *Cohesion in English*. London: Longman

Halliday, M.A.K. and Hasan, R. 1985. *Language, Context and Text: Aspects of Language in a Social-Semiotic Perspective*. Victoria: Deakin University Press

Halliday, M.A.K. and James, Z.L. 1993. A quantitative study of polarity and primary tense in the English finite clause. In J. Sinclair, M. Hoey and G. Fox (eds) *Techniques of Description*. London: Routledge, pp.32-82

Halliday, M.A.K. and Matthiessen, C.M.I.M. 2004. *An Introduction to Functional Grammar – Third Edition*. London: Arnold

Hughes, R. Carter, R. & McCarthy, M. 1994. Discourse Context as a Predictor of Grammatical Choice. In D. Graddol & S. Thomas (eds.) *Language in a Changing Europe*. Clevedon: Multilingual Matters, pp.47-54

Hymes, D. 1972. On Communicative Competence. In J.B. Pride & J. Holmes (eds.) *Sociolinguistics*. Harmondsworth: Penguin. pp.269-293

Karamanis, N., Mellish, C., Poesio, M. and Oberlander, J. 2008. Evaluating Centering for Information Ordering Using Corpora. *Computational Linguistics* 35/1 p.29-46

Martin, J.R. 1992. *English Text: System and Structure*. Amsterdam: John Benjamins

Martin, J.R. & Rose, D. 2007. *Working with Discourse. 2nd Edition*. London: Equinox

Martin, J.R. & Rose, D. 2008. *Genre Relations: Mapping Culture*. London: Equinox

Matthiessen, C.M.I.M. 1995. *Lexicogrammatical Cartography: English Systems*. Tokyo: International Language Science Publishers

Matthiessen, C.M.I.M. 2005. Frequency profiles of some basic grammatical systems: an interim report. In G. Thompson & S. Hunston (eds.) *System and Corpus - Exploring Connections*. London: Equinox, pp.103-142

McCarthy, M. 1991. *Discourse Analysis for Language Teachers*. Cambridge: Cambridge University Press.

Mitkov, R. 2000. Pronoun resolution: The practical alternative. In S. Botley and T. McEnery (eds.), *Corpus-based and Computational Approaches to Discourse Anaphora*. Amsterdam: John Benjamins, p.129-144

Mitkov, R., Lappin, S. and Boguraev, B. 2001. Introduction to the special issue on computational anaphora resolution. *Computational Linguistics 27/4* p.473-477

Moore, N. 2008. Bridging the metafunctions: Tracking participants through taxonomies. In C. Jones & E. Ventola (eds.) *From Language to Multimodality: New Developments in the Study of Ideational Meaning*. London: Equinox, pp.111-129

Moore, N. 2010. *Structuring Information in Written English*. Unpublished PhD Thesis, University of Liverpool.

O'Donnell, M. 2010. *UAM CorpusTool*. Software available from http://www.wagsoft.com/CorpusTool/download.html on 20 Jan 2010

Paltridge, B. 2006. *Discourse Analysis: An Introduction*. London: Continuum.

Prince, E. 1981. Toward a taxonomy of given-new information. In S. Cole (ed.) *Radical Pragmatics*. New York: Academic Press, pp.223-255

Stubbs, M. 1983. *Discourse Analysis*. Oxford: Basil Blackwell.

Taboada, M. and Zabala, L.H. 2008. Deciding on units of analysis within Centering Theory. *Corpus Linguistics and Linguistic Theory* 4/1, pp.63-108.