Textual Relations and Topic-Projection: Issues in Text Categorization

Samir Karmakar
Jadavpur University
samir.krmkr@gmail.com

Lahari Chatterjee
Jadavpur University
lahari.chatterjee@gmail.com

Abahan Dutta
Jadavpur University
abahanjiriya@gmail.com

Abstract

Categorization of text is done on the basis of its *aboutness*. Understanding *what a text is about* often involves a subjective dimension. Developments in linguistics, however, can provide some important insights about what underlies the process of text categorization in general and topic spotting in particular. More specifically, theoretical underpinnings from formal linguistics and systemic functional linguistics may give some important insights about the way challenges can be dealt with. Under this situation, this paper seeks to present a theoretical framework which can take care of the categorization of text in terms of relational hierarchies embodied in the overall organization of the text.

1 Introduction

Multiplicity of text is the consequence of the way textual components are selected and combined into coherent wholes, apart from the factors contributing to the content. One could be suspicious about this structure-centric investigation; but it is really hard to give up structure-centricity particularly in a position when structure is crucial both in formation and representation of the text.

Distinguishibility of one text from the other depends on what sorts of textual components are selected and how are they combined into the complex structure of a text. This complex weaving of *what*s and *how*s is often termed as `textuality` - the property because of which a text attains its uniqueness. Interpretation of text, therefore, arises through the gradual unpacking of textuality. Silverman (1994) argues, "[t]he interpretation of the text brings the textuality . . . outside the text, so as to specify and determine the text in a particular fashion."

Specification and/or the determination of a text in a particular fashion possess(es) a daunting challenge to linguistics in general and computational linguistics in particular. In linguistics, this finds its way through the study of discourse, text *etc.* (Beaugrande and Dressler, 1981); whereas in computational linguistics, interest is developed due the growth of text categorization, information structure *etc.* (Nomoto and Matsumoto, 1996). A careful investigation of these two lines will reveal the fact that their respective queries and approaches are built on the question of how a text is structured: Since a structure is the embodiment of different structuring principles out of which it is made up of, explicating the process through which a text comes into being remain a central concern. Note that textuality as an account of constituents and combinatorial principles is intrinsic to the text. Therefore, the questions of specification and/or determination of a text is translated into the way the respective textuality is.

Under this situation, then, this paper is interested in understanding what textuality is. One among many ways to investigate this question is to answer how the topic of a text is projected through the characteristic but hierarchical associations of its constituents. In other words, the projection of a topic in a text in some way brings out a nexus of text-internal relations holding among constituent statements of it. If so, then, topic-spotting in one sense is an act to categorize a text with an emphasis on textuality.

If we consider topic-spotting as the single most important criterion in categorizing a text, then the paper seeks to develop an analytical account of *how the weaving of statements into a network results into the projection of its topic*. This, in turn, plays a crucial role in identifying the way a text is categorized.

Since text and textuality are inseparable from each other, our task will be of two folds: *Firstly*,

S Bandyopadhyay, D S Sharma and R Sangal. Proc. of the 14th Intl. Conference on Natural Language Processing, pages 43–49,
Kolkata, India. December 2017.

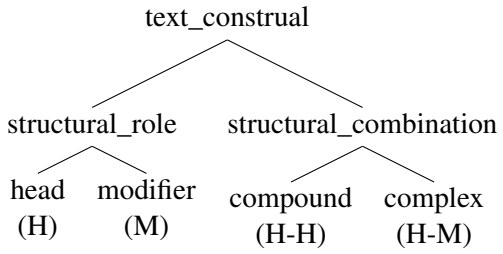
we need to come up with an analysis of the relations holding among the constituents of the text; and, *secondly*, how these relations are hierarchically organized in a text. Taken them togather, a general description of the textuality will evolve with a linguistic answer to the problem of topic-spotting.

Above mentioned two tasks will be performed over the news reports called brief. Broadsheet newspaper – generally designed in 8 columns – contains some brief news in the left most column. They are typically restricted into single column and consists of 5 to 10 sentences. The space, alloted to the briefs, is termed as Doric column for having symbolic resemblance with an archaic form of architectural order developed in Greece and Rome. A Doric column contains three to four brief news and they are ordered according to their significance which may vary from one news paper to another depending on the editorial policies. Briefs are structured.

2 Topic Projection and Textual Relations

To Taboada and Mann (2006), the communicative function of a written text is the consequence of the way words, phrases, grammatical structures and other relevant linguistic entities are getting involved into a coherent whole. In other words, a text-construal consists of certain structural roles performed by the constituents reflecting dependencies among the statements within a text and the way these roles are combined togather with the help of relations. As a result, following schematic representation of a text-construal is surfaced:

1.



As per this scheme, structural roles could be of two types namely (a) head (= H) and (b) modifier (= M). It is often noticed that head is projected as the topic. Relative saliencies of statements over each other often depends on their respective structural roles in a structural combination. For example, in a text-construal if the statements are connected with each other with the expressions - like 'and', 'or', 'but'*etc.* - chances are high for both the statements to enjoy the status of head (ex. *the moon was bright and the temperature was moderate*). This type of coordinating structure will be classified as compounded. In contrast to compounding, complex type structural combination distinguishes constituent clauses as either principle or subordinate (ex. *A guest is unwelcome when he stays too long*). Complexities involved in identifying topic of text-construal is often considered as the contribution of different statements within a network of textual relations.

In an investigation, where the topic-based categorization of text is being talked about, one need to know what is meant by topic. The topic of a text is its aboutness. More technically, topic is a statement which is entailed by the text. Following Djik (1977), we can formulate the following formal definition for the topic of a text - where text is a collection of statements in particular order.

2. A statement σ_i is the `TOPIC` of a sequence of statements $\Sigma = \langle \sigma_1, \sigma_2, \ldots, \sigma_n \rangle$ iff for each $\sigma_i \in \Sigma$ there is a subsequent Σ_k of Σ such that $\sigma_i \in \Sigma_k$ and for each successive Σ_k there exists σ_i such that $\Sigma_k \models \sigma_i$.

In other words, a statement will have the status of the topic, iff it is entailed globally by a text which it is a part of. By this, it is meant that all other local entailment relations will never have the status of topic in virtue of not being able to succeed across all subsequent collections of the statements to the ultimate text.

The most important criterion, *i.e.* the concept of entailment (represented with $\models$), for a statement to be the topic of a sequence of statements manifested through a text-construal needs some clarification: A statement *p* entails *q* when the truth of the first (*p*) guarantees the truth of the second (*q*), and the falsity of the second (*q*) guarantees the falsity of the first (*p*) (Saeed, 2009):

3.

p	$\models$	*q*
T	$\rightarrow$	T
F	$\rightarrow$	T or F
F	$\leftarrow$	F
T or F	$\leftarrow$	T

In continuation to our discussion, it could be said that a statement will have the status of topic in a text-construal, iff it is assigned to truth. As per composite truth table of entailment (3), the absolute truth of the entailed statement is subject to

the truth of a statement which is entailing the entailed. In other words, in order to have the status of topic, a statement must be in congruence with another statement in terms of the truth values. Note that the implementation of the criterion assumes a careful dissociation between *sentence* and *statement*. We will comment on this issue in our discussion below.

In a text, the other statements, acting as entailers, are used in modifying the entailed sentence (i.e. the topic) in various capacities - say for example elaborating, extending, enhancing etc. We will call them `relators`. It is not always necessary for a relator to be expressed explicitly in a text construal. What is worth mentioning is the fact that a statement with the status of topic is modified by modifiers in a structured manner. Having said this, it is emphasized that topic of a text hierarchically organizes various other functions with which it is associated. As an example consider the following piece from a brief published in Anandabazar on July 31, 2017:

4. (a) jhARkhaNDe bhArI briSTir AshangkA
Jharkhand.loc heavy rain.poss fear
nA thAkAy Dibhisir jal chArA
not having DVC.poss water release
kAmeche
reduce.perf.pres.3
Having no fear of heavy rain in Jharkhand, water release of DVC has reduced

(b) rAjya prashAsan tAi trAner
State administration therefore relief.poss
kAje bARti najar dicche
work-loc extra attention give.impf.pres.3
State administration, therefore, is giving more attention to the relief work.

Note that (4a) and (4b) both of them separately have their own topics: When (4a) is about the reduction in the release of water, (4b) is about paying extra attention to the flood situation. However, when these two statements are put togather in a text, the topic or aboutness of the resultant text is determined by the characteristic relation holding between (4a) and (4b). A careful look into (4) will reveal the fact that the topic of it is (4a). Being a topic, (4a) will enjoy the status of head (H) and (4b), the status of modifier (M). But what type of relation (4a) and (4b) are in? We can define the relation as `therefore`, because (4a) is reporting a situation that logically results into the situation reported in (4a). Moreover, (4b) contains an explicit lexical item *tai* in it to show how (4b) is modifying (4a). Further, reporting of (4a) is more central to reporter's purpose in putting forth the H-M combination than the reporting of (4b). Thus the relation could be termed as `therefore` - which licenses complex strucural combination. This discussion can be summarized in the following way:

5. $\texttt{therefore}(4a, 4b) \models 4a$
interpreted as ((4a), `therefore` (4b)) entails (4a) with a reference to a text-construal

To distinguish a statement from a sentence, we will introduce Greek alphabet σ with the provisions of using subscripts. Later on, it would be shown that a single sentence can have more than one statements in it. For this time, lets consider the statement expressed in (4a) is σ_i and the statement expressed in (4b) is σ_j. As a result, (5) will be converted into

6. $\texttt{therefore}(\sigma_i, \sigma_j) \models \sigma_i$
where `therefore` is a relator connecting a head statement with its modifier statement with a reference to a sequence of statements corresponding to a text-construal

The local relation(s) holding between two statements are getting modified when a third sentence is added with it. Consider the following sentence as the part of (4):

4. (c) rabibAr goghATe jal nAmleo
Sunday Goghat.loc water decreasing.prt
nadIgulir jalastar beshi thAkAy ghATAl
river.pl.poss water more having Ghatal
o khAnAkule teman nAmeni
and Khanakul.loc such decreasing.neg
On Sunday, in Goghat, though the water level decreased, in Ghatal and Khanakul no such change is noticed due to the high water level in the rivers.

Inclusion of (4c) will have its impact on the existing relational pattern because of effecting the distribution of roles and their combinatorial pattern in the resultant text-construal: (4c) represents a statement which is contradictory to the text-construal comprised of (4a) and (4b). This time no relation is explicitly mentioned. We will name this relation `contrarily` - since (4c) is providing an information which is contradicting with the previously stated information. Because of being compound type structural combination, resultant text-

construal will entail both (4a) and (4c): In compound type structural combination both the statements have the status of heads. As a consequence both (5) and (6) will be augmented or modified in the following ways:

7. $\texttt{contrarily}(\texttt{therefore}(4a, 4b), 4c)$
$\models \begin{cases} 4a \\ 4c \end{cases}$
Interpreted as,
(((4a) `therefore` (4b)), contrarily (4c)) entails 4a and 4c with respect to a text-construal

Conversion to the relational scheme of corresponding statements will give us the following result:

8. $\texttt{contrarily}(\texttt{therefore}(\sigma_i, \sigma_j), \sigma_k)$
$\models \begin{cases} \sigma_i \\ \sigma_k \end{cases}$ with respect to a sequence of statements corresponding to the text-construal

Note that the entailment relation is changed with the addition of newer statement. If this is a deviation from what is claimed in (2), then one should have some satisfactory answer to the question of how topics are licensed to percolate from one text-construal to its successive text-construals. No doubt, the answer to this problem has to come from the characteristic interactions holding between the structural aspect (= syntactic) and the meaning aspect (= semantic) of a text-construal. By strucural aspect, different combinations of H(ead) and M(odifier) are meant; whereas the meaning aspect is primarily concerned about the topic as well as entailment relations. The proposed solution to this problem will be explained in Section 4.

2.1 Sentence Internally Topic Projection

Though we are concerned about the topic spotting with a focus on the sequences of statements primarily at the sentential level, it is possible to trace back the topic from the sub-sentential level analysis - because a sentence can contain more than one statements. Therefore, to trace back our analysis from the subsentential level, we need to identify the subsentential constituents in the following way:

4(a) $\sigma_{i.1}$: jhARkhaNDe bhArI briSTir aAshangkA
Jharkhand.loc heavy rain.poss fear
nA thAkAy
not having
Having no fear of heavy rain in Jharkhand,

$\sigma_{i.2}$: Dibhisir jal chArA kameche
DVC.poss water release reduce.perf.pres.3
water release of DVC has reduced.

4(b) σ_j: rAjya prashAsan tAi trAner
State administration therefore relief.poss
kAje bARti najar dicche
work-loc extra attention give.impf.pres.3
State administration, therefore, is giving more attention to the relief work.

4(c) $\sigma_{k.1}$: rabibAr goghATe jal nAmleo
Sunday Goghat.loc water decreasing.prt
On Sunday, in Goghat, though the water level decreased

$\sigma_{k.2}$: nadIgulir jalastar beshi thAkAy
river.pl.poss water-level more having
having high water level in the rivers

$\sigma_{k.3}$: ghATAl o khAnAkule teman
Ghatal and Khanakul.loc such
nAmeni
decreasing.neg
in Ghatal and Khanakul no such change is noticed

Now, consider the case of (4a) which is a collection of following two statements: (i) there is no fear of heavy rain (= $\sigma_{i.1}$), and (ii) DVC is reducing the water release (= $\sigma_{i.2}$). Here in this case the former one is the modifier and the latter one is the head. First one is the reason for the second one. Alternatively, we can say, second one is the consequence of the first one:

9. $\texttt{consequently}(\sigma_{i.1}, \sigma_{i.2}) \models \sigma_{i.2}$

Similarly, (4c) as a complex sentence is made up of three distinct statements: (i) decreasing of the water level in Goghat region on Sunday ($\sigma_{k.1}$), (ii) having more water in the rivers (= $\sigma_{k.2}$), and (iii) not decreasing water levels in Ghatal and Khanakul regions (= $\sigma_{k.3}$). Here, (iii) is the head modified with (ii).

10. $\texttt{consequently}(\sigma_{k.2}, \sigma_{k.3}) \models \sigma_{k.3}$

Being in contrast with topic-projection of (10), the statement $\sigma_{k.1}$ will also have the status of head. As a result, along with $\sigma_{k.3}$, $\sigma_{k.1}$ will also be entailed by the resultant sequence of statements:

11. $\texttt{contrarily}(\sigma_{k.1}, \texttt{consequently}(\sigma_{k.2}, \sigma_{k.3}))$
$\models \begin{cases} \sigma_{k.1} \\ \sigma_{k.3} \end{cases}$

Later on, in Section 4, the rest of this story of topic projection will be presented. Here, in this point of our discussion, we would rather like to turn towards the questions of how a particular relation existing between two statements are identified, and how a statement is assigned to the topic. To address these issues, in Section 3, the underlying conceptual framework for topic extraction is explained.

3 Conceptual Framework for Topic Extraction

The process involved in the categorization of text in terms of its topic extraction, as is described in Section 2, can be conceptualized in the following way: As per our understanding, any text (like, brief) can be conceived as the sequence of statements. Each of these statements in isolation has a topic - no matter, how trivial it may sound. As a part of a text-construal, each one of them is related with some other statement. As is discussed earlier, (i) either one of the each pair has the status `head` and the other is the `modifier` (as in subordination), or (ii) both of them are of head status (as in coordination). Relator along with the concept of structural saliencies plays a crucial role in determining the topic of the text-construal made up of the constituent statements. Formally, a relator can be defined as,

12. $\mathfrak{R} : \mathfrak{S} \times \mathfrak{T}$ where $\mathfrak{S}$ is the set of statements and $\mathfrak{T}$ is the set of topics;

Assignment of a topic τ_i to a statement σ_i is a subjective task. This subjective dimension can be discussed in terms of a characteristic function:

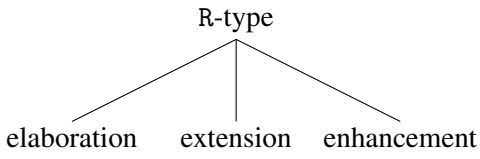
13. $\mathfrak{f} : \mathfrak{R} \rightarrow \{1, 0\}$

Before getting into our final section, lets have a look into the issues of relations.

3.1 Textual Relations

What seems to be central to the conceptual framework discussed in Section (3) is an account of different types of textual relations - for which we have relators of both implicit and explicit types. Few such relations - like `therefore`, `contrarily`, and `consequently` - are already mentioned in our discussion. Relations are important for the formation of complex statements. More specifically, relators are crucial in maintaining the coherence of a text. Three major types of relators will be discussed in section following the proposal of Systemic Functional Linguistics (Eggins, (Eggins, 2004)):

14.



These three types of relations are crucial in explaining the way text-construal incorporates the newer forms of information through the gradual increment of the relational network. While *elaborating* we are restating a statement for better clarity. In case of *extending*, additional statements are supplied; and the act of *enhancing* is used to indicate the further development of the meaning already communicated by a statement. Being the part of the typological classification of the relators, each of them are defined set-theoretically over a set of relators. In (15), these three types of relators are represented with their respective members:

15. (a) elaboration
`clarify`, `restate`, `exemplify`, `instantiate`, `illustrate`, `in_other_words`, `to_be_more_precise`, `as_a_matter_of_fact`, `actually`, `in_fact` etc.
(b) extension
`and`, `but`, `additionally`, `furthermore`, `moreover`, `excepting`, `apart_from_that`, `alternatively`, `on_the_other_hand`, `on_the_contrary`, `instead` etc.
(c) enhancement
`after`, `before`, `next`, `then`, `therefore`, `simultaneously`, `sequentially`, `until`, `since`, `now`, `similarly`, `yet`, `still`, `despite`, `though`, `consequently` etc.

The algebraic system underneath the process of topic sorting as a most important criterion to categorize a text, then, is a tuple $\langle \mathfrak{S}, \mathfrak{T}, \mathfrak{R}, \mathfrak{f} \rangle$ - which consists of a non-empty set of statements, a set of topics, a class of relations defined over the cartesian product of $\mathfrak{S}$ and $\mathfrak{T}$, and a characteristic function assigning each of the relations to

the set of values. In addition to the issues discussed above, what seems to be most crucial for the relations mentioned in (15) is their classification either as compound or complex type relations. More specifically, it is important to know which type of relations permit coordinate structures and which ones permit subordinate structures. From a gross observation, it seems the relations categorized as elaboration type and enhancement type permits subordination; whereas extension type relations are useful in constructing coordination.

4 Discussion

This section is dedicated to the hierarchical organization of the relations in a text-construal. This is done with the help of attribute value matrix. As per the conventions, set earlier, H and M are used to mark the organizational roles associated with the statements. Functional dependencies existing between the statements, along with their respective entailments, are represented with the aid of the relations. Apart from these, T and τ are used as subscripts to mark the status of the statements as topic. A statement marked with T is likely to percolate as topic to the next stage of construal as against the statement with subscipted τ confined within the text-construal which it is a part of. Note the identification of a topic (τ) either as a head (H) or as a modifier (M) is determined solely by the semantics of the relations. The topic is marked as T when it is identified as H. In addition to this, a concept of *rank* needs to be introduced here: If a matrix embeds another matrix in it, then the former one would be considered as of higher rank construal in comparison to the latter one. Consider the following examples:

16. matrix with higher rank with respect to (17):

$$\begin{bmatrix} \mathrm{H}_{\boxed{T}}{:}\texttt{consequently}(\sigma_{i.1},\sigma_{i.2}) \models \sigma_{i.2} & \begin{bmatrix}\mathrm{M}_{\boxed{\tau}}{:}\sigma_{i.1} \models \sigma_{i.1}\end{bmatrix} \\ & \begin{bmatrix}\mathrm{H}_{\boxed{T}}{:}\sigma_{i.2} \models \sigma_{i.2}\end{bmatrix} \end{bmatrix}$$

(16) is considered as a matrix of higher rank with respect to the matrices enumerated in (17):

17. matrices with lower ranks

(a) $\begin{bmatrix}\mathrm{M}_{\boxed{\tau}}{:}\sigma_{i.1} \models \sigma_{i.1}\end{bmatrix}$ (b) $\begin{bmatrix}\mathrm{H}_{\boxed{T}}{:}\sigma_{i.2} \models \sigma_{i.2}\end{bmatrix}$

However, (16) will be considered as a matrix with lower rank with respect to (18).

18. Topic percolation through the network of textual relations:

$$\begin{bmatrix} \Sigma & \boxed{1}\begin{bmatrix} \mathrm{H}_{\boxed{T}}{:}\mathfrak{t}(\sigma_i,\sigma_j) \models \sigma_i & \begin{bmatrix}\mathrm{H}_{\boxed{T}}{:}\mathfrak{c}(\sigma_{i.1},\sigma_{i.2}) \models \sigma_{i.2} & \begin{bmatrix}\mathrm{M}_{\boxed{\tau}}{:}\sigma_{i.1} \models \sigma_{i.1}\end{bmatrix} \\ & \begin{bmatrix}\mathrm{H}_{\boxed{T}}{:}\sigma_{i.2} \models \sigma_{i.2}\end{bmatrix}\end{bmatrix} \\ & \begin{bmatrix}\mathrm{M}_{\boxed{\tau}}{:}\sigma_j \models \sigma_j\end{bmatrix} \end{bmatrix} \\ & \boxed{2}\begin{bmatrix} \mathrm{H}_{\boxed{T}}{:}\mathfrak{C}(\sigma_{k.1},\mathfrak{c}(\sigma_{k.2},\sigma_{k.3})) \models \begin{matrix}\sigma_{k.1}\\ \sigma_{k.3}\end{matrix} & \boxed{3}\begin{bmatrix}\mathrm{H}_{\boxed{T}}{:}\sigma_{k.1} \models \sigma_{k.1}\end{bmatrix} \\ & \boxed{4}\begin{bmatrix}\mathrm{H}_{\boxed{T}}{:}\mathfrak{c}(\sigma_{k.2},\sigma_{k.3}) \models \sigma_{k.3} & \begin{bmatrix}\mathrm{M}_{\boxed{\tau}}{:}\sigma_{k.2} \models \sigma_{k.2}\end{bmatrix} \\ & \begin{bmatrix}\mathrm{H}_{\boxed{T}}{:}\sigma_{k.3} \models \sigma_{k.3}\end{bmatrix}\end{bmatrix} \end{bmatrix} \end{bmatrix}$$

For the sake of the brevity and the ease of the presentation, in (18), we have used following abbreviations: $\texttt{therefore} = \mathfrak{t}$, $\texttt{consequently} = \mathfrak{c}$, $\texttt{contrarily} = \mathfrak{C}$.

Needless to say, matrices mentioned in (17) would be of *lowest* rank because of not embedding any other matrix; on the other hand, (18) will be of *greatest* rank in virtue of not being embedded in other matrix. Significance of relative ranks is useful in explaining how the projection of topic is taking place within the text-construal.

As per attribute value matrix of (18), the topic of a sequence of statements, Σ corresponding to a text construal (4), will be that statement which is embedded in all successive matrices of higher ranks as head (= H). In other words, topic of a modifier is not licensed to be the topic of a matrix with higher rank within which the modifier is embedded. As per this assertion, then, it is not hard to argue why the topics of the lowest rank text-construals as modifier fail to percolate as a topic in the text-construals which are in immediately higher ranks.

As the analytical framework outlined and discussed above, the text mentioned in (4) have two distinct topics:

19. $\Sigma \models \begin{cases} topic_{4a-b}\ marked\ as\ 1 \\ topic_{4c}\ marked\ as\ 2 \end{cases}$

These two topics are projected by two different text-construals which are in equal rank as is obvious from the attribute-value matrix of (18). Similar situation prevails in case of the matrices marked with 3 and 4. If two topics are of equal rank, then chances are high for the respective texts to be independent of each other. In other words, two topics with equal rank are combined togather into a text-construal with the aid of those relations which are crucial in producing compounded structures.

On the basis of this discussion, we can argue that Dijk's criterion (1977) for topic identification

mentioned in (2) seeks the following modification: Any text as a sequence of statements will have one and only one topic iff the constituent textual relations are in complex type structural combination. Compounding of constituent relations will indicate their respective projections as independent of each other. In such a situation, the text-construal can be broken into two independently occurring texts. This could be used as a potential clue to the auto-editing of briefs in particular and news reporting in general.

References

R. A. De Beaugrande and W. U. Dressler 1981. *Introduction To Text Linguistics* Longman, New York.

M. Taboada and W. C. Mann 2006. Rhetorical Structure Theory: Looking Back and Moving Ahead *Discourse Studies*, 8(3), pp. 423-459.

T. A. van Djik 1977. *Text and Context: Explorations in the semantics and pragmatics of discourse* Longman, London.

J.I. Saeed 2009 *Semantics* (3rd edition) Wiley-Blackwell, Oxford.

H. J. Silverman. 1994. *Textualities: Between Hermeneutics and Deconstruction* Routledge, London.

T. Nomoto and Y. Matsumoto 1996 Exploiting Text Structure For Topic Identification *Workshop On Very Large Corpora*, 101-112

S. Eggins 2004 *An Introduction to Systemic Functional Linguistics* Second Edition Continuum, New York