Locating Topics in Text Processing

Eleni Miltsakaki
Univeristy of Pennsylvania

Abstract

In this paper we are concerned with the location of topics in text processing and the determination of the update unit in looking up topic continuations and topic shifts. Using key elements of the Centering Model of local discourse coherence and empirical evidence from Modern Greek and Japanese we argue that the appropriate update unit for topic tracking is the sentence in its traditional sense and not the finite clause, thus providing an account for the status of the subordinate clause in the calculation of topic transitions. We bring forth an argument from English, Modern Greek (MG) and Japanese for keeping topic and information structure distinct. We briefly discuss the significance of the current work to automated essay scoring and coreference-based summarization systems.

1 Introduction

This paper is concerned with the issue of identifying the location of topics in text processing. Adopting the framework of the Centering Model, we discuss the importance of defining the appropriate update unit in tracking topics and topic shifts in discourse. Based on empirical findings from Modern Greek and Japanese we argue that the prominent location of topics is the main clause (the highest ranked entity in the main clause, see section 2). We define the update unit as the sentence in the standard grammatical sense, contra Kameyama (1998) who treats tensed adjunct clauses as independent update units. We briefly discuss the role of the location of subordinate clauses relative to main clauses in English, Modern Greek and Japanese and suggest that topic structure and information structure are two distinct aspects of text processing. We argue that it is precisely blurring this distinction that has misled topic identification efforts. Finally, we bring further evidence from a study on automated essay scoring (Miltsakaki and Kukich (2000a) and offer some pointers to the potential benefits for coreference based summarization systems.

The paper is structured as follows. Section 2 provides an overview of the Centering framework. In section 3, we discuss previous work related to the determination of the update unit. In section 4, we motivate and present our definition of the update unit for the purposes of identifying topic structure. We discuss examples from English, Modern Greek and Japanese. In section 5, we discuss the results of a Centering study in MG. In the light of these results we reevaluate the arguments presented in section 3 in favour of earlier formulations. In section 6 we discuss further evidence from an automated essay scoring system and the potential significance of the current work to summarization systems. We conclude in section 7.

2 The Centering Model

Centering was originally proposed as a model of the complexity of inferencing involved in discourse when speakers process the meaning of an utterance and integrate it into the meaning of the previous discourse (Joshi and Kuhn (1979), Joshi and Weinstein (1981)). From a different perspective, Grosz and Sidner (Sidner (1979), Grosz (1977), Grosz and Sidner (1986)) identified the 'attentional state' as a basic component of discourse structure and proposed that it consisted of two levels of focusing: global and local. For Grosz and Sidner, Centering Theory provided a model for monitoring local focus. A synthesis of these two approaches yielded a Centering model which was designed to account for the difference in the perceived coherence of discourses such as in (1) and (2) below (examples from Hudson-D'Zmura (1988)).

(1) a. John went to his favorite music store to buy a piano.
b. He had frequented the store for many years.
c. He was excited that he could finally buy a piano.
d. He arrived just as the store was closing for the day.

(2) a. Josh went to his favorite music store to buy a piano.

b. It was a store John had frequented for many years.

c. He was excited that he could finally buy a piano.

d. It was closing just as John arrived.

Discourse (1) is intuitively more coherent than discourse (2). Discourse (1) centers a single individual whereas discourse (2) seems to flip back and forth among several different entities. The perceived difference is seen to arise from the varying degree of continuity in the topic structure of the discourse.

We now turn to the basic components of the Centering Model.

2.1 Segments and entities

Discourse consists of a sequence of textual segments and each segment consists of a sequence of utterances. In Centering Theory, utterances are designated by $U_i - U_n$. Each utterance U_i evokes a *set* of discourse entities, the FORWARD-LOOKING CENTERS, designated by $Cf(U_i)$. The members of the Cf set are ranked according to discourse salience (the ranking is given in Section 2.3). The highest-ranked member of the Cf set is the PREFERRED CENTER, Cp. A BACKWARD-LOOKING CENTER, Cb, is also identified for utterance U_i. The highest ranked entity in $Cf(U_{i-1})$ *realized* in U_i is called the BACKWARD-LOOKING CENTER, Cb. The BACKWARD-LOOKING CENTER is a special member of the Cf set because it represents the discourse entity that U_i is about. The BACKWARD-LOOKING CENTER can be seen as the Centering version of what in the literature is often called the 'topic' (Reinhart (1981), Horn (1986)).

The Cp for a given utterance may be identical with its Cb, but not necessarily so. In fact, the computation of local coherence in discourse is dependent on the distinction between looking back in the discourse with the Cb and projecting preferences for interepretations in the subsequent discourse with the Cp.

2.2 Centering transitions

Four types of transitions, reflecting four degrees of topic continuity, are defined in Centering. They are computed as shown in Table 1 and ordered according to the ordering rule in (3).

(3) Transition ordering rule:
Continue is preferred to Retain, which is preferred to Smooth-Shift, which is preferred to Rough-Shift.

	Cb(Ui) = Cb(Ui-1)	Cb(Ui) ≠ Cb(Ui-1)
Cb(Ui) = Cp(Ui)	Continue	Smooth-Shift
Cb(Ui) ≠ Cp(Ui)	Retain	Rough-Shift

Table 1: Table of transitions

2.3 Cf ranking

As mentioned earlier, the PREFERRED CENTER of an utterance is defined as the highest ranked member of the Cf set. The ranking of the Cf members is determined by the salience status of the entities in the utterance and may vary cross-linguistically. It has been proposed (Kameyama (1985), Brennan et al. (1987))that the Cf ranking for English is determined by grammatical function as follows:

(4) Ranking of forward-looking centers:
SUB>IND/OBJ>OBJ>OTHERS

Later cross-linguistic studies based on empirical work (Di Eugenio (1998), Turan (1995), also Kameyama (1985)) determined the following more refined ranking, with QIS standing for quantified indefinite subjects (e.g. 'many people', 'every boy') and PRO-ARB for arbitrary plural pronominals (e.g. '*We* should respect human rights').

SUB>IND/OBJ>OBJ>OTHERS>QIS,
PRO-ARB

As regards the ranking of entities within complex NPs (e.g. his mother, software industry), the working hypothesis is that they are ranked from left to right (e.g. Walker and Prince (1995)).

2.4 Utterance: the update unit

In the earlier formulation of Centering the length of the *utterance* (henceforth the update unit) was not defined explicitly. While it was clear that a unit was needed for updating the Cf list and the Cb assignment, the precise size of this unit was left undetermined.

Before Kameyama (1993), the update unit was informally understood to be the tensed clause. Kameyama (1993) also defined it as, roughly, the tensed clause with the exception of relative clauses and clausal complements which she argued were part of the update unit containing the matrix clause. We will discuss this issue extensively in the next section.

3 Related work

The Centering model has been used and modified by many researchers working on the interpretation of pronominals (anaphora resolution). Anaphoric elements occur in all types of clauses. Naturally, utterance level issues soon became central in the development of algorithms for anaphora resolution. It was crucial that the size of units were constrained to enable handling intrasentential anaphora. As a result, defining the appropriate unit was often dictated by needs specific to anaphora resolution algorithms.

However, if we maintain Centering as a model of local discourse coherence it would be desirable to ensure that the proposed modifications yield transitions that reflect our intuitions about perceived discourse coherence as well as the degree of the processing load required by the hearer/reader at any given time in discourse processing. Modifying Centering definitions in the interest of a successful anaphora resolution algorithm is a valid effort in its own right but one that is orthogonal to modifying Centering definitions to reflect discourse coherence. Successful anaphora resolution algorithms do not need to account for the processing load involved in text processing. Language users have at their disposal a wealth of resources in resolving anaphora successfully but the processing cost may vary in each case. Simply put, some anaphoric elements are easier to resolve than others and anaphora resolution algorithms strive to handle all such cases equally well.

We now turn to Kameyama (1993) (also Kameyama (1998)) who was concerned with the problem of intrasentential Centering and, in particular, the definition of the appropriate update unit when processing complex sentences. Roughly, her suggestion was to break up complex sentence according to the following hypotheses: conjoined and adjoined tensed clauses form independent units whereas tenseless subordinate clauses, report complements and relatives clauses belong to the update unit containing the matrix (superordinate) clause.

Let us now turn our attention to the tensed adjunct hypothesis which is our major concern here. Kameyama brings support of her hypothesis from backward anaphora. She argues that, with respect to backward anaphora, the tensed adjunct hypothesis predicts that the pronoun in the fronted adjunct clause is anaphorically dependent to an entity already introduced in the immediate discourse and not to the the subject of the main clause it is attached to:

(3) NULL:[1] Kern began reading a lot about the history and philosophy of Communism

(4) ESTABLISH (Cb=Jim Kern): but never 0 felt there was anything he as an individual could do about it.

(5) CHAIN (Cb=Jim Kern) When he attended the Christina Anti Communist Crusade school here about six months ago

(6) NULL: Jim became convinced that an individual can do something constructive in the ideological battle

(7) ESTABLISH (Cb=Jim Kern): and 0 set out to do it.

The above argument is weak in two respects. First it is not empirically tested that in cases of backward anaphora the antecedent is found in the immediate discourse. As counter-evidence we present a naturally occurring example, (8), taken from an e-mail correspondence from the organizer of a reading-group at University of Pennsylvania. Clearly, the antecedent of *he* in (8) cannot be identified in the immediate previous discourse. The referent of the pronominal *he* is identified in the subject of the subsequent main clause.

(8) There has been a slight change of plan since I just today realized how late in the month it is. Because **he** will be leaving us soon, **Yuji** has kindly agreed to talk to us tomorrow about some aspects of what he will present this weekend at the PLC. If there is time ...

Such examples are not a problem in our account. Backward anaphora can easily be resolved to the subject of the main clause, in fact the Cp in both (5)-(6) and (8).

Secondly, this account leaves the use of a full NP in Kameyama's main clause (6) unexplained. Full NPs occurring in Continue transitions have been observed to signify a segment boundary. Assuming that segment boundaries do not occur between a main clause and a subordinate clause associated with it, the use of a full NP in (6) remains puzzling.

Empirical evidence in support of Kameyama's hypothesis that tensed adjunct clauses should be treated as independent processing units comes from Di Eugenio (1990) and Di Eugenio (1998). Di Eugenio carried out Centering studies in Italian. Before discussing her evidence some background on her studies are in order.

In Italian (as in MG) there are two pronominal systems: weak pronouns that must cliticize to the verb and strong pronouns that are syntactically

[1] We present here Kameyama's own example retaining the terminology of the system she proposes.

similar to full NPs. Null subjects are allowed and belong to the system of weak pronouns.

Di Eugenio (1990) proposed that the alternation of null and overt pronominal subjects could be explained in terms of centering transitions. Typically, a null subject signals a CONTINUE, and a strong pronoun a RETAIN or a SHIFT. [2]

Di Eugenio (1998) tested her earlier conclusions on naturally occurring Italian data. Following Kameyama (1993) she treated tensed adjuncts as independent update units. Her motivation for doing so came from the following example where the use of a strong pronoun in the main clause cannot be explained if the preceding adjunct is not treated as an independent update unit. The translation (taken from Di Eugenio (1998)) is literal but not word for word. For the utterance preceding (4) the Cb(Ui-1)=vicina-j (neighbor-fem) and Cf(Ui-1)=vicina-j.

(9) Prima che i pigroni-i siano seduti a tavola a far colazione,
'Before the lazy ones-i sit down to have breakfast,'

(10) lei-j e via col suo-j calessino alle altre cascine della tenuta.
'she-j has left with her-j buggy for the other farmhouses on the property.'

We will further discuss this example and offer an alternative explanation in section 5 i the light of the conclusions drawn from the MG Centering study.

Suri et al. (1999) address the problem of developing and assessing algorithms for tracking local focus and proposing referents for anaphora resolution. In particular, they are concerned with the appropriate treatment of complex sentences and report results with regard to sentences of the form “SX because SY” where SX and SY are simple sentences. They propose what they call the SSD (Semantically Slanted Discourse) Methodology to test how an anaphora resolution algorithm should be extended to capture the focusing structure pertaining to complex sentences. They apply their SSD Methodology to extend their RAFT/RAPR algorithm. The algorithm prefers to resolve a subject pronoun in a simple sentence so that it corefers with the Subject Focus of the previous sentence. To address the question of how to process “SX because SY” sentences they constructed discourses of the form:

(S1) simple sentence
(S2) SX because SY
(S3) simple sentence

Based on their quantitative results they propose the 'Prefer SX Hypothesis' as an extension to their anaphora resolution algorithm. This is because their results show that a subject pronominal in (S3) is resolved to the subject of SX independently of the form and referent of the subject in SY. When semantic/pragmatic reasons dictate that such resolution is not plausible the discourse was judged infelicitous by their subjects. The relevant discourses and judgements are given below:

Discourse 1
(S1) Dodge was robbed by an ex-convict the other night.
(S2) The ex-convict tied him up because he wasn't cooperating.
(S3) Then he took all the money and ran.

Discourse 2
(S1) Dodge was robbed by an ex-convict the other night.
(S2) The ex-convict tied him up because he wasn't cooperating.
(S3) #Then he started screaming for help.

Notice that their results are exactly compatible with our hypothesis about the appropriate update unit without having to stipulate a special 'extension.' In our approach, “SX because SY” forms a single update unit with the subject of SX occupying the favourite focus(topic) position. In section 5 we argue that, roughly, topic shifts must be established in the main clauses. It is therefore not surprising that a pronominal subject in (S3) will resolve to the referent of the subject in SX. If the writer wanted to shift focus (or retain an old one) to an entity introduced in SY s/he should opt for a full NP (or special stress in spoken discourse).

Discourse 3
(S1) Dodge was robbed by an ex-convict the other night.
(S2) The ex-convict tied him up because he wasn't cooperating.
(S3) Then Dodge started screaming for help.

[2] Di Eugenio collapsed the distinction between Smooth and Rough Shifts. However, the reader is referred to Miltsakaki and Kukich (2000b) for a discussion of the special role of Rough-Shifts with respect to data where text coherence is under evaluation and therefore cannot be assumed. Miltsakaki and Kukich discuss data from writing tests evaluating students' essay writing skills.

4 Redefining the update unit

In this section we propose that in identifying the topic structure of texts the relevant unit in which topics are located is the sentence, i.e. the unit containing the matrix clause and all the dependent (i.e. subordinate) clauses associated with it. The Cf list contains all the entities evoked in the sentence, the most salient of which is the subject of the main clause. It follows that the entities evoked in the subordinate structures are less salient, assuming a more salient role only when they are promoted to the higher positions of the Cf list of the following unit in accordance with the definition of the Cb.

At a preliminary stage of the Centering study in MG (section 5) we adopted Kameyama's hypotheses. The update unit was soon identified erroneous as it yielded a highly and counter-intuitively incoherent MG discourse. This was noted especially with regard to time adjuncts. Consider the discourse spanning over (10) and (13) (taken from the MG data).

(11) Ki epeza me tis bukles mu
and I-was-playing with the curls my

'And I was playing with my hair.'
Cb=I, Cp=I, Tr=Continue

(12) Eno ekini pethenan apo to krio
while they were-dying from the cold

'While they were dying from the cold,'
Cb=none, Cp=THEY, Tr=Rough-Shift

(13) Ego voltariza stin paralia
I was-strolling on-the beach

'I was strolling on the beach.'
Cb=NONE, Cp=I, Tr=Rough-Shift

(14) Ki i eforia pu esthanomun
and the euphoria that I-was
den ihe to teri tis
feeling not have the partner its

'And the euphoria that I was feeling was unequalled.'
Cb=I, Cp=EUPHORIA, Tr=Rough-Shift

If we treat the adjunct clause in (12) as an independent unit, the resulting transitions (3 Rough-Shifts) yield a highly incoherent discourse, contra the actual perceived coherence. The picture changes dramatically if we treat (12) and (13) as one unit. The resulting transitions are Continue-Continue-Retain.

It turns out that the tensed adjunct hypothesis is not problematic to the Modern Greek text alone. Consider the constructed example from English shown in Table 2.

John had a terrible headache. When the meeting was over, he rushed to the pharmacy store	
John had a terrible headache	
Cb	?
Cf	John>headache
Tr	none
When the meeting was over	
Cb	none
Cf	meeting
Tr	Rough-Shift
He rushed to the pharmacy store	
Cb	none
Cf	John>store
Tr	Rough-Shift

Table 2: Sequence: main-subordinate-main

Allowing the subordinate clause to function as an update unit yields a highly incoherent discourse in English (two Rough-Shifts), reflecting a high degree of discontinuity counter to the perceived coherence of the discourse in Table 2. If indeed there are two Rough-Shift transitions in this discourse the use of the pronominal in the third unit is puzzling. In addition, reversing the order of the clauses, as shown in Table 3, results in a highly coherent discourse, in sharp contrast with the discourse of Table 2. Assuming that the two discourses demonstrate a similar degree of continuity in the topic structure (they are both *about* 'John', we would expect the transitions to reflect this similarity when, in fact, they do not.

John had a terrible headache. He rushed to the pharmacy store as soon as the meeting was over.	
John had a terrible headache	
Cb	?
Cf	John>headache
Tr	none
He rushed to the pharmacy store	
Cb	none
Cf	John
Tr	Continue
as soon as the meeting was over	
Cb	John
Cf	none
Tr	Retain

Table 3: Sequence:main-main-subordinate

We conclude that the introduction of a new discourse entity, 'meeting' in this case, in the time-clause does not interfere with the topic structure of the discourse nor does it project a preference for a shift of topic, as the Cp normally does when it instantiates an entity different from the current Cb.

Further evidence in support of our definition of the update unit and our hypothesis about the location of topics comes from Japanese.[3] In Japanese, topics and subjects are lexically marked (wa and ga respectively) and null subjects are allowed. Note that subordinate clauses must precede the main clause. Consider the Japanese discourse (16)-(18). Crucially, the referent of the null subject in the second main clause resolves to the topic marked subject of the first main clause, ignoring the subject-marked subject of the intermediate subordinate clause.

(15) Taroo wa tyotto okotteiru youdesu
Taroo TOP a-little upset look
'Taroo looks a little upset.'

(16) Jiroo ga rippana osiro o
Jiroo SUB great castle OBJ
tukutteiru node
is-making because
'Since Jiroo is making a great castle,'

(17) ZERO urayamasiino desu
ZERO jealous is
'(He-Taroo) is jealous.'

It is also worth mentioning that cataphoric structures (backward anaphora) are far more common in subordinate rather than paratactic structures shown in (19), with the exception of certain modal contexts, shown in (20) [4]

(18) As soon as he arrived, John jumped into the shower.

(19) #He arrived and John jumped into the shower.

(20) He-i couldn't have imagined it at the time but John Smith-i turned out to be elected President in less than 3 years.

Assuming that the position of the subordinate clause does not affect the topic structure, we would like to ask ourselves what determines, if anything, the linear position of the relative clause. In what follows, we give a brief outline of a tentative explanation.

Let us first focus on the role of dependent (i.e., subordinate) clauses. Traditional grammar books classify clauses into two main categories, namely dependent and independent clauses. Independent clauses are matrix clauses. Dependent clauses come in three flavors: adverbial, nominal and relative. Nominal clauses hold complement positions with respect to the verb of the matrix clauses and adverbial clauses are tensed adjunct clauses, in Kameyama's terminology. Informally, dependent clauses are understood as tools used by the speaker/writer to add or specify information in relation to the proposition expressed in the main clause and are therefore anaphorically dependent on the main clause. If this is true, then it is not surprising that they do not interfere with the topic structure of the discourse.

Still, we haven't answered the question about their linear order with respect to the main clause. Let us, briefly, turn our attention to the surface word order within a single clause. It is commonly assumed that for each language there is an underlying canonical order of the basic constituents. In an SVO language like MG, the canonical order of the verb and its arguments is subject-verb-object. This, of course, is not always the attested surface order. In syntactic theories, it is commonly assumed that surface word order is derived by various movement operations. Some movement operations are dictated by the syntax of each language and are necessary to yield grammatical sentences. It is also common, however, especially in free word order languages, that movement is syntactically optional and the surface word order is used to satisfy information packaging needs (for example to arrange the information into old-new, ground-focus etc). Note that when this happens, it is only the surface word order that is altered and not the basic relation of the arguments to the predicate. To give an example from English, in (15) the internal argument of the verb (the object) has been fronted but its original relation to the verb has remained the same.

(21) Chocolate Mary hates.

Moving to the sentential level, we entertain the hypothesis that the same principle dictates the position of the clauses relative to each other. Each dependent clause stands in a specific relation to the main clause and this relation is not altered by the order in which the clause appears on the surface.

Conceptually, this approach is similar to the the discourse LTAG treatment of subordinate conjunctions. In discourse LTAGs subordinate conjunctions are treated as predicates, anchoring initial trees containing the main and the subordinate clause as arguments. Each subordinate conjunction may anchor a family of trees to reflect

[3] Thanks to Kimiko Nakanishi for providing me with the data. In a Centering study she conducted in Japanese she also concluded that treating subordinate clauses as independent units would yield a highly incoherent Japanese discourse.

[4] Thanks to Ellen Prince for pointing out this example.

variations of the surface order of the substituted argument clauses but the predicate argument relation remains the same. (Webber and Joshi (1998), Webber et al. (1999a), Webber et al. (1999b)).

How does the above discussion relate to the definition of the Centering update unit? Recall that the Centering model keeps track of the topic structure. In other words it keeps track of discourse salience. If we dissociate salience from information structure the relevant unit for calculating salience is at the sentence level, horizontally (see Figure 1). The relative order of independent/dependent clauses is determined by information structuring, a process orthogonal to the computing of salience. Subordinate links are not relevant to the salience mechanism. Salience is calculated paratactically.

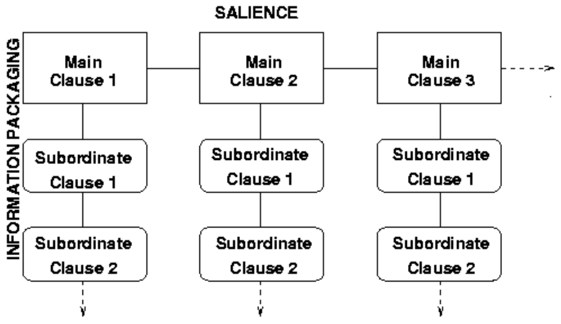

Figure 1:

A corollary of such a model is that you can introduce referents in the vertical level without affecting the status of the salient entity on the horizontal level. It follows that changes of topic must be established at the horizontal level.

Such a conception of the salience structure suggests that text processing is not strictly incremental as commonly assumed. The Cf list may be constructed incrementally but the final ranking is determined only after the sentence is complete.

To wrap up, in this section we looked at three very different languages, namely English, Modern Greek and Japanese and we concluded that in all three languages the surface order of tensed adjunct clauses (tensed subordinate clauses in our terminology) relevant to the main clause does not affect the topic structure of the discourse. We entertained the hypothesis that the surface order of clauses is determined by information structure needs so that the information is appropriately arranged with respect to the current discourse model of the hearer (e.g old-new). The distinction between information and salience structure is often blurred by the inevitable overlap between information status and salience as topics themselves tend to be discourse old. Obviously, more work is needed in order to determine the nature of the interaction between discourse salience and information structure. However, we showed that there is a significant gain in better understanding discourse structure by keeping the two processes distinct.

5 Centering in Modern Greek

The text used in this study is a Modern Greek novel titled *I won't do this favor for you* by C. A. Chomenidis, a young Greek novelist. Its length amounts to 6,000 words and was chosen for its richness in pronominal tokens (393 in total excluding epithets).

5.1 Establishing the Cf ranking

To establish the Cf ranking in MG we use Rambow (1993)'s diagnostic. We show that in MG, as in Turkish (Turan (1998)) surface word order does not contribute to the Cf ranking. The relevant indicator of salience in the Cf list is subjecthood. Consider examples (22) and (23). Note that the null pronominal in (22b) and (23b) resolves to the subject irrespective of its surface position. Gender and lexical considerations are controlled. Both *economical policy* and *arrangement* are feminine and they can both be *inadequate*.

(22) a. I prosfati diefthetisi-i tha
the recent arrangement will
veltiosi tin ikonomiki politiki-j?
improve the economic policy?
'Will the recent arrangement improve the economic policy?'

b. Ohi, (null-i) ine aneparkis.
No, (it) is inadequate.
'No, it is inadequate.'

(23) a. Tin ikonomiki politiki-j tha ti-j
the economic policy will CL-it
veltiosi i prosfati diefthetisi-i?
improve the recent arrangement?
'Will the recent arrangement improve the economic policy?'

b. Ohi, (null-i) ine aneparkis.
No, (it) is inadequate.
'No, it is inadequate.'

The examples above indicate that in MG subjects rank higher than non-subjects. Also, in MG, as in Turkish, a strong pronominal or a full NP must be used if the object of Ui-1 gets promoted to the subject position of Ui. We take this a further evidence that objects rank lower than subjects. The following example demonstrates the point:

(24) O Yannis-i proskalese ton Yorgo-j.
the John invited the Yorgo.

'John invited George.'

a. null-i tu-j prosfere ena poto.
he him offered a drink.
'He-i offered him-j a drink.'

b. #null-j #tu-i prosfere ena poto.
he him offered a drink.
'He-j offered him-i a drink.'

c. O Yorgos tu-i prosfere ena poto.
the George him offered a drink.
'George offered him-i a drink.'

d. Ekinos-j tu-i prosfere ena poto.
he-strong him offered a drink.
'HE-j offered him-i a drink/'

5.2 Transitions

We follow the standard Centering transitions as defined in section 2 with the following modification. In cases of Cb(Ui-1)=none and Cp(Ui-1)=Cp(Ui) we coded the transition as Continue to reflect the intuition that when no Cb can be identified at a discourse medial Ui-1 but the Cp of Ui-1 equals the Cp of Ui the transition in question is highly coherent. The Cp in Ui-1 establishes a new center and then in the following Ui the transition keeps the same center in the privilleged Cp position. Also, note that, unlike other Centering studies, we chose not to pre-segment the text. This decision was made in order to avoid arbitrariness given that the notion of segment is not well understood and consequently not clearly defined in the Centering literature.

5.3 Annotation

A total of 474 units, as defined here, were coded and for every two consecutive utterances the transitions were calculated. In each unit the elements of the Cf list were coded as shown in Table 4.

CODE	GLOSS
null	null pronominal (subject)
weak	weak pronominal
poss	weak possessive
null-Q	quantified indefinite phrase is realized as null pronominal
weak-Q	quantified indefinite phrase is realized as weak pronominal
full	full noun phrase
strong	strong pronominal
full/strong-poss	possesive with full NP or emphatic anaphoric element
epithet	epithet

Table 4: Codes and glosses

5.4 Results

Table 5 shows the distribution of form selection over Centering transitions. As expected, null and weak forms predominate in continue transitions.

	C	R	S-S	R-S
null	207	22	51	22
weak	21	2	4	4
poss	23	2	5	5
total	251	26	60	31
full	2	12	3	57
strong(+poss)	9	3	5	8
epithet	1	1	1	4
total	12	16	9	69

Table 5: Transitions over non-segmented text

However, note that the number of Rough-Shifts with weak forms is surprisingly high. Table 6 shows the distribution of Rough-Shifts in detail.

focus pops	19
Ui-1 overt argument missing	2
two character scenes	3
other	6

Table 6: Classification of Rough-Shifts

As 'focus-pops' we classified instances of Rough-Shifts at the boundaries of switching mode of writing from dialogue to narrative and vice-versa and at the boundaries of parenthetic setting-descriptions. Recall that the text was not pre-segmented. These results provide support for the kind of focus-pop theory developed by Grosz and Sidner (1986). We will not discuss this result any further as it is not central to the concerns of this paper.

Another interesting result is the distribution of strong pronominals: the number of Continue transitions is surprisingly high. In fact, the difference in the distribution of Continue and Rough-Shift transitions is insignificant. In Table 7 we have classified the instances of strong pronominals associated with continue transitions.

	poset	relative	emphasis	other
strong	6	1	1	1

Table 7: Strong pronouns and Continue transitions

Under 'poset', partially ordered set, we have classified instances where the relevant entities stand in what is commonly described as a 'contrast' relationship to some other entity in the discourse. Here, we follow Prince (1981) who argues that 'contrast' is not a primitive notion. A 'contrast' relation arises 'when alternate members of some salient set are evoked and, most importantly, when there is felt to be a salient opposition of what is predicated of them.' (Prince (1998)).

Although the number of strong pronominals is small to draw any definitive conclusions with regard to the use of strong pronominals, Table 7 indicates that, at least in Modern Greek, one of the uses of strong pronominals is to signify this type of contrast. Examples (25)-(26) demonstrate the point where in its context, the propositional opposition is between *them* thinking that she was suffering when *she* was actually experiencing pleasure from killing without being caught. A similar contrast is demonstrated in the discourse (10)-(13) given in section 4. [5]

(25) ke agonizondan na me
and were-trying-they subjun-prt me
parigorisun.
console-they

'and they were trying to console me.(SMOOTH-SHIFT)'

(26) Omos ego iha epitelus vri ton eafto
however I had finally found the self
mu...
my...

'However, I had found myself... (CONTINUE)'

Turning to the remaining categories of Table 7, in the case of relative clauses, the use of a strong pronominal is obligatory and dictated by the grammar of the language. The emphatic instance is also controlled by the grammar. In this case, the strong form appears after the phrase 'ute ke' (not even) which is necessarily followed by either a full NP or a strong pronominal.

We can now turn to Di Eugenio (1998)'s motivation for treating tensed adjunct clauses as independent update units. We repeat the relevant example for convenience:

(27) Prima che i pigroni-i siano seduti a tavola a far colazione,
'Before the lazy ones-i sit down to have breakfast,'

(28) lei-j e via col suo-j calessino alle altre cascine della tenuta.
'she-j has left with her-j buggy for the other farmhouses on the property.'

The example above is simply another instance of using a strong pronominal to contrast the salient entity with some other entity in the discourse. It is plausible that in Italian, like in MG, the strong pronominal is not used to signify a Rough-Shift but to contrast 'she' the salient entity in (28) with the 'lazy ones', in (27). That the salient entity in the previous discourse is the 'vicina' is also verified by the immediately preceding discourse, shown in (29)-(31). [6]

(29) NULL-j e' una donna non solo graziosa ma anche energica e dotata di spirito pratico;
'and not only is she-j pretty but also energetic and endowed with a pragmatic spirit;'

(30) NULL-j e la combinazione di tutto cio' e', a dir poco, efficace.
'and the combination of all these qualities is effective, to say the least.'

(31) NULL-j si alza all'alba per sovrintendere a che si dia da mangiare alle bestie, si faccia il burro, si mandi via il latte che deve essere venduto; una quantita' di cose fatte mentre il piu' della gente se la dorme della grossa,
'she gets up at dawn to supervise that the cows are fed, that the butter is made, that the milk to be sold is sent away; a lot of things done while most people sleep soundly '

[5] Dimitriadis (1996) argues that strong pronominals in MG , categorically, indicate that the antecedent is NOT the Cp of the previous utterance. For reasons of space we cannot go into the details of his analysis here but we will point out that there is ample evidence that strong pronominals do, in fact, pick the Cp of the previous utterance as their antecedent precisely in the cases identified here. The following example, taken from a Greek newspaper on line (*Eleftherotipia 10/3/2000*) clearly shows that Dimitriadis's claim overlooks the contrastive function of strong pronominals.

(1) To idio kani ke i N.D-i.
the same does and the N.D.

'N.D.-i (our note: Greek opposition political party) do the same.

(2) Null-i gnorizi alla den null-i lei.
null knows but not say.

'They-i (literally, she-i) know but they-i don't say.'

(3) Aoristos null-i iposhete oti **afti-i** tha diahiristi
Vaguely null promises that she will manage
kalitera tin meta ONE epohi me to epihirima
better the after ONE era with the argument
oti null-i ine to kat' exohin evropaiko
that null is the pre dominately European
komma.
party.

'They-i vaguely promise that THEY (our note: contrasting governing party) will manage the after ONE (European Currency Unification) era with the argument that they are the predominantly European political party'

[6] Many thanks to Barbara Di Eugenio (personal communication) for providing me with the extra data in (29)-(31)

6 The significance of the update unit in NLP applications

In this section we will briefly discuss some more evidence coming from an educational application, namely automated writing evaluation, and offer some pointers of the usefulness of the salience update unit to coreference based text summarization.

In Miltsakaki and Kukich (2000a) we show that a metric of text incoherence based on the proportion of Rough-Shifts to the total amount of realized transitions improves the performance of *e-rater* (Burstein et al. (1998b), Burstein et al. (1998a)), an automated essay scoring system developed at ETS. The results were statistically significant and the proposed algorithm was performed on a total of 100 student essays . For a small amount of essays we constructed a questionnaire for writing experts and asked them to give us their evaluation of those essays, isolating essay coherence from the other factors that affect the final score of an essay. Their evaluation was consistent with our coherence score when transitions were identified and marked according to the definition of the updated unit proposed here. While there were cases where the number of Rough-Shifts did not increase due to splitting main clauses from tensed subordinate clauses (because the same Cb carried over to the subordinate clause) in cases where it mattered (when different entities were evoked in subordinate clauses) treating subordinate clauses as independent units produced coherence scores that conflicted with the evaluation of the experts. Due to the small number of essays evaluated by human experts specifically for coherence, we do not have quantified results to report. However, the compatibility of the Rough-Shift metric with the corresponding human evaluation of incoherence remains impressive despite the small sample.

In text summarization, various researchers utilize coreference chains for automated summarization. Azzam et al. (1999)'s approach is based on the intuition that texts are about some central entity or entities which can be viewed as the topic of the discourse. They first build coreference chains. Then, sentence selection (for inclusion in the summary) is done on the basis of criteria relating to the length, spread and start of chain. Baldwin and Morton (1998) develop a system for constructing summaries of documents containing information relevant to a query. In their system, coreference relations, including pronominal resolution, are identified between the query and the document under consideration. Sentence selection is decided upon a scoring system which assigns scores based on various relations of the coreference chains and various elements of the query and the document in question. Morton (1999) utilizes coreference chains to retrieve answers to user-made queries. The task for his system is to identify and retrieve the text that contains the answer to the query. Coreference chains are first constructed in single documents so that the discourse context where these entities appear is not overlooked. To improve the relevance of the summary and also to control its length all of the above systems implement additional mechanisms: a)focus chains, (b)exclusion of various structures such as prepositional phrases, relative clauses etc, and c)models for determining salience, respectively). We suggest that such additional mechanisms may be dispensed with if coreference chains are built only for the most salient entities. Furthermore, sentence selection can be restrained by the status of the clause where topics appear, thus giving preference to main clauses.

7 Conclusion

This paper presented empirical evidence for the location of topics in text processing. We discussed our motivation for keeping information structure distinct from salience structure and hypothesized that surface clause order phenomena can be accounted for by a better understanding of information packaging strategies. Viewing topic structure as a distinct phenomenon helps us draw a clearer picture and achieve a better understanding of how topics and topic shifts are identified in text processing. For an automated system of topic tracking it is crucial that we know where to look for potential topics. Using key concepts from the Centering Model of local discourse coherence and empirical findings from Modern Greek and Japanese we defined the relevant update unit as the traditional sentence, elaborating on issues concerning the status of subordinate clauses. This larger unit, in combination with the Centering notions of Cb and Cp better reflects the way we perceive topic transitions in natural discourse and, at the same time, it simplifies considerably the job of automated topic tracking systems. A challenging research direction in modelling topic structure would involve a better understanding of a)the salience status of events and possibly other non entity-like topics and b)how topic structure in general interacts with other aspects of text processing such as information packaging, rhetorical relations and intentional structure. We leave this for future work.

References

S. Azzam, K. Humphreys, and R. Gaizauskas. 1999. Using coreference chains for text summarization. In *Proceedings of the Workshop 'Coreference and Its Application'*.

B. Baldwin and T. S. Morton. 1998. Dynamic coreferenced-based summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP ;98)*.

S. Brennan, M. Walker-Friedman, and C. Pollard. 1987. A Centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162. Stanford, Calif.

J. Burstein, K. Kukich, S. Wolff, and M. Chodorow. 1998a. Enriched automated essay scoring using discourse marking. In *Discourse Relations and Discourse Markers Workshop, Annual Meeting of the Association for Computational Linguistics, Montreal, Canada*, August.

J. Burstein, K. Kukich, S. Wolff, M. Chodorow, L. Braden-Harder, M.D. Harris, and C. Lu. 1998b. Automated essay scoring using a hybrid feature identification technique. In *Annual Meeting of the Association for Computational Linguistics, Montreal, Canada*, August.

B. Di Eugenio. 1990. Centering theory and the italian pronominal system. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING 90)*, pages 270–275. Helsinki.

B. Di Eugenio. 1998. Centering in Italian. In *Centering Theory in Discourse*, pages 115–137. Clarendon Press, Oxford.

A. Dimitriadis. 1996. When pro-drop languages don't: Overt pronominal subjects and pragmatic inference. In *Proceedings of CLS 32*.

B. Grosz and C. Sidner. 1986. Attentions, intentions and the structure of discourse. *Computational Linguistics*, 12:175–204.

B. Grosz. 1977. The representation and use of focus in language underastanding. Technical Report No. 151, Menlo Park, Calif., SRI International.

L. Horn. 1986. Presupposition, theme and variations. In *Chicago Linguistics Society*, volume 22, pages 168–192.

S. Hudson-D'Zmura. 1988. *The Structure of Discourse and Anaphor Resolution: The Discourse Center and the Roles of Nouns and Pronouns.* Ph.D. thesis, University of Rochester.

A. Joshi and S. Kuhn. 1979. Centered logic: The role of entity centered sentence representation in natural language inferencing. In *6th International Joint Conference on Artificial Intelligence*, pages 435–439.

A. Joshi and S. Weinstein. 1981. Control of inference: Role of some aspects of discourse structure: Centering. In *7th International Joint Conference on Artificial Intelligence*, pages 385–387.

M. Kameyama. 1985. *Zero Anaphora: The Case of Japanese.* Ph.D. thesis, Stanford University.

M. Kameyama. 1993. Intrasentential Centering. In *Proceedings of the Workshop on Centering.* University of Pennsylvania.

M. Kameyama. 1998. Intrasentential Centering: A case study. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, pages 89–112. Clarendon Press: Oxford.

E. Miltsakaki and K. Kukich. 2000a. Automated evaluation of coherence in student essays. In *Proceedings of the Workshop on Language Rescources and Tools in Educational Applications, LREC 2000.*

E. Miltsakaki and K. Kukich. 2000b. The role of centering theory's rough shift in the teaching and evaluation of writing skills. In *Proceedings of ACL 2000, Hong-Kong (to appear).*

T. S. Morton. 1999. Using coreference for question ansering. In *Proceedings of the Workshop 'Coreference and Its Application'.*

E. Prince. 1981. Topicalization, focus-movement, and Yiddish-movement: A pragmatic differentiation. In D. Alford et al, editor, *Proceedings of the Seventh Annual Meeting of the Berkeley Linguistics Society*, pages 249–264.

E. Prince. 1998. On the limits of syntax, with reference to left-dislocation and topicalization. In P. Culicover and L. McNally, editors, *The Limits of Syntax*, volume 29 of *Syntax and Semantics.* NY: Academic Press.

O. Rambow. 1993. Pragmatic aspects of scrambling and topicalization in German. In *Workshop on Centering Theory in Naturally Occuring Discourse.* Institute of Research in Cognitive Science, University of Pennslylvania.

T. Reinhart. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica*, 27:53–94.

C. Sidner. 1979. Toward a computational theory of definite anaphora comprehension in English. Technical Report No. AI-TR-537, Cambridge, Mass. MIT Press.

L. Suri, K. McCoy, and J. DeCristofaro. 1999. A methodology for extending focusing frameworks. *Computational Linguistics*, 25(2):173–194.

U. Turan. 1995. *Null vs. Overt Subjects in Turkish Discourse: A Centering Analysis.* Ph.D. thesis, University of Pennsylvania.

U. M. Turan. 1998. Ranking forward-looking centers in Turkish: Universal and language specific properties. In A. Joshi M. Walker and

E. Prince, editors, *Centering Theory in Discourse*, pages 139–160. Clarendon Press, Oxford.

M. Walker and E. Prince. 1995. A bilateral approach to givenness: A hearer-status algorithm and a Centering algorithm. In T. Fretheim and J. Gundel, editors, *Reference and Referent Accessibility*. Amsterdam: John Benjamins.

B. Webber and A. Joshi. 1998. Anchoring a lexicalized tree adjoining grammar for discourse. In *ACL/COLING Workshop on Discourse Relations and Discourse Markers*. Montreal, Canada.

B. Webber, A. Knott, M. Stone, and A. Joshi. 1999a. Discourse relations: A structural and presuppositional account using lexicalized tag. In *1999 Meeting of the Association for Computational Linguistics*. College Park MD.

B. Webber, A. Knott, M. Stone, and A. Joshi. 1999b. What are little texts made of? A structural and presuppositional account using lexicalized tag. In *International Workshop on Levels of Representation in Discourse (LORID '99)*.

View publication stats