Centering in Greek

Eleni Miltsakaki
University of Pennsylvania

Abstract

This paper presents a corpus-based analysis on the discourse functions of weak and strong forms of referring in Greek. We focus on null subjects as well as overt weak and strong pronominal forms. The distribution of the pronominal paradigms in a Greek corpus reveals multiple discourse functions. Specifically, null pronouns signify continuation on the same topic or return to the main/earlier topic after an interruption or other embedded structure has occurred. Strong pronominals prompt reference to a non-salient entity and alert to an upcoming switch, sometimes abrupt, to a new topic. A second function of strong pronouns is to signify a contrastive relationship of an entity to some other salient entity or set of entities, previous evoked in the discourse. We analyze the data and model the discourse constraints on their distribution with respect to the Centering Model of attention in discourse (Grosz, Joshi, & Weinstein, 1995).

1 Introduction

In this paper an empirical quantitative study is presented on the distribution and discourse functions of nominal and pronominal forms in Modern Greek. The study contributes to the investigation of the complexities involved in form-function mappings and shows that, in fact, form-functions mappings are not unitary. Instead, linguistic forms can and are used to serve multiple functions in the organization of the discourse. We demonstrate how these functions are partially modelled in the Centering framework while others need more elaborate semantic representations. The current work has interesting implications for computational approaches to discourse, which space restrictions will not allow us to discuss. We only note that understanding the complexities of one-to-many mappings of linguistic form to discourse function is crucial for the successful design and development of discourse models in natural language processing and leave the question open to future research. The paper is organized as follows.

In Section 2, we give a brief overview of Centering Theory, a model of attentional state in discourse which relates discourse coherence with choice of referring expression. Section 3 offers a brief description of the pronominal paradigm in Greek. Starting with the assumption that certain entities in discourse are more salient than others, we investigate the factors which determine the relative salience of discourse entities in Greek in Section 4. We argue that the strongest salience factor in Greek is not surface word order, as would be expected, but grammatical role, in particular subjecthood. We support this claim using Rambow's (1983) diagnostic on salience and confirm it with corpus-based results. Next, we discuss the notion of Center update unit (Centering's *utterance*) and show that contra Kameyama's (1998) initial hypothesis, tensed adjunct clauses in Greek do not form independent update units. In other words, new entities introduced in subordinate structures do not override the salience status of the entities evokes in main clauses. We support this claim with empirical evidence.

Having established the Centering update unit and the salience ranking of discourse entities in Greek, we proceed to the corpus analysis of the nominal and pronominal forms in Section 5. It is shown that null subjects and weak pronominals are used to signal topic continuation and return to a superordinate unit after an embedded unit is closed off (e.g. parenthetical text and switches from narrative to direct speech). Strong pronominals are used to refer to low salience entities, warn for an abrupt topic shift, and signal a partially ordered set relationship (e.g. contrast) of the current entity to a previously evoked salient set of entities. The analysis of the data is formally modelled with regard to the Centering Model of attention in discourse.

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

Table 1: Table of Centering Transitions

2 Overview of Centering

Centering was developed as a model of the center of attention between speakers in natural language discourse. The model aimed at modeling the interaction between 'attentional state', inferential complexity and the form of referring expression. The formulation of Centering Theory resulted from the synthesis of two main lines of work. Originally, Joshi, Kuhn, and Weinstein (Joshi & Kuhn, 1979; Joshi & Weinstein, 1981) proposed Centering as 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. Grosz and Sidner (Sidner, 1979; Grosz, 1977; Grosz & Sidner, 1986) recognized what they called 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 of attention. A synthesis of these two approaches yielded the Centering model which was designed to account for those aspects of processing that are responsible for the difference in the perceived coherence of discourses as those demonstrated in (1) and (2) below (examples from (Grosz et al., 1995)).

(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. John 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). This difference may be seen to arise from the different degrees of continuity in what the discourse is about. Discourse (1) centers a single individual, *John*, whereas discourse (2) seems to center in and out different entities, *John, store, John, store*. Centering is designed to capture these fluctuations in continuity.

Also, contra earlier assumptions based on purely semantic or inferential theories of discourse understanding, (Hobbs, 1985). Centering also predicts that discourses (3) and (4) below, differ in coherence despite the fact that there is no semantic ambiguity at the time the discourses are fully processed and therefore the referents of the pronouns should be equally easy to retrieve.

(3) a. Jeff helped Dick wash the car.
b. He washed the windows and Dick waxed the car.
c. He soaped a pane

(4) a. Jeff helped Dick wash the car
b. He washed the windows and Dick waxed the car.
c. He buffed the hood

The pronominal subject in 4c can only be interpreted as *Dick* because the semantics of *buffing* is associated with the *waxing* event. Still, by using a pronoun in (4c), the speaker is only confusing the reader because up to utterance (4a) *Jeff* has been the center of attention and therefore the most likely referent of the pronoun in (4c). It is only when the hearer gets to the word *buff* that s/he realizes that the referent must be *Dick*.

In what follows we present the basic concepts and data structures of the model to demonstrate how Centering evaluates discourse coherence and its interaction with choice of referring expression.

2.1 The Centering Model

The Centering view of discourse is very simple. Discourse consists of a sequence of textual segments and each segment consists of a sequence of utterances. 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 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 the previous utterance, $Cf(U_{i-1})$, that is *realized* in the current utterance, U_i, is its designated 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, 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. It is precisely this distinction between looking back in the discourse with the Cb and projecting preferences for interpretations in the subsequent discourse with the Cp that provides the key element in computing local coherence in discourse within the Centering framework.

Centering rules and transitions. Since Centering is designed to model attentional state, it followis that it also defines changes or shifts in attention. Four transitions from one attentional state to another are defined which also reflect four degrees of coherence: Continue, Retain, Smooth-Shift, and Rough-Shift. The rules for computing the transitions between two adjacent utterances are shown in Table 1. They correspond to the four combinations of two variables: whether the 'topic' of the current utterance, i.e., Cb(Ui), is the same as the 'topic' of the previous utterance, i.e., Cb(Ui-1), and whether the 'topic' of the current utterance, Cb(Ui), is realized in a position saved for salient entities, Cp(Ui), the highest ranked entity in the Cf set. In English, for example, that position has been argued to be the subject position. Finally, Centering transitions are ordered according to degree of coherence as defined in the Transition Ordering rule, shown below.

Transition Ordering Rule:
Continue is preferred to Retain, which is preferred to Smooth-shift, which is preferred to Rough-shift.

Centering, also, defines a rule, known as the **Pronoun Rule**,which constrains the choice of referring expression in certain conditions and at the same time makes a testabe prediction for the theory:

Pronoun Rule:
If some element of the Cf of the previous utterance is realized as a pronoun in the current utterance, then so is the Cb of the current utterance.

The Pronoun Rule captures the intuition that pronominalization is one way to indicate discourse salience and that Backward-looking centers are often deleted or pronominalized. Later studies in pro-drop languages like Japanese (Kameyama, 1985) or Turkish (Turan, 1995) showed that the Pronoun Rule for such languages must be reformulated to accommodate zero pronouns: If some element of the Cf of the previous utterance is realized as a zero pronoun in the current utterance, then so is the Cb of the current utterance.

The Pronoun Rule and the Centering Transitions predict that the interpretations that hearers will prefer when processing discourse are those requiring minimal processing effort. For example, an instance of a Continue transition followed by another Continue transition requires minimal effort for interpretation, as the hearer only needs to keep track of one main entity which is both the Cb and the Cp of the current utterance. Below, we demonstrate how the Centering Rules apply to discourses (3) and (4), repeated here in Table 2.

a. Jeff helped Dick wash the car.	a. Jeff helped Dick wash the car.
Cb=none Cf=Jeff,Dick, car Transition=none	Cb=none Cf=Jeff,Dick, car Transition=none
b. He washed the windows and Dick waxed the car.	b. He washed the windows and Dick waxed the car.
Cb=**Jeff** Cf=Jeff,windows, Dick, car Transition= Continue	Cb=**Jeff** Cf=Jeff,windows, Dick, car Transition= Continue
c. He soaped the pane.	c.' He buffed the hood.
Cb=**Jeff** Cf=Jeff Transition= Continue	Cb=**Dick** Cf=Dick Transition= Smooth-shift

Table 2:

Utterance (b) is a Continue transition because the Cb is the same as in (a) and the Cp in (b) is the same as the Cb in (b), namely *Jeff*. In contrast, (c') is a Smooth-shift transition, because the Cb has changed from (b), but the Cp is the same as the Cb. According to the Centering Model, the discourse with the (c') continuation is less coherent that the one with (c). The Continue transition identified in the (b) utterances is interpreted as an indication by the speaker that s/he intends to Continue talking about the *Jeff*. Instead, the speaker, shifts attention (with a Smooth-shift transition) to

Cf Ranking Rule (1)
SUBJ>IND. OBJ>OBJ>OTHERS
Cf Ranking Rule (2)
(TOPIC)>(EMPATHY)>SUBJ>IND. OBJ>OBJ>OTHERS>(QIS, PRO-ARB)

Dick). This is misleading for the hearer who first interpetes the pronoun *he* in (c') as the Cp of the precious utterance (cf the Pronoun Rule) and then has to revise this interpetation. (Walker, Joshi, & (eds), 1998), (Hudson-D'Zmura & Tanenhaus, 1998) show that this corresponds to both an increase in processing time and an increase in subjectts' judgment that the discourse with the (c') continuation doesn't make sense.

2.2 Setting Centering Parameters

In the previous section we illustrated the basic Centering definitions with constructed examples from English. However, attempting to test the theory on naturally occurring data brings to the surface issues that were left unspecified in Centering. Open issues in Centering are, for example, the rule determining the Cf ranking and the definition of the *utterance*. However, determining the salience factors or what textual chunk constitutes a single update unit in discourse can be and maybe is best seen as Centering parameters that need to be set for each language under investigation. There is no a priori reason to maintain that such specifications are universal and therefore they will be best studied on empirical grounds across languages.

Cf ranking (Kameyama, 1985) and (Brennan, Walker-Friedman, & Pollard, 1987) proposed that the Cf ranking for English is determined by grammatical function as shown in the Cf Ranking Rule (1). Later crosslinguistic studies based on empirical work (Di Eugenio, 1998; Turan, 1995; Kameyama, 1985) yielded the Cf ranking shown in Rule (2). The parentheses indicate optionality depending on the language.

TOPIC and EMPATHY were first introduced for Japanese,(Walker, Iida, & Cote, 232), where it was proposed that entities marked with the TOPIC marker in Japanese rank higher than entities that speaker marks as EMPATHY (LOCUS).[1] (Turan, 1995) suggests that the notion of EMPATHY is also relevant to Western languages in cases of non-agentive psychological verbs such as *interest* and *seem*, perception verbs such as *feel* and *appear*, and in general expressions that refer to a character's point of view such as *The thought crossed her mind*. Also, Turan found that in her Turkish corpus quantified indefinite subjects (QIS) and arbitrary plural pros (PRO-ARB), such as dropped subjects representing non-specific instances of *we* or *you*. We investigate the factors determining discourse salience in Greek in Section 4.1.

Utterance. In early formulations of Centering Theory, the 'utterance' was not defined explicitly. In subsequent work, (Kameyama, 1998) was concerned with intra-sentential centering and the problem of defining the appropriate update units when processing complex sentences. (Kameyama, 1998) formulated a set of hypotheses which can, roughly, be summarized as follows: main clauses, tensed adjunct and tensed conjunct clauses constitute independent processing units. Most of the subsequent work on Centering assume this hypothesis without further investigation. (Notable exceptions are (Miltsakaki, 1999), (Miltsakaki & Kukich, 2000) and (Poesio, Cheng, Henschel, Hitzeman, Kibble, & Stevenson, 2000)).

3 The Pronominal System in Greek

The pronominal system in Greek consists of two pronominal paradigms: strong and weak. Greek also allows null subjects, which we will classify in the weak paradigm. Weak as well as strong pronouns obey certain syntactic constraints which we will now enumerate.

Null pronouns are only allowed in subject position. When direct and indirect objects are realized with weak pronouns, they must cliticize to the left of the verb. The order of the pronominal clitics is also constrained with the indirect object always preceding the direct object. Strong pronominals are obligatory in prepositional phrases and also when heading a relative clause. Both weak and strong pronouns are morphologically marked for case, number and gender. Greek has three genders, masculine, feminine or neuter. Noun phrases with human referents are normally marked as male of female (except for infants and kids). However, other animate and all inanimate objects can be either masculine or feminine or neuter.

[1] In Japanese, the EMPATHY LOCUS marks the entity who the speaker identifies with or takes his pespective in the discourse.

Strong and weak forms are also available in possessive noun phrases (NPs). Weak possessive NPs consist of the head noun followed by a weak form in genitive, shown in example (5). Strong possessive NPs are constructed with an emphatic form preceding the head noun and marked with the same case as the head noun, example (6). In this study, we have classified as strong possessive, and therefore as a strong form, instances where both the possessor and the possessee are nouns as in (7). Finally, another paradigm of a strong forms that we have considered is the anaphoric *o idhios* which is also morphologically marked for gender, number and case. An example of *idhios* is shown in (8).

(5) I mitera mu
the mother my
'My mother.'

(6) I diki mu mitera
the own my mother.
'MY mother.'

(7) I mitera tis Marias.
the mother the-gen Maria-gen
'Maria's mother.'

(8) I idhia ostoso ihe apoliti sinesthisi.
the herself however had absolute awareness.
'(She) herself however was fully aware.'

4 Specifying Centering parameters in Greek

4.1 Salience factors

The salience status of an entity is determined by a number of factors which, as already suggested, may vary cross-linguistically. This is because languages may choose different linguistic strategies and/or encoding to single out entities that are intended to be more salient in discourse. In English, for example, Grosz, Joshi and Weinstein (1983, 1986, 1995) (also, (Kameyama, 1985) and (Brennan et al., 1987)) have proposed that the Cf list is partially determined by the grammatical configuration hierarchy and that subjects rank higher than objects. (Rambow, 1993), on the other hand, claims that in German, the salience of the entities appearing between the finite and non-finite verbs (Mittelfeld) is determined by word order and used a diagnostic to confirm this claim.

One would expect that possibly in free word order languages, in general, the relevant salience of entities would be reflected by choices in word order. Contra expectations, it turns out that, in fact, this hypotheses does not hold. In what follows, we use Rambow's diagnostic to explore the salience factors in Greek, which indicates that word order in Greek does not affect salience. Subjects seem to rank higher than objects independently of their surface position. (Turan, 1995) and (Prasad & Strube, 2000) have reached similar conclusions for Turkish and Hindi respectively (both free word order languages).

In discourses (9) and (10), word order and grammatical function are contrasted. The discourse initial question in (9a) introduces two entities, i.e. *prosfati diefthetisi* and *ikonomiki politiki* in subject and object positions respectively. Both entities are ambiguous morphologically as they are both marked feminine. In (9b), the reply to the question contains a dropped subject and the predicate *aneparkis* is such that can take either *prosfati diefthetisi* or *ikonomiki politiki* as its subject. The test involves a native speaker's interpretation of the dropped subject in the reply under two condition. In the first condition, shown in discourse (9), the subject appears in the preverbal position. In the second condition, shown in discourse (10), the object has been fronted and the subject appears post-verbally. If the interpretation of the dropped subject in the reply changes according to the surface position of the entities, we will take it as an indication that word order affects the salience of entities in Greek. If, on the other hand, the dropped subject in both conditions is interpreted as the subject of the preceding questions, we will take it as an indication that grammatical function determines salience in Greek, with subjects ranking higher than objects. It turns out that the dropped subject is interpreted as the subject of the preceding question in both conditions so we, at least tentatively, conclude that grammatical function determines discourse salience in Greek with subjects ranking higher than objects.

(9) a. I prosfati diefthetisi-i tha veltiosi tin ikonomiki politiki-j?
the recent arrangement will 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.'

(10) a. Tin ikonomiki politiki-j tha ti-j veltiosi i prosfati diefthetisi-i?
the economic policy will CL-it 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.'

Additional evidence for the lower ranking of objects with respect to subjects is provided in example (11) and the continuations in (11a)-(11d). We observe that the object in (11) is unavailable for the interpretation of the dropped subject in (11a). Instead, subject reference to *Yorgo* can be achieved either with a full noun phrase, the proper name in this case, or with a strong pronominal. This behaviour is reminiscent of the Promotion Rule proposed in (Turan, 1998) for Turkish. The Promotion Rule states that in order for the object to be realized as a null subject in subsequent discourse it first needs to be 'promoted' to the subject position as a full NP. Preliminary evidence shows that the same rule holds in Greek except that promotion to subject position can be achieved with any strong form, including strong pronouns.

(11) 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/'

Based on the observations in this section, we formulate the hypothesis that salience in Greek is determined by grammatical function and will assume the following Cf ranking: **SUBJECT>OBJECT>OTHER**. In the course of this study we also found evidence that EMPATHY is also a factor in Greek, in cases of verbs such as *like* where the dative experiencer (marked with genitive case in Greek) is more salient than the grammatical subject. Also, as in Turkish, quantified indefinite expressions and impersonal uses of *we* and *you* are also ranked low independently of their grammatical function.

4.2 Determining the size of the *utterance*

At a preliminary stage of the current study, the working hypothesis with regard to the size of the *utterance*, was Kameyama's set of hypotheses. It soon became obvious, however, that Kameyama's tensed adjunct clause hypothesis did not hold with respect to the Greek data. The tensed adjunct clause hypothesis treats tensed adjunct clauses as independent Centering units. [2]Treating such clauses as independent units in Greek yielded discourses with several counter-intuitive Rough-shift transitions. Rough-shift transitions are the least coherent transitions that are rarely found, especially in written discourses.

By way of demonstration, let us consider (12). If we treat the adjunct clause in (12b) as an independent unit, we end up with three Rough-Shift transitions, contra the perceived continuity of the discourse (at least in our judgment). The picture changes dramatically if we treat (12b) and (12c as one unit. The resulting transitions are Continue-Continue-Retain.

(12) a. 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

[2]Tensed adjunct clauses are dependent clauses such as 'time', 'concession' clauses etc, excluding relative and complement clauses.

b. 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

c. Ego voltariza stin paralia
I was-strolling on-the beach
'I was strolling on the beach.'
Cb=NONE, Cp=I, Tr=Rough-shift

d. Ki i eforia pu esthanomun den ihe to teri tis
and the euphoria that I-was feeling not have the partner its
'And the euphoria that I was feeling was unequalled.'
Cb=I, Cp=EUPHORIA, Tr=Rough-shift

More interestingly, it turns out that treating tensed adjunct clauses as independent processing units is problematic in other languages as well. For example, in the English disocures (13) and (14), allowing the time clause to be an independent update unit yields a highly incoherent discourse (two Rough-shifts). Besides, 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 (14), yields an improved discourse where one Continue transition has now replaced one Rough-shift in (13). Assuming that the two discourses demonstrate a similar degree of continuity regarding their attention structure (they are both *about* 'John'), we would expect the transitions to reflect this similarity when, in fact, they do not. It seems, then, that the introduction of a new discourse entity, 'meeting', in the time-clause does not affect the salience of *John*. Neither does it project a preference for an attention shift, as the Cp normally does when it instantiates an entity different from the current Cb.

(13) a. John had a terrible headache.
Cb=none, Cp=John, Transition=none

b. As soon as the meeting was over,
Cb=none, Cp=meeting, Transition=Rough-shift

c. he rushed to the pharmacy store
Cb=none, Cp=John, Transition=Rough-Shift

(14) a. John had a terrible headache.
Cb=none, Cp=John, Transition=none

b. He rushed to the pharmacy store
Cb=John, Cp=John, Transition=Continue

c. as soon as the meeting was over.
Cb=none, Cp=meeting, Transition=Rough-Shift

Further evidence in support of incorporating tensed adjunct clauses with the main clause comes from Japanese.[3] In Japanese, topics and subjects are lexically marked ('wa' and 'ga' respectively), null subjects are permissible and subordinate clauses must precede the main clause. Consider the Japanese discourse (15). Crucially, the referent of the null subject in the second main clause resolves to the topic marked subject of the first main clause and not to the competing antecedent in the intermediate subordinate clause.

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

b. Jiroo ga rippana osiro o tukutteiru node
Jiroo SUB great castle OBJ is-making because
'Since Jiroo is making a great castle,'

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

We conclude that tensed adjunct clauses do not form independent processing units and will assume that a single *utterance* is defined as the main clause accompanied by its dependent clauses, including tensed adjunc clauses. In each tensed clause in the *utterance*, the entities introduced therein will be ranked according to the Cf ranking we assume for Greek. The set of entities introduced in the main clause rank higher than the set of entities introduced in a dependent clause.

[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.

5 Corpus and findings

The corpus in this study comprises a short story of approximately 6,000 words. The short story is an excerpt of the collection titled 'I won't do this favor for you' (my translation), authored by the modern Greek writer C.A. Chomenides. The text was chosen for its richness of nominal and pronominal expressions as the story involves multiple characters. The text was not pre-segmented. Centering theory as a theory of local discourse coherence is designed to apply within local textual segments. However, we decided to compute Centering transitions across the entire text for two reasons: a)because there is no principled way of segementing the text and b)because we wanted to be able to study the behaviour of nominals and pronominals across segment boudnaries, especially those between switches in the mode of writing from direct to indirect speech and vice versa and across paragraph boundaries. A total of 467 units(utterances) were identified, containing 371 weak forms (null subjects and weak pronominals) and 96 strong forms (full NPs and strong pronominals). In each unit, the elements of the Cf list evoked in that unit were coded according to the following coding schema.

null	null subjects
weak	weak pronouns, weak possessives, quantified indefinite phrases realized as null and quantified indefinite phrases realized with a weak pronoun
full	full noun phrases (including proper names)
strong	strong pronouns, strong possessives, epithets, emphatics (i.e. *idhios*

Table 3: Coding

For every two consecutive processing units, Centering transitions were computed according to the Transition Table shown in Table 1, section 2.1. The results of the distribution of forms with respect to Centering Transitions are shown in Table 4.

	continue	retain	smooth-shift	rough-shift
null	203	22	52	29
weak	44	1	9	10
total	243	23	61	39
full	6	8	3	54
strong	1	2	6	6
total	7	20	9	60

Table 4: Distribution of weak and strong referring expressions

As a first approximation, Table 4 reveals an overall tendency for null subjects and weak forms to be associated with Continue and Smooth-shift transitions rather than Retain and Rough-shift transitions. Continue and Smooth-shift transitions have one important property in common: the Preferred Center, in our data realized by either a null subject or a weak pronominal, is the Backward-looking Center (Cb), i.e. it is the current topic of the discourse. Therefore, reduced forms of referring, such as null subjects or in general weak Cps are used when the intended referent is the most salient entity of the discourse, namely, the topic. On the other hand, full forms are preferred when the topic link between the current and previous discourse unit is broken as the case is in Rough-shift transitions.

On closer inspection of the results in Table 4, however, we observe that the number of instances of weak forms in Rough-shift transitions is unexpectedly high. Rough-shift transitions are unexpected because, being the least coherent transitions, they are identified when all links between the current, previous and subsequent discourse are broken. It is therefore surprising that any weak forms of reference would ever be used in these cases. So, we did a second pass of the data, focusing specifically on those cases. The results of the distribution of weak forms over Rough-shift transitions are shown in Table 4.

Following (Grosz & Sidner, 1986), as 'focus pops' are classified cases where a parenthetic or other embedded description has been completed and the current utterance signals the closing off of the interrruption and the return to the previous, super-ordinate discourse segment. Such descriptions or interruptions halt temporarily the flow of the narrative and are sometimes used to give background information of a new setting when the narrative changes setting

focus pops	11
Mode switches	13
Missing arguments	6
deictic links	2
other	4

Table 5: Classification of rough-shifts

or additional non-contemporaneous information related to a main character. For illustration, an example of a focus-pop is given in (17). The immediately preceding discourse is given in translation in (16). The discourse spanning over (17a)and (17b) temporarily freezes the narrative to provide additional information about the hotel and then (17c) resumes the narrative and temporally returns to the discourse in (16), immediately preceding the interruption.

(16) *I took him to a hotel for lovers in Victoria Square, where I used to go at the time of my relationship with Elias, the only boyfriend I ever had who didn't have a vacation house or at least a car.*

(17) a. Mesa se okto hronia o enikiazomenos peristerionas tu erota ihe ekmondernisti
in to eight years the rentable pigeon-loft of love had been-modernized.
'Within eight year the rentable pigeon-loft of love had been modernized.'

b. Ihane vali tileorasis sta domatia ke sistima exaerismou.
had-they put TVS in-the rooms and system of-air-condition.
'They had installed TVs and air-conditioning.'

c. Akinitopiisa to asanser anamesa ston proto kai ston deftero.
immobilized-I the elevator between to-the first(floor) and to-the second(floor).
'I stopped the elevator between the first and second floor.'

Mode switches in Table 5 are switches from narrative to direct speech and vice versa. Moving to the next category, missing arguments are cases where an argument is only implicitly realized in the discourse. So the source of the Rough-shift transition in such cases was the fact that the entity linking the current and the previous discourse was not overtly realized. While this is not a problem for Centering Theory in the sense that a realized entity need not be overtly realized, it is a thorny issue in computational approaches to discourse representation as such entities are hard to retrieve. Deictic links are cases where the link between two utterances is established by discourse deixis, i.e. the use of a demonstrative pronoun like *afto* to refer to a previous textual segment. Discourse deixis and the formulation of its contribution to discourse coherence as well as its interaction with entity-based coherence accounts is an open research area in anaphor resolution in general and in Centering Theory in particular. Adding to the complexity of the phenomenon, human annotators often disagree on the selection of textual segments which serve as antecedents to discourse deictic anaphors (Eckert & Strube, 1999). Finally, The category *other* includes cases such as description of two person scenes where the dialogue between two characters contained only first and second person reference and other such hard to classify in one category cases.

For the sake of clarity we classified focus-pops and mode-switches as two distinct categories. However, mode switches are a type of focus pops as a textual segment of direct speech embedded in a narrative forms a hierarchically distinct segment in the discourse. They constitute a kind of interruptions to the flow of the narrative. Under this treatment the results in Table 5 (a) provide further evidence for Grosz and Sidner's claim that discourse is hierarchically organized in super- and embedded segments and b) indicate that weak forms can be used to signal the closing off of an embedded segment and return to the topic of the previous discourse before the embedding or interruption started. A corollary of this observation is also that embedded segments are easily identified in discourse processing as hearers seem to keep the preceding discourse and the salient entities therein on hold until the interruption is complete.

Turning to strong pronominals, we see that the number of continue transitions is surprisingly high. In fact, the difference between Continue and rough-shift transitions is insignificant, indicating that the distribution of strong pronominals is not solely controlled by transition type. During a second pass of the data, we focused on the distribution of strong forms, in particular, strong pronominals. Table 6 shows the classification of strong forms in 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

	poset (contrast)	relative	
strong	6	1	
	Extra info	Parathesis	Boundary
Full NPs	3	2	1

Table 6: Strong forms in Continue transitions

are evoked and, most importantly, when there is felt to be a salient opposition of what is predicated of them.' ((Prince, 1998)).

Table 6 then indicates that strong pronominals are used to signify this type of contrast. Example (18) is indicative: given its prior context, the propositional opposition is between *them* thinking that 'she' was suffering when *she* was actually experiencing pleasure from killing without being caught. [4]

(18) a. ke agonizondan na me parigorisun.
and were-trying-they subjun-prt me console-they
'and they were trying to console me.(SMOOTH-SHIFT)'

b. Omos ego iha epitelus vri ton eafto mu...
however I had finally found the self my...
'However, I had found myself... (CONTINUE)'

c. O dikos tis iroikos thanatos den ihe tosi simasia oso i diki mu tapinosi.
the own her heroic death not had that-mush importance as the own my humiliation.
'HER heroic death was not as important (to her) as MY humiliation. (CONTINUE)'

The use of a strong pronoun when heading a relative clause is controlled by the grammar of the language. In this case, no other form is an option so the rough-shift is insignificant and we can easily write a rule to modify the transition. The same is true for the use of strong pronouns when picking an entity from a previously evoked set. The emphatic use in the corpus is also easily identified linguistically as it is preceded by the phrase 'ute ke' (not even) after which only strong forms or full NPs are allowed.

6 Conclusions

In this paper, we presented a corpus-based analysis of the distribution of weak and strong forms of referring in Greek. We concluded that weak and strong forms in Greek, and most likely in other languages, are not unitary phenomena but have multiple functions in discourse. Specifically for Greek, null subjects are used to refer to the most salient entity in discourse, arguably the discourse 'topic', whereas strong pronominals are used to refer to a non-salient entity in discourse. We showed how these functions can be modeled within the framework of Centering's approach to discourse coherence. On the other hand, null subjects are also used to signal the closing off of an embedded structure in discourse and the intention to 'pick-up' the discourse where it was left before the interruption. Also, strong pronominals are used to signal a, for example, 'contrastive' relationship to some other entity be already a member of an evoked set of entities. Identifying the multiplicity of discourse functions that linguistic forms have in a language is a burdensome but crucial step that needs to be taken in order to better understand how to build integrated discourse models that are both linguistically informed and cognitively/computationally possible.

References

Brennan, S., Walker-Friedman, M., & Pollard, C. (1987). A Centering Approach to Pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pp. 155–162. Stanford, Calif.

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

Dimitriadis, A. (1996). When Pro-Drop Languages Don't: Overt Pronominal Subjects and Pragmatic Inference. In *Proceedings of CLS 32*.

[4](Dimitriadis, 1996) argues that strong pronominals in Greek are used to indicate that the antecedent is NOT the Cp of the previous utterance. While we agree that this is indeed one of the functions of strong pronominals as confirmed by our data, too, we would like to point out that this is not the unique function of strong pronominals. There is evidence that strong pronominals do, in fact, pick the Cp of the previous utterance as their antecedent precisely in the cases identified here. For a naturally occurring example of such cases the reader is referred to (Miltsakaki, 1999).

Eckert, M., & Strube, M. (1999). Resolving Discourse Deictic Anaphora in Dialogues. In *Proceeding of EACL'99*.

Grosz, B. (1977). The Representation and Use of Focus in Dialogue Underastanding. Tech. rep. No. 151, Menlo Park, Calif., SRI International.

Grosz, B., Joshi, A., & Weinstein, S. (1995). Centering: A Framework for Modeling Local Coherence in Discourse. *Computational Linguistics*, *21*(2), 203–225.

Grosz, B., & Sidner, C. (1986). Attentions, Intentions and the Structure of Discourse. *Computational Linguistics*, *12*, 175–204.

Hobbs, J. (1985). On the Coherence and Structure of Discourse. Tech. rep. CSLI-85-37, Center for the study of language and information, Stanford University.

Horn, L. (1986). Presupposition, Theme and Variations. In *Chicago Linguistics Society*, Vol. 22, pp. 168–192.

Hudson-D'Zmura, S., & Tanenhaus, M. (1998). Assigning Antecedents to Ambiguous Pronouns: The Role of the Center of Attention as a Default Assignment. In Walker, M., Joshi, A., & Prince, E. (Eds.), *Centering Theory in Discourse*, chap. 11. Oxford University Press.

Joshi, A., & Kuhn, S. (1979). Centered Logic: The Role of Entity Centered Sentence Representation in Natural Language Inferencing. In *6th International Joint Conference on Artificial Intelligence*, pp. 435–439.

Joshi, A., & Weinstein, S. (1981). Control of Inference: Role of Some Aspects of Discourse Structure: Centering. In *7th International Joint Conference on Artificial Intelligence*, pp. 385–387.

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

Kameyama, M. (1998). Intrasentential Centering: A Case Study. In Walker, M., Joshi, A., & Prince, E. (Eds.), *Centering Theory in Discourse*, pp. 89–112. Clarendon Press: Oxford.

Miltsakaki, E. (1999). Locating Topics in Text Processing. In *Proceedings of Computational Linguistics in the Netherlands (CLIN'99)*.

Miltsakaki, E., & Kukich, K. (2000). The Role of Centering Theory's Rough Shift in the Teaching and Evaluation of Writing Skills. In *Proceedings of ACL 2000, Hong-Kong*.

Poesio, M., Cheng, H., Henschel, R., Hitzeman, J., Kibble, R., & Stevenson, R. (2000). Specifying the Parameters of Centering Theory: a Corpus-Based Evaluation using Text from Application-Oriented Domains. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, ACL 2000*.

Prasad, R., & Strube, M. (2000). Pronoun Resolution in Hindi. In *Working Papers in Linguistics*, Vol. 6. University of Pennsylvania.

Prince, E. (1981). Topicalization, Focus-Movement, and Yiddish-Movement: A Pragmatic Differentiation. In et al, D. A. (Ed.), *Proceedings of the Seventh Annual Meeting of the Berkeley Linguistics Society*, pp. 249–264.

Prince, E. (1998). On the Limits of Syntax, with Reference to Left-Dislocation and Topicalization. In Culicover, P., & McNally, L. (Eds.), *The Limits of Syntax*, Vol. 29 of *Syntax and Semantics*. NY: Academic Press.

Rambow, O. (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.

Reinhart, T. (1981). Pragmatics and Linguistics: An Analysis of Sentence Topics. *Philosophica*, *27*, 53–94.

Sidner, C. (1979). Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse. Tech. rep. No. AI-TR-537, Artificial Intelligence Laboratory, Cambridge:Massachusetts Institute of Technology.

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

Turan, U. M. (1998). Ranking Forward-Looking Centers in Turkish: Universal and Language Specific Properties. In Walker, M., Joshi, A., & Prince, E. (Eds.), *Centering Theory in Discourse*, pp. 139–160. Clarendon Press, Oxford.

Walker, M., Iida, M., & Cote, S. (193-232). Japanese Discourse and the Process of Centering. In *Computational Linguistics*, Vol. 20/2.

Walker, M., Joshi, A., & (eds), E. P. (1998). *Centering Theory in Discourse*. Clarendon Press: Oxford.