FACHLITERATUR
EDITION ROMIOSINI
ΕΠΙΣΤΗΜΗ

ICGL12 | 12th INTERNATIONAL CONFERENC ON GREEK LINGUISTICS
16 – 19 SEPTEMBER 2015
FREIE UNIVERSITÄT BERLIN, CEMOG

Proceedings of the ICGL12

vol. 2

The International Conference on Greek Linguistics is a biennial meeting on the study and analysis of Greek (Ancient, Medieval and Modern), placing particular emphasis on the later stages of the language.

PROCEEDINGS OF THE ICGL12
ΠΡΑΚΤΙΚΑ ΤΟΥ ICGL12

Thanasis Georgakopoulos, Theodossia-Soula Pavlidou, Miltos Pechlivanos, Artemis Alexiadou, Jannis Androutsopoulos, Alexis Kalokairinos, Stavros Skopeteas, Katerina Stathi (Eds.)

PROCEEDINGS OF THE 12TH INTERNATIONAL CONFERENCE ON GREEK LINGUISTICS

ΠΡΑΚΤΙΚΑ ΤΟΥ 12ΟΥ ΣΥΝΕΔΡΙΟΥ ΕΛΛΗΝΙΚΗΣ ΓΛΩΣΣΟΛΟΓΙΑΣ

VOL. 2

Vertrieb und Gesamtherstellung: Epubli (www.epubli.de)
Satz und Layout: Rea Papamichail / Center für Digitale Systeme, Freie Universität Berlin
Gesetzt aus Minion Pro
Umschlaggestaltung: Thanasis Georgiou, Yorgos Konstantinou
Umschlagillustration: Yorgos Konstantinou

ISBN 978-3-946142-35-5
Printed in Germany

Online-Bibliothek der Edition Romiosini:
www.edition-romiosini.de

ΠΕΡΙΕΧΟΜΕΝΑ

CORPUS C58 AND THE INTERFACE BETWEEN INTRA- AND INTER-SENTENTIAL LINGUISTIC INFORMATION

Alexandros Tantos, Giorgos Chatziioannidis, Katerina Lykou,
Meropi Papatheohari, Antonia Samara & Kostas Vlachos
Aristotle Univerity of Thessaloniki
alextantos@lit.auth.gr, georgidc@lit.auth.gr, kate_lyk@hotmail.com,
papathem@lit.auth.gr, antosam93@gmail.com, vlackons@gmail.com

Περίληψη

Το παρόν άρθρο έχει δύο κύριους στόχους: να παρουσιαστούν α) τα κύρια χαρακτηριστικά του Corpus 58 (C58), του πρώτου επισημειωμένου με κειμενικές σχέσεις σώματος κειμένων για την ελληνική γλώσσα, και τις προκλήσεις που καλούνται να αντιμετωπίσουν οι συντελεστές ενός αντίστοιχου σώματος κειμένων, και β) τα πρώτα ποσοτικοποιημένα ερευνητικά αποτελέσματα που υποδεικνύουν τη διεπίδραση ενδοπροτασικών και εξωπροτασικών γραμματικών παραγόντων (βλ. Asher & Lascarides (1996), Asher & Lascarides (2003), Danlos (2001), Tantos (2008)).1Πιο συγκεκριμένα, εξετάζοντας το C58, προκύπτει ότι το είδος της κειμενικής σχέσης ανάμεσα σε δυο εκφωνήματα παρουσιάζει σχέση εξάρτησης με τον θεματικό ρόλο των ορισμάτων των ρημάτων, τη ρηματική όψη και το είδος του υποκειμένου, κενού ή λεξικά εκπεφρασμένου.

Keywords: Corpus 58, computational linguistics, discourse relations, verb valency, aspect, thematic roles, subject form, data analysis

1 Corpus 58 is named after the number of annotated texts in its first version.

1. Introduction

Over the last two decades there has been a strong tendency in Computational Linguistics [CL] and Natural Language Processing [NLP] to use linguistically annotated corpora, in order to build classification systems that mine texts in a smarter manner, (cf. Pustejovsky and Stubbs (2012), Marcu and Echihabi (2002)). However, the current state of the art in corpus annotation and exploitation stays at the simplest level of linguistic description and uses most of the times lexically or phrasally annotated textual data. Ignoring the still unresolved -to some extent- problem of defining grammatical categories in linguistic theorizing, the vast majority of existing linguistically annotated resources stay at the morphological or syntactic annotation level and rarely semantically annotated resources are created and exploited at all. Therefore, nowadays efforts for integrating more linguistic information into NLP systems ought to focus on two dimensions:

1. the semantics of utterances,
2. the annotation of more linguistic description levels, such as the discourse one.

Discourse annotation is an unresolved and challenging annotation task for a system to undertake. The present paper aims to point out the need for considering the formal semantic and pragmatic machinery developed within the theoretical linguistic tradition in order to deal with challenging issues related to linguistic annotation at the textual level that could be beneficiary for NLP systems, such as question-answering or summarization systems. Throughout this paper we will present some aspects of the first manually and discourse-annotated corpus for Modern Greek, the C58 corpus that brings up the complexity of discourse annotation and indicates the main corpus design practices and choices made for the compilation of C58. Its second part presents the first data-driven research results that unveil the close and intricate relationship between inter- and intra-sentential levels of linguistic analysis.

Next section describes briefly Segmented Discourse Representation Theory [SDRT], (Asher (1993), Asher and Lascarides (2003)) a formal discourse semantic theory that underlies C58's annotation guidelines. SDRT provides a rich toolset for describing utterance interdependencies based on formal criteria for defining discourse relations and for using graph-based representations. Section 3 describes the main features of C58, while sections 4 and 5 present the annotation principles of C58 and one of the most important challenges any discourse relations' corpus needs to face, namely discourse segmentation. Sections 6 and 7 present and discuss the first results that indicate

a clear connection between the inter- and intra-sentential grammatical factors drawn upon C58's annotations.

2. SDRT and Discourse Relations

SDRT is a formal discourse theory (Asher (1993), Asher and Lascarides (2003)) that lies within the semantics and pragmatics and provides us with the formal tools to analyze discourse structure. Although formal linguistic theories, such as SDRT, have been superseded during the last decade and the current trend in CL is to use ML or other stochastically-based methods for extracting knowledge at the textual level, it seems that a few theoretical axioms are sufficient to interpret the discourse structure appropriately in cases where these methods are incapable of predicting (cf. Asher (1993)).

Discourse relations denote the semantic and pragmatic connection between two discourse units [DUs] or utterances in tradiotional terms. Theories that study discourse structure aim to explain the way utterances or DUs combine with each other and create a coherent text (e.g. Polanyi (1988), Asher (1993), Marcu (2000)). Therefore, *coherence* and *cohesion* are partly formalized notions between the existing formal discourse semantic theories and they start to become tangible ideas that can be studied within discourse-annotated corpora.

Discourse relations are defined based on definite semantic effects that they have on textual semantic interpretation and are classified as either *coordinating* or *subordinating* relations. Briefly, the distinction between the two types of discourse relations has an intuitive motivation: some parts of a text play a subordinate role relative to other parts, and, thus, new information is added at the same 'level of detail' with the previous utterances or not.

Moreover, the distinction between *subordinating* and *coordinating* relations dictates discourse anaphoric accessibility and attachment availability. SDRT's representation of discourse is graph-based and illustrates the hierarchical information flow of discourse. Such a representation clearly demonstrates the basic distinction between discourse coordination and subordination.

(1) π1 John had a great evening last night.
π2 He had a great meal.
π3 He ate salmon.

π4 He devoured lots of cheese.
π5 He won a dancing competition.
π6 ??It was a beautiful pink.

In the classic example of Asher and Lascarides (2003) in (1), π1 to π5 are the existing DUs integrated in the discourse before π6, the current DU, enters it and gets discourse-related with at least one of the previous five DUs. The manner and the number of DUs with which we interpret π6 to be related dictate the strength of coherence. The graph-based semantic representation of figure 1 represents subordinating relations through vertical lines and coordinating ones through horizontal lines.[2]

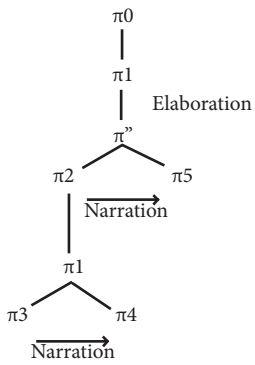

Figure 1 | The graph-based SDRT's representation of (1).

3. C58: main characteristics

C58 is the first corpus annotated with discourse relations for Greek, based on SDRT's theoretical framework for the implementation of its annotation scheme.[3] C58 consists of 58 journalistic texts (more than 1000 annotated DUs), sampled from the *Corpus of*

2 For more details about how figure 1 is constructed, cf. Asher and Lascarides (2003).
3 C58 lies in http://brat.lit.auth.gr/brat/#/, where one can visit all annotated texts and their graph-based annotations. More information and news about the corpus can be found on a separate site, http://projectgdrc.wix.com/c58project. Until now, brat's exported annotation files have been translated into XML files, based on a customized inline XML tagging scheme that bypasses the graph-based reality of our intended representations.

Modern Greek Texts, a reference corpus of Greek compiled by the *Center of Greek Language*.[4] This corpus is a text collection sampled from two popular Greek newspapers, *Μακεδονία* and *Τα Νέα*.[5] It includes approximately 7.000 and 4.500 articles from each newspaper accordingly. We decided to collect sample texts proportionally from both of them in order to neutralize any possible dialectical factors. The initial corpus of the *Center of Greek Language* is divided into 56 text genres, essentially covering the whole extent of the journalistic genre type, ranging from *book review* to *economic news*. We narrowed down the scope of C58 texts to 27 genres, based on the mean size of their texts, in an attempt to meet the balance and representativeness criteria in the sample.

4. The Annotation Cycle

C58 has been annotated a) for discourse relations in the inter-sentential level and b) for thematic roles of the verbal arguments, grammatical aspect (perfective vs. imperfective) and the subject form (null vs. overt) in the intra-sentential level. Annotation of thematic roles has been based partly on Verbnet's thematic roles' inventory (Schuler 2006).[6]

The annotation cycle of C58 consisted of two phases. In the first phase we defined the annotation scheme that includes a) the segmentation of texts into discourse units [DU] used as first order annotation terms for C58, b) the discourse relations' and c) the thematic roles' inventories. The diversity of discourse relations and the difficulty of defining the appropriate set of thematic roles (Levin 2005), expected in any such attempt, led to a series of improvement versions of the annotation scheme until the final set of segmentation criteria and annotation categories were decided. The second phase targeted at maximizing inter-annotator agreement and improving annotation quality. Annotation has been done on brat, an online rapid annotation tool that enables linking of annotations and eases the annotation process, since it resembles many parts of the theoretical representations that lie behind the annotation of C58.

4 The corpus is available online at http://www.greeklanguage.gr.

5 *Μακεδονία* is published in Thessaloniki, northern Greece, while *Τα Νέα* in Athens, southern Greece.

6 Verbnet's thematic roles are a mixture of the classic thematic roles' hierarchies, as defined in the syntax-semantics interface and the Framenet's set of world-knowledge inspired roles. C58 follows a safer annotation strategy for defining and annotating thematic roles within the sentence: whenever verbal semantics clearly and unambiguously assigns thematic roles to verbal arguments, we prefer linguistically motivated thematic roles, adopted in classical by now linguistic literature on thematic role hierarchies (e.g. Van Valin (2005)); For the rest, ambiguous cases C58 resorts to Verbnet's thematic roles.

5. Challenges

Indicatively, some challenging issues related to the annotation of C58 include thematic roles' assignment, impersonal expressions and gerunds. However, the most important –and partly unresolved at the present time- issue that needs to be addressed is the discourse segmentation task. In the next section, we will focus on that issue and will leave out for a discussion in the future the previously mentioned challenges.

5.1 Discourse Segmentation

The annotation of discourse relations is the most advanced type of annotation, since the difficulties start at the segmentation level. There are a number of definitions and approaches as to what constitutes a discourse unit based on a number of various criteria, both syntactic and semantic (cf. Marcu (1998), Marcu et al. (1999), Marcu (2000), Polanyi et al. (2004), Asher (1993)). Moreover, the existing sentence segmentation algorithms, both supervised and unsupervised, such as Kiss and Strunk (2006), have been trained and reached high levels of accuracy measures, with a reasonably good balance between precision and recall. However, discourse units' definition should not be based on sentence boundaries, since they are not identified conceptually with sentences or clauses. Our definition of DUs is the following:

> • *DUs are semantically meaningful units for discourse inference and interpretation in the sense that their presence serves the discourse relatedness and promotes coherence.*

Based on this definition, the borders of DUs are not easily formally defined and their definition resembles that of utterances elsewhere in the semantics' literature. Initially, it seems that it is an almost impossible task to pin down formal criteria for annotating DUs. However, a number of linguistically motivated observations based on C58's texts suggest that a number of clues can in fact lead us to successfully differentiate sentences or clauses from utterances in the relevant way, namely in whether they affect or not discourse coherence.

Two important examples of linguistically inspired segmentation criteria are complex NPs and relative clauses. The usual tacit one-to-one mapping from verb-denoting expressions to clauses or sentence tokens does not hold between verb-denoting expres-

sions and DUs. Complement clauses of complex NPs, such as *the fact that* in (2) (from C58), are not annotated as separate DUs, since the verb that belongs to the complement clause does not affect discourse structure. Its function is of secondary importance for discourse semantic purposes and, therefore, although it retains its syntactic autonomy, it does not obtain a DU status.

(2) *Το γεγονός ότι* η απόκτηση επιθετικού κρίνεται απαραίτητη μετά την παραχώρηση του Καπετάνου και η έλλειψη αξιολόγης εναλλακτικής λύσης, αυξάνει τις πιθανότητες απόκτησης του Σαπουί.
'*The fact that* the acquisition of a forward player is deemed necessary and the lack of a reliable alternative solution increases the chances of acquiring Sapui.'

A second case where grammar is helpful in determining DUs is the case of restrictive relative clauses, namely clauses necessary for sentence interpretation, as in (3). The restrictive clause *that he wants to exploit* places the main clause in the right context, since without it the pronoun *this* cannot easily be resolved within its context. However critical and indispensable for the sentential meaning they are, restrictive clauses are not DUs. In other words, restrictive clauses function as intermediate means for semantically and pragmatically relating the main clause to other utterances rather than being autonomous and able to relate to other utterances separately on their own.

(3) Αυτό είναι το στοιχείο που θέλει να εκμεταλλευτούν.
'This is the evidence that he wants them to exploit.'

On the other hand, nonrestrictive relative clauses are not necessary for the sentence to be interpreted and, therefore, their role is additional and not complementary to the main clause.

The appropriateness of the segmentation criteria of C58 is closely related to the degree that they reveal patterns of discourse inference and interpretation and its interface to sentential grammatical factors. Next section is devoted to a number of independence tests that will answer the central question of the paper; namely whether grammatical factors within sentences influence or drive discourse inference and interpretation.

6. Independence tests between discourse relations and intra-sentential grammatical factors

Manual annotation of a corpus allows us to a) quantitatively explore the data, b) trace regularities and derive new evidence in favor or against given hypotheses and c) feed supervised learning algorithms with valuable domain-specific features. Based on C58 we will focus on the exploration of hypotheses regarding the interface between inter- and intra-sentential grammatical factors. More specifically, we will test hypotheses for the discourse structure As a first step toward a data-analytic approach to annotated data, we conducted a series of independence tests to unveil the first interdependencies between our variables of observation, namely discourse relations on the inter-sentential level and thematic roles, grammatical aspect and subject form on the intra-sentential level. In other words, this section attempts to answer the following question: *Is there any interdependence between intra- and inter-sentential linguistic factors?*

C58 includes 1705 annotated discourse relations. Our data is grouped based on verb valency in the two related utterances (Valency_<1,2 or 3>) and their combinations are coded accordingly, as in table 1.

The reason why we chose verb valency as the main criterion for classifying and studying pairs of related utterances is that transitivity is claimed to influence discourse inference and interpretation (Danlos 2001). The differing number of arguments of the main verbs of each utterance may influence the phenomenon under scrutiny, namely the way that two utterances are interpreted. Particularly, verb valency states a) the number of participants in the main eventuality described by each utterance and, therefore, b) information about the thematic roles of the participants and the way they influenced the event which may reveal essential information about the lexical aspect of the verb that in its turn influences discourse structure (Krifka 2008).

Out of all sixteen possible *valency* combinations, we focused on the four groups of data with the highest frequency, more appropriate for conducting independence tests, namely:[7]

a) transitive verbs in both utterances (*Valency_2 – Valency_2*)

7 Corpus size is a critical issue, especially important when we deal with multivariate representations of data with numerous categories for each variable (cf. the data sparcity problem).

Data Combination Types (first_utterance – second_utterance)	**Annotation Frequencies**
Valency_2 – Valency_2	616
Valency_2 – Valency_1	234
Valency_1 – Valency_1	189
Valency_1 – Valency_2	249
Valency_2– Valency_3	44
Valency_1 – Valency_3	23
Valency_3 – Valency_2	49
Valency_3 – Valency_1	11
Valency_2 – Topicalized cluster	34

Table 1 | Grouped annotations and their frequencies based on verb valency in C58.

b) transitive verb in the first and intransitive verb in the second utterance[8] (*Valency_2 – Valency_1*)
c) intransitive verb in the first and intransitive verb in the second utterance (*Valency_1 – Valency_1*)
d) intransitive verb in the first and transitive verb in the second utterance (*Valency_1 – Valency_2*)

Moreover, in order to improve the quality of the conclusions, we grouped discourse relations into three groups according to their frequency of appearance, since many of them appear rarely in C58. Next four subsections include figures of discourse relations' frequencies for each subgroup in the four above-mentioned groups of data along with the chi-square tests of independence for the three groups of the discourse relations. Section 7 sums up these first results.

6.1 Valency_2 – Valency_2

Starting with the most frequent valency combination, namely when both related utterances include transitive verbs, we can see in figure 2 that *Continuation*, a coordinating

8 *First* and *second* utterances are considered in a linear textual fashion, namely the one that precedes textually is the first utterance, etc.

relation, is the most frequent discourse relation between the relevant DUs, while *Commentary* and *Elaboration*, two subordinating relations, follow with considerable difference in frequency. Therefore, no clear pattern related to the hierarchical nature of the discourse relations can be traced for this case. Zeroing in on the inter-sentential grammatical factors and its, four of them, the thematic role, the verbal aspect and the subject form of the second utterance and the subject form of the first utterance seem to interact with the factor of discourse relations. This may be an indication that the information flow of the discourse imposes restrictions or designates much stronger grammatical preferences in the second utterance. Although the chi-square test does not provide us with a detailed view on the specific kinds of relations and values of the above-mentioned inter-sentential variables, it offers a positive answer as to whether discourse structure and grammar interact and functions as a starting point for further research.

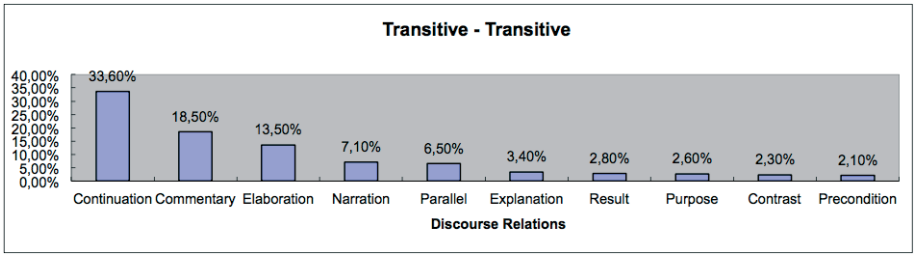

Figure 2 | Discourse relations' frequency for the Valency_2 – Valency_2 combination

Associations or Relations	Chi square statistic	P-value
Discourse_Rel * Them_Role_A1	X2(9) = 16.637	p = 0.055
Discourse_Rel * Them_Role_A2	X2(15) = 22.962	p = 0.085
Discourse_Rel * Them_Role_B1	X2(9) = 16.763	p = 0.053
Discourse_Rel * Them_Role_B2	X2(12) = 33.837	p = 0.001
Discourse_Rel * Asp_A	X2(3) = 3.691	p = 0.297
Discourse_Rel * Asp_B	X2(3) = 17.259	p = 0.001
Discourse_Rel * Subj_Form_A	X2(3) = 11.904	p = 0.008
Discourse_Rel * Subj_Form_B	X2(3) = 26.778	p < 0.001

Table 2 | Chi square tests for the Valency_2 – Valency_2 combination

6.2 Valency_2 – Valency_1

The second valency combination appears to have identical frequency pattern for the discourse relations in figure 3 but, interestingly, dissimilar when it comes to the interaction between specific grammatical factors and discourse relations. Although the statistically significant cases, again, relate discourse relations and grammatical factors in the second utterance, as in the previous case, there seem to be no interaction between the verbal aspect in the second utterance and discourse relations.

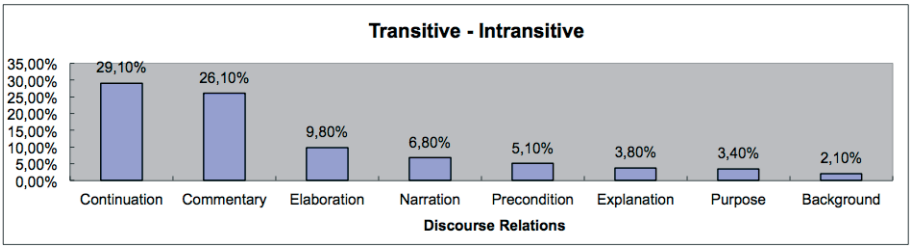

Figure 3 | Discourse relations' frequency for the Valency_2 – Valency_1 combination

Associations or Relations	Chi square statistic	P-value
Discourse_Rel * Them_Role_A1	X2(4) = 1.313	p = 0.859
Discourse_Rel * Them_Role_A2	X2(6) = 4.056	p = 0.669
Discourse_Rel * Them_Role_B1	X2(6) = 14.993	p = 0.02
Discourse_Rel * Asp_A	X2(2) = 0.641	p = 0.726
Discourse_Rel * Asp_B	X2(2) = 0.6558	p = 0.038
Discourse_Rel * Subj_Form_A	X2(2) = 0.864	p = 0.649
Discourse_Rel * Subj_Form_B	X2(2) = 9.793	p = 0.007

Table 3 | Chi square tests for the Valency_2 – Valency_1 combination

6.3 Valency_1 – Valency_1

The valency combination of intransitive verbs in both utterances shows the same frequency pattern for the types of discourse relations as in the previous two combinations. As far as the interaction between grammatical factors and discourse relations is concerned, the thematic role and the subject form of the second utterance interact

with the type of discourse relation. The tendency observed in the previous cases is confirmed again, namely that the grammatical preferences in the second utterance play a significant role in the discourse relatedness.

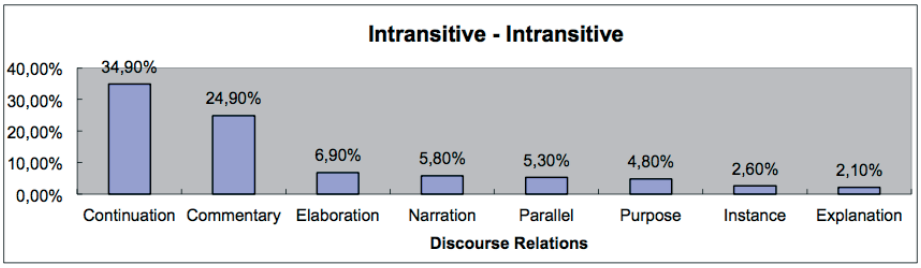

Figure 4 | Discourse relations' frequency for the Valency_1 – Valency_1 combination

Associations or Relations	Chi square statistic	P-value
Discourse_Rel * Them_Role_A1	X2(9) = 13.173	p = 0.155
Discourse_Rel * Them_Role_B1	X2(9) = 5.336	p = 0.804
Discourse_Rel * Asp_A	X2(3) = 6.511	p = 0.089
Discourse_Rel * Asp_B	X2(3) = 7.225	p = 0.065
Discourse_Rel * Subj_Form_A	X2(3) = 3.218	p = 0.359
Discourse_Rel * Subj_Form_B	X2(3) = 3.454	p = 0.327

Table 4 | Chi square tests for the Valency_1 – Valency_1 combination

6.4 Valency_1 – Valency_2

Although the second and third place in the frequency ranking have changed in the last of the four valency combinations, there is no essential change in the hierarchical pattern of the discourse relations, since both *Elaboration* and *Commentary* are subordinating relations. Both grammatical factors of the second utterance interact with the type of discourse relation again and that appears as a persistent pattern in all valency combinations.

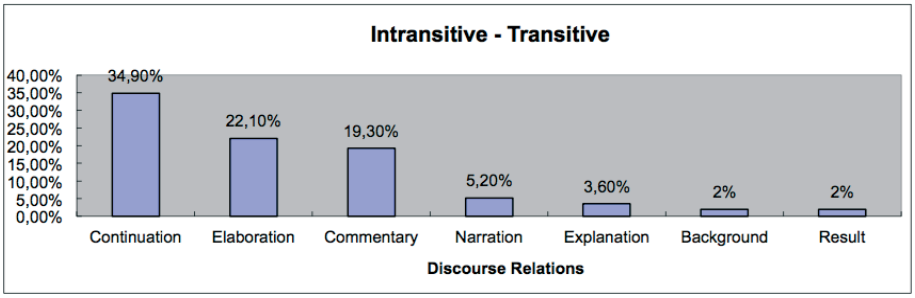

Figure 5 | Discourse relations' frequency for the Valency_1 – Valency_2 combination

Associations or Relations	Chi square statistic	P-value
Discourse_Rel * Them_Role_A1	X2(9) = 20.217	p = 0.017
Discourse_Rel * Them_Role_B1	X2(6) = 20.254	p = 0.002
Discourse_Rel * Them_Role_B2	X2(12) = 27.985	p = 0.006
Discourse_Rel * Asp_A	X2(3) = 4.876	p = 0.181
Discourse_Rel * Asp_B	X2(3) = 14.935	p = 0.002
Discourse_Rel * Subj_Form_A	X2(3) = 1.850	p = 0.604
Discourse_Rel * Subj_Form_B	X2(3) = 13.257	p = 0.004

Table 5 | Chi square tests for the Valency_1 – Valency_2 combination

7. Discussion

In the previous section, the figures and tables 2-5 show that in both groups, the highest appearance frequencies of discourse relations are consistently for *Continuation*, *Elaboration* and *Commentary*. The first and important conclusion drawn from the study is that our data confirm the pre-theoretic intuition that the journalistic genre describes discourse in a specific way, whereby the subordinating nature of *Elaboration* and *Commentary,* that provides more details about events, succeeds the coordinating nature of *Continuation*, of story telling. Additionally, it is worth-noting that the consistent ranking of frequencies shows independence of discourse relations and the transitivity pattern of the utterances' verbs. A compelling project is to expand C58 across different genres, in order to evaluate the importance of the current finding and eventually trace new associations between discourse relations and various text genres.

Highlighted with red in the third column of the tables 2-5 are the cases where a statistically significant association has been observed between discourse relations and one of the variables. Therefore, the question of section 6 is not only answered positively, namely there is strong evidence that the two levels of linguistic description interact, but also the interdependence of the two is observed in specific cases and engaging conclusions may be drawn upon. Summing up the statistically significant rows in table 6, we note the following:

Utterance combinations	Statistically significant cases
Valency_2 – Valency_2:	• The *thematic role of the object and the aspect of the second utterance, both subject types for the related utterances* and the type of discourse relation are interdependent.
Valency_2 – Valency_1:	• The *thematic role of the subject and the subject type of the second* ut*terance* and the type of discourse relation are interdependent.
Valency_1 – Valency_1:	• No dependence is traced!
Valency_1 – Valency_2:	• Both *thematic roles, the verbal aspect and the subject type of the second utterance* and the type of discourse relation are interdependent.

Table 6 | Statistically significant cases for all valency combinations.

8. Conclusion

A systematic way of extracting annotation patterns even in more complicated levels of description, such as in the discourse level, may considerably boost the effectiveness of modern applications in the long run, as a part of an intermediate semi-automatic training phase, and allow us to understand even further the various linguistic interfaces. The paper aimed to stress the usefulness of linguistically informed discourse annotation for improving CL systems and for providing a data-analytic way in order to study the interface between inter- and intra-sentential levels. Real use utterances sampled within C58, the first corpus of annotated discourse relations for Greek, have been exploited quantitatively for evaluating qualitative domain-specific knowledge, such as the interdependence

between discourse relations and a number of intra-sentential grammatical factors (i.e. grammatical aspect, thematic roles, subject type). In section 5 we included only a small part of the difficulties we faced throughout the process of annotating the data and drew some basic annotation guidelines for discourse annotation that can be exploited within the frame of any other discourse annotation project. The second part of the paper parsed through some first independence tests for tracing associations between the observed frequencies of the annotated grammatical factors and the type of discourse relations.

The next step for a data-analytic exploration of our annotated resource, C58, will be to focus on the details of these interdependencies and prioritize them. The first conclusions have shown that there is a clear pattern in the interdependence of inter- and intra-sentential variables that need to be further investigated. In other words, the next question to be answered is *which of the statistically significant intra-sentential factors seem to be more important for deriving discourse inferences and which discourse relations dictate intra-sentential preferences?*

References

Asher, Nicholas, and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge: Cambridge University Press.

Asher, Nicholas. 1993. *Reference to Abstract Objects in Discourse, Studies in linguistics and philosophy*. Dordrecht, Boston (Mass.), London: Kluwer Academic.

Gardent, Claire, and Bonnie Webber. 1998. "Describing Discourse Semantics." In *Proceedings of the 4th TAG Workshop*. Philadelphia: University of Pennsylvania.

Kiss, Tibor, and Jan Strunk. 2006. "Unsupervised Multilingual Sentence Boundary Detection." *Computational Linguistics* 32(4):485–525.

Krifka, Manfred. 2008. "Basic Notions of Information Structure." Acta Linguistica Hungarica 55:243-276.

Levin, Beth and Rappaport Hovav, M. 2005. *Argument Realization*. New York: Cambridge University Press.

Marcu, Daniel. 1998. "A Surface-Based Approach to Identifying Discourse Markers and Elementary Textual Units in Unrestricted Texts". In *Proceedings of the COLING/ACL'98 Workshop on Discourse Rela-*

tions and Discourse Markers, edited by Manfred, Leo Werner, and Eduard Hovy, 1-7. New Brunswick, NJ: Association for Computational Linguistics.

Marcu, Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA, USA: MIT Press.

Marcu, Daniel, and Abdessamad Echihabi. 2002. "An Unsupervised Approach to Recognizing Discourse Relations." In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 368–375. Stroudsburg, PA, USA: Association for Computational Linguistics. Accessed December 2016. http://dx.doi.org/10.3115/1073083.1073145.

Marcu, Daniel, Magdalena Romera, and Estibaliz Amorrortu. 1999. "Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices, Issues." In *Workshop on Levels of Representation in Discourse*, 71–78. University of Maryland.

Polanyi, Livia. 1988. "A Formal Model of the Structure of Discourse." *Journal of Pragmatics* 30(2):35–175.

Polanyi, Livia, Chris Culy, Martin van den Berg, Gian Lorenzo T., and David Ahn. 2004. "Sentential Structure and Discourse Parsing." In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, 80–87. Stroudsburg, PA, USA: Association for Computational Linguistics. Accessed December 2016. http://dl.acm.org/citation.cfm?id=1608938.1608949.

Pustejovsky, James, and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: a Guide to Corpus-Building for Applications*, Sebastopol, CA: O'Reilly Media.

Schuler, Karin K. 2006. "VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon." Unpublished Phd diss., University of Pennsylvania .

Van Valin, Robert D. Jr. 2005. *Exploring the Syntax–Semantics Interface*. Cambridge: Cambridge University Press.