Russian Journal of Linguistics

2021 Vol. 25 No. 2 478—506
http://journals.rudn.ru/linguistics

DOI: https://doi.org/10.22363/2687-0088-2021-25-2-478-506

Research article

Inter-annotator agreement in spoken language annotation: Applying $_u\alpha$-family coefficients to discourse segmentation[1]

Salvador PONS BORDERÍA and Elena PASCUAL ALIAGA

Universidad de Valencia
Valencia, Spain

Abstract

As databases make Corpus Linguistics a common tool for most linguists, corpus annotation becomes an increasingly important process. Corpus users do not need only raw data, but also annotated data, submitted to tagging or parsing processes through annotation protocols.

One problem with corpus annotation lies in its reliability, that is, in the probability that its results can be replicable by independent researchers. Inter-annotation agreement (IAA) is the process which evaluates the probability that, applying the same protocol, different annotators reach similar results. To measure agreement, different statistical metrics are used. This study applies IAA for the first time to the *Valencia Español Coloquial* (Val.Es.Co.) discourse segmentation model, designed for segmenting and labelling spoken language into discourse units. Whereas most IAA studies merely label a set of in advance pre-defined units, this study applies IAA to the Val.Es.Co. protocol, which involves a more complex two-fold process: first, the speech continuum needs to be divided into units; second, the units have to be labelled. Kripendorff's $_u\alpha$-family statistical metrics (Krippendorff et al. 2016) allow measuring IAA in both segmentation and labelling tasks. Three expert annotators segmented a spontaneous conversation into subacts, the minimal discursive unit of the Val.Es.Co. model, and labelled the resulting units according to a set of 10 subact categories. Kripendorff's $_u\alpha$ coefficients were applied in several rounds to elucidate whether the inclusion of a bigger number of categories and their distinction had an impact on the agreement results. The conclusions show high levels of IAA, especially in the annotation of procedural subact categories, where results reach coefficients over 0.8. This study validates the Val.Es.Co. model as an optimal method to fully analyze a conversation into pragmatically-based discourse units.

Keywords: *corpus annotation, inter-annotator agreement, Kripendorff's $_u\alpha$-coefficients, discourse segmentation, Val.Es.Co. Model, subacts*

For citation:

Pons Bordería, Salvador & Elena Pascual Aliaga. 2021. Inter-annotator agreement in spoken language annotation: Applying $_u\alpha$-family coefficients to discourse segmentation. *Russian Journal of Linguistics* 25 (2). 478–506. DOI: https://doi.org/10.22363/2687-0088-2021-25-2-478-506

[1] This paper was made possible by the research project Project FFI2016–77841-P, *Unidades discursivas para una descripción sistemática de los marcadores del discurso en español* (UDEMADIS), funded by the *Ministerio de Economía y Competitividad/ AEI and the ERC.*

Научная статья

Согласие между аннотаторами при аннотировании разговорной речи: применение $_{u}\alpha$-коэффициентов к сегментации дискурса[2]

Сальвадор ПОНС БОРДЕРИА, Елена ПАСКУАЛЬ АЛИАГА

Валенсийский университет
Валенсия, Испания

Аннотация

Благодаря появлению баз данных корпусная лингвистика становится привычным инструментом для большинства лингвистов. Именно поэтому аннотирование корпусов приобретает все большую значимость. Пользователям корпусов нужны не только сырые, но и аннотированные данные, т. е. размеченные с применением протоколов аннотирования и методов синтаксического анализа (парсинга). Одна из проблем, с которой сталкиваются исследователи при аннотировании корпуса, – это проблема надежности, то есть возможности воспроизведения результатов исследования независимыми исследователями. Согласие между аннотаторами (IAA) – это методика оценивания вероятности того, что, применяя один и тот же протокол, разные аннотаторы получат одинаковые результаты. Для измерения согласия используются разные статистические показатели. Представленное исследование впервые применяет IAA к модели сегментации дискурса *Valencia Español Coloquial* (Val.Es.Co.), предназначенной для сегментации и разметки единиц устного разговорного дискурса. В отличие от преимущественного большинства исследований IAA, в которых только маркируется набор заранее определенных единиц, в данном исследовании IAA применяется в рамках Val.Es.Co.-протокола, предусматривающего более сложный двухступенчатый процесс: во-первых, речевой континуум разделяется на дискурсивные единицы; во-вторых, осуществляется разметка дискурсивных единиц. Статистические показатели $_{u}\alpha$ -семейства Криппендорфа (Krippendorff et al. 2016) позволяют измерять IAA как в задачах сегментации, так и в задачах разметки. Три эксперта-аннотатора разделили спонтанную речь на субакты, минимальные дискурсивные единицы Val.Es.Co.-модели и разметили полученные единицы в соответствии с набором из 10 подкатегорий. $_{u}\alpha$-коэффициенты Криппендорфа применялись в нескольких экспериментах, чтобы выяснить, повлияло ли включение большего числа категорий и их различие на результаты IAA. Мы получили высокие уровни IAA, особенно в аннотации процедурных категорий субактов, где результаты достигают коэффициентов выше 0,8. Таким образом, исследование подтверждает, что Val.Es.Co.-модель является оптимальным методом для полной сегментации речи на прагматически мотивированные дискурсивные единицы.

Ключевые слова: *аннотирование корпусов, согласие между аннотаторами, $_{u}\alpha$-коэффициенты Криппендорфа, сегментация дискурса, Val.Es.Co. Model, субакты*

[2] Написание этой статьи стало возможным благодаря исследовательскому проекту Project FFI2016–77841-P, *Unidades discursivas para una descripción sistemática de los marcadores del discurso en español* (UDEMADIS), финансируемому Министерством экономики и конкурентоспособности */ AEI и ERC.*

Для цитирования:
Pons Bordería S., Pascual Aliaga E. Inter-annotator agreement in spoken language annotation: Applying $_u\alpha$-family coefficients to discourse segmentation. *Russian Journal of Linguistics.* 2021. Vol. 25. № 2. P. 478–506. DOI: https://doi.org/10.22363/2687-0088-2021-25-2-478-506

1. Introduction

Consulting electronic corpora to retrieve examples of use has become a standard research method for linguists working in fields like Pragmatics. This retrieval process depends on a previous annotation of such corpora. Today, annotation can be a completely automatized process in simpler tasks like tagging words, yet more complex tasks (like determining discourse relationships among words, sentences or paragraphs) require human intervention: trained linguists analyze and annotate corpora, alone or in teams, guided by annotation protocols. The more reliable the protocol, the better the annotation of the corpus.

At this point, a question arises, related to how reliable (meaning 'objective') an annotation protocol can be. 'Objective', in turn, means 'replicable', that is, able to produce the same results if repeated by independent groups of researchers. Inter-annotator agreement (henceforth, IAA) is the process whereby the reliability and the replicability of a corpus annotation protocol are tested (Arstein & Poesio 2008, Artstein 2017). *Reliability* and *replicability* are evaluated by seeking whether the same annotation protocol leads to the same annotation results when applied independently by two or more annotators. Agreement among annotations is measured using chance-correction statistical metrics such as Cohen's kappa (Cohen 1960, 1968, Carletta 1996), Scott's pi (Scott 1955, *cf.* Fleiss 1971) or Krippendorff's-alpha (Krippendorff 1970, 2013). As a result of this measurement, annotation labels, segmentation and the annotation protocol can be validated and thus accepted or rejected.[3] This paper presents a study with the aim to assess the reliability of a specific corpus annotation protocol: the Val.Es.Co. model of discourse units (Briz & Grupo Val.Es.Co. 2003, Grupo Val.Es.Co. 2014), designed for analyzing spoken, spontaneous conversations.

Beyond its multiple applications, in the field of corpus linguistics IAA has been successfully applied so far[4] to two main aspects of corpus annotation: to the labelling of discourse markers (Crible & Degand 2019a, 2019b, Zufferey and Popescu-Belis 2004), and to the recognition of discourse relations, either explicitly conveyed by discourse connectives or not. In this field, IAA has made use of annotated corpora such as the *Penn Discourse Treebank 3.0* (PDTB) (Prasad et al. 2019, Miltsakaki et al. [2004], Prasad et al. [*2*008]), the Rhetorical Structure Theory Discourse Treebank (RST-DT) (Marcu et al. [1999], Carlson et al. 2003a, 2003b)

[3] A discussion on the complex relationship between reliability and validity can be found in Krippendorff (2013), Spooren and Degand (2010) and van Enschot *et al.* (in press).

[4] Other fields of research on discourse also make use of the IAA methodology: for example, Riou (2015) on the topic transitions in turn-constructional units, Grisot (2015, 2017) on verb tenses.

and the *Prague Discourse Treebank 2.0* (PDiT 2.0) (Rysová et al. 2016; Mírovský et al. [2010]).

In most previous works, IAA is used to measure the fit of a set of labels onto a set of units: in Crible & Degand (2019a), a set of 423 tokens of discourse markers (henceforth, DM) is annotated independently by two expert annotators into thirty-four functional labels hierarchically distributed. Likewise, in Scholman et al. (2016), 40 non-expert annotators annotate Discourse Relations in 36 excerpts containing pre-delimited segments, taking as a basis the theory of the cognitive approach to coherence relations (Sanders et al. 1992, 1993). In this study, 12 hierarchically distributed categories are assigned to each pairing of segments. Common to both studies is the fact that the annotators operate with two closed sets: DM or pairs of utterances, on the one hand, and discourse relationships, on the other.

Valuable as these efforts might be, IAA achieves an extra layer of complexity when the process implies a previous identification of the units to be labelled. In this case, the annotation process involves two consecutive steps, *segmentation* and *labelling*:

a) segmentation means identifying units by setting their boundaries in a given continuum (e. g. in a text or in a conversation);

b) labelling is the assignment of a specific category to each unit.

This twofold procedure constitutes the main endeavour of *discourse segmentation models* (Pons Bordería 2014), which are theoretical proposals aimed at fully dividing speech into units and subunits, just as syntactic analyses do with sentences and phrases. The calculation of IAA is an important step to evaluate the fit of a given model and to compare it to other models on an objective basis. However, IAA has not been applied to *both* processes simultaneously, as this paper does.

To better illustrate this two-step annotation process, recall example (1), where two speakers (S1 and S2) discuss about their preferences regarding two supermarket chains, *Consum* and *Mercadona*:

(1) S1: no me gustan las de Consum me gustan más las de Mercadona
S2: a mí también pero mi madre compró en Consum ayer
[S1: I don't like the ones from Consum I prefer the ones from Mercadona
S2: me too but my mother shopped in Consum yesterday]

Excerpt (1) can be analyzed by two different annotators, say A and B. Their analysis comprises two different tasks: the first one consists of dividing the text into linguistic units, as shown in (1'). The second task consists of labelling the units from a closed set of alternates {x, y, z,…, n}, as shown in (1''). With respect to the first task, differences in interpretation can produce different segmentations. In (1'), annotator A interprets a sequence *abc* as a single unit ([abc]), whereas annotator B analyzes the same sequence as two units ([ab][c]):

(1')

Ann. A	{no me gustan las de Consum me gustan más las de Mercadona} [{I don't like the ones from Consum I prefer the ones from Mercadona}][5]
Ann. B	{no me gustan las de Consum} {me gustan más las de Mercadona} [{I don't like the ones from Consum} {I prefer the ones from Mercadona}]

Ann. A	{a mí también pero mi madre compró en Consum ayer} [{me too but my mother shopped in Consum yesterday}]
Ann. B	{a mí también} {pero mi madre compró en Consum ayer} [{me too} {but my mother shopped in Consum yesterday}]

Divergences may arise also in the second task of labelling, as annotators A and B can interpret his sequence differently ([$_{x}$abc$_{x}$] *vs*. [$_{y}$ab$_{y}$][$_{z}$c$_{z}$]), as shown by (1''):

(1'')

Ann. A	{no me gustan las de Consum me gustan más las de Mercadona} $_{DSS}$ [{I don't like the ones from Consum I prefer the ones from Mercadona}$_{DSS}$]
Ann. B	{no me gustan las de Consum}$_{DSS}$ { me gustan más las de Mercadona}$_{DSS}$ [{I don't like the ones from Consum}$_{DSS}$ {I prefer the ones from Mercadona}$_{DSS}$]

Ann. A	{ a mí también pero mi madre compró en Consum ayer$_{DSS}$ } [{me too but my mother shopped in Consum yesterday}$_{DSS}$]
Ann. B	{a mí }$_{DSS}$ {pero mi madre compró en Consum ayer}$_{SSS}$ [{me too}$_{DSS}$ {but my mother shopped in Consum yesterday}$_{SSS}$]

Examples (1') and (1") illustrate the complexity of an annotation process involving segmentation and labelling. Most research on IAA consists of matching a set of labels (pragmatic functions, or discourse relationships) onto a pre-defined set of units (DM, turns or punctuation-delimited sentences). In discourse segmentation, the units themselves have to be established independently by each annotator. Here, agreement is much harder to reach, for not only a good match in the labels-onto-units projection is needed, but this match is dependent on a previous agreement on the segmentation of discourse units. The analysis in this paper reveals a complex approach to IAA, especially considering that i) the object of study are spontaneous conversations, a place where contextual cues must be taken into account for properly identifying units; and ii) the segmentation makes use of syntax, prosodic, semantic and pragmatic information (see 2.2).

[5] The translation of the examples in this paper are segmented, except in those cases where this would lead to an incorrect segmentation, due to the different structures in Spanish and English. In other cases, the translation changes significantly the structure of the Spanish sentence to ensure understandability. In both cases, a correct segmentation would imply a parallel analysis of the English translation, which is far from the goals of this paper.

The annotation process described so far becomes even more complex when more than two annotators are implicated, as the potential sources of divergence multiply and therefore good results are harder to achieve[6].

To sum up, three parameters can be implied in an annotation process:

a) The number of annotators.

b) The segmentation (or not) of the linguistic units as part of the annotation process.

c) The number of labels to be applied.

The complexity of the process is largely dependent on the numbers assigned to these variables. For instance, two annotators labelling a same set of discourse markers with a set of five categories face a total of 2*1*5 = 10 variables. Two annotators labelling a set of eleven discourse relationships on the same pairs of sentences face a total of 2*1*11 = 22 variables. Alternatively, three annotators dividing a full conversation into units – units which can be coincident or not – and assigning a set of eight labels to each unit face a total of 3*2*8 = 48 variables. It is evident that, the more parameters are included in the annotation, the greater differences might be expected.

The metrics selected in this paper are Kripendorff's $_u\alpha$-family coefficients (Krippendorff et al. 2016) and the units to be tested are the subacts, the minimal segments in the Val.Es.Co. model (see 2.3). As subacts organize the distribution of conceptual and procedural information[7] in speakers' turns, IAA evaluates one key feature of a discourse segmentation model, namely the extent to which both kinds of meaning can be robustly accounted for by a single, pragmatically-based analysis.

In what follows, section 2 presents some previous literature on discourse segmentation (§ 2.1) and brings into play the applicability of IAA to proposals of discourse segmentation models. More specifically, the Val.Es.Co. model (§ 2.2), and the statistical techniques for measuring IAA (§ 2.3) are presented in detail. Section 3 explains the methodology in this study. Section 4 shows the results obtained in IAA measurement, and sections 5 and 6 sum up the results and the main findings of this study.

2. When discourse segmentation models met Krippendorf's $_u\alpha$-family coefficients

2.1. Current annotation proposals by discourse segmentation models

Since spoken discourse began to be a focus of interest for linguistic research, it became evident that traditional syntax was too narrow as a segmentation tool (Pons Bordería 2014: 1). Units such as *sentence* or *clause* proved inadequate for

[6] Artstein and Poesio (2005) prove that, as regards tests such as Fleiss'κ and a generalized Cohen's κ, including more annotators is a good way to decrease the so-called annotator bias – the individual preferences of annotators. See also Artstein and Poesio (2008: 570–573).

[7] The conception of procedural meaning used in this paper is limited to *non-propositional procedural meaning*, what equals it with discourse markedness (Briz and Pons Bordería 2010). For a more comprehensive account of procedural meaning, see Wilson (2011) and Grisot (2017).

analyzing spoken language, where some "deviant" language uses ("unachieved" syntactic structures, multifunctional discourse markers or unusual word ordering, just to mention a few) are not the exception, but the rule (Sornicola 1981, Blanche-Benveniste & Jeanjean 1987, Narbona 1986, 1992, 2012, Briz 1998).

The need for a *new syntax* (Narbona 1992) to account for spoken language set the grounds for an emerging area of research on models for discourse segmentation. As Pons Bordería (2014: 1) explains, efforts attempting to find new units for analyzing spoken discourse have been made in particular from Romance languages, where Latin grammar has been traditionally influential. This is evident in the proliferation of various segmentation models[8] in French, Spanish or Italian such as those of Geneva (Roulet et al. 1985, Roulet, Fillietaz & Grobet, 2001), the Sorbonne (Morel & Danon-Boileau 1998), the Val.Es.Co. Research Group (Briz & Grupo Val.Es.Co. 2003, Grupo Val.Es.Co. 2014), Leuven (Degand & Simon 2009a) and Freiburg (Groupe de Fribourg 2012). All these models, while offering different units and divergent criteria to identify them, have in common one aim: segmenting spoken language without leaving any segments unanalyzed.

Segmenting spoken language becomes especially challenging when it comes to smaller-scope units (Degand & Simon [2005, 2009a], Grupo Val.Es.Co. [2014: 12], Briz [2011]). Contrary to higher-scope units such as *turn* or a *dialogue*, identifying smallest scope units requires considering diverse parameters such as prosodic cues, syntactic boundaries or pragmatic information, which must be properly balanced to achieve a sound result. Evaluating such complex segmentation and labelling practices by means of IAA techniques provides a handle for assessing and improving any discourse segmentation proposal.

Despite its beneficial potential, discourse segmentation models have barely made use of IAA techniques. Being most of them theoretical, studies showing the results of applying a segmentation model are the exception (Degand & Simon 2011, 2009b, Latorre 2017, Pascual 2015a, 2015b). To the authors' knowledge, no model has applied IAA to test protocols for segmenting discourse into units.

We believe that IAA contributes to providing a robust way of identifying discourse units, a goal at which segmentation models should aim. Testing the segmentation protocol becomes crucial for developing theories and more robust protocols. This study applies IAA to the Val.Es.Co. model — more specifically, to the unit *subact.*

2.2. The Val.Es.Co. model (VAM) of discourse segmentation

The Val.Es.Co. model of discourse units (henceforth, VAM) (Briz & Grupo Val.Es.Co. 2003, Val.Es.Co. Group 2014) relies on different approaches (Conversation Analysis [Sacks et al. 1974], Discourse Analysis, [Sinclair &

[8] Pons (2014) also explains that proposals made by the segmentation models are based on various fields that lay the foundations of the new units: macrosyntax (Van Dijk 1977), transphrastic approaches (Stati 1990), Conversation Analysis (Sacks et al. 1974) or Discourse Analysis (Sinclair and Coulthard 1975).

Coulthard 1975], the Sorbonne Group [Morel and Danon-Boileau 1998], the Geneva Group [Roulet 1985, Roulet 1991, Roulet et al. 2001]). Since 2003, this framework has been applied to different problems, such as the polyfunctionality of discourse markers (Briz 1998, Briz & Pons 2010, Estellés 2011, Pons 2008), the study of intensification and hedging devices (Albelda 2007, Albelda & Gras 2011), or diachronic approaches in grammaticalization or constructionalization (Pons & Estellés 2009, Pons 2014, Salameh 2021).

The VAM comprises eight hierarchical units (*discourse*, *turn-taking*, *turn*, *dialogue*, *exchange*, *intervention*, *act* and *subact*) located into three dimensions (*social*, *structural* and *informative*) and two levels (monologic and dialogic), as the following table illustrates (Table 1).

Table 1

Units, leveis and dimensions of the VAM (Val.Es.Co. Group 2014: 14

Level	Dimension		
Dialogic	Social	Structural	Informative
	Turn-taking	Discourse Dialogue Exchange	
Monologic	Turn	Intervention Act	Subact

In this top-to-bottom model, wider-scope units have scope over smaller-scope units (e.g. interventions have scope over acts, exchanges have scope over interventions, and so forth). Speaking is conceived as an activity involving three dimensions: first, speaking is a *social* activity, where speaker and hearer interact; second, speaking is a *structural* activity, consisting of uttering language (including disfluency phenomena such as false starts or truncated segments); finally, speaking is and an *informative* activity, whereby information is packed into units.

The *act* and *subact* units are monological, whereas *exchange*, *turn*, *turn-taking*, *discourse* and *dialogue* are dialogical units. In turn, the unit *intervention* is, at the same time, monological and dialogical, as the maximal projection in speaker's production and, at the same time, the minimal content aimed at interacting with other participants. Dimensions, levels and units are interrelated and allow for a complete segmentation of a conversation.

The IAA study in this paper focuses on the smallest unit in the VAM – the *subact* – conceived as the smallest piece of information delivered by a speaker. As such, it is perhaps the most difficult unit to identify, since the boundaries of informative units intertwine with the syntactic ones (Briz & Grupo Val.Es.Co. 2014)[9].

[9] As exemplified by the traditional definition of a sentence as "a unit with full meaning" (Bello 1847), or the identification of subordinated clauses with "secondary" meaning, for instance, in the case of conditional clauses.

2.3.1. Subact: definition and types

A subact is defined as the smallest monological and informative unit. Subacts are hierarchically subordinated to a wider-scope unit called *act*; therefore, a subact or a group of subacts constitute an *act,* defined as the host of an illocutionary force (Grupo Val.Es.Co., 2014: 54). Notation-wise, subacts are indicated by braces ({ }) whereas acts are indicated by the hash sign (#).

Subacts are classified into two main categories, depending on the type of information they convey: *substantive subacts* (SS) convey conceptual information, and *adjacent subacts* (AS) convey procedural information. SS are, in turn, subdivided into *directive substantive subacts* (DSS), *subordinated substantive subacts* (SSS) and *topicalized subordinated substantive subacts* (TopSSS). DSS carry the weight of the main content in the act; SSS host semantically secondary or dependent information; TopSS are instances of prosodically or informatively detached constituents:

(2) A: # {y al cine→}$_{\text{TopSSS}}$ {¿vas a venir?}$_{\text{DSS}}$ #
B: # {No puedo}$_{\text{DSS}}$ {porque tengo que estudiar}$_{\text{SSS}}$
[A: # {and to the cinema→}$_{\text{TopSSS}}$ {are you coming?}$_{\text{DSS}}$ #
B: # {I cannot go}$_{\text{DSS}}$ {because I should prepare for my exam}$_{\text{SSS}}$ #]

In example (2), the TopSSS "and to the cinema→" is prosodically detached from the segment that conveys the main illocutionary force: "are you coming?". At the same time, the TopSSS is informatively dependent on the DSS (otherwise, the prototypical ordering of the utterance might be "and are you coming to the cinema?"). On the other hand, the SSS "because I should prepare for my exam" depends on the DSS "I cannot go" (as shown by the subordination conjunction *because*) and contains the explanation derived from the negative assertion made by A (Salameh, Estellés & Pons, 2018: 115). This SSS could be removed without changing the illocutive force of the intervention — a refusal; its subordinated nature lies on the fact that B would not be able to answer to A's previous intervention with just the SSS, as shown in (2'):

(2') A: # {y al cine→}$_{\text{TopSSS}}$ {¿vas a venir?}$_{\text{DSS}}$ #
B: # {porque tengo que estudiar}$_{\text{SSS}}$
[A: # {and to the cinema→}$_{\text{TopSSS}}$ {are you coming?}$_{\text{DSS}}$ #
B: # {because I should prepare for my exam}$_{\text{SSS}}$ #]

AS convey procedural information and can be further divided into *Textual Adjacent Subacts* (TAS), *Modal Adjacent Subacts* (MAS) and *Interpersonal Adjacent Subacts* (IAS): TAS (like *then*, *moreover*, or *hence*) relate chunks of message. MAS (like *well*, *oh*, or *just*) convey the relationship between the speaker and his own message. Finally, IAS (like *see?, right?,* or *look*) convey the relationship between speakers and hearers:

(3) A: # {las llaves↑}$_{TopSSS}$ {bueno}$_{MAS}$ {es quee}$_{TAS}$ {no te las puedo dar}$_{DSS}$ / {porque las necesito}$_{SSS}$ {¿sabes?}$_{IAS}$
[A: # {the keys↑}$_{TopSSS}$ {well}$_{MAS}$ {es quee}$_{TAS}$ {I cannot give them to you}$_{DSS}$ / {because I need them}$_{SSS}$ {you know?}$_{IAS}$]

Together, these six labels (DSS, SSS, TopSSS, TAS, MAS and IAS) account for most of the distribution of information in a spontaneous conversation. However, in spontaneous conversations, some constituents remain unachieved, reflecting processes in language-planning (Ochs 1979, Sornicola 1981). These fragmentary units pose a problem for any discourse segmentation model, since by nature of their unachieved status, they cannot be classified as AS or SS. According to their degree of completion, the Val.Es.Co. model classifies them as XSS (an incomplete constituent with conceptual content), ASX (an incomplete constituent with procedural content), XXS (an incomplete constituent whose conceptual or procedural nature cannot be established), and R (a sub-structural, residual element in the analysis)[10] (Pons Bordería [2016] and [Pascual 2018, 2020]). Example (4) shows some of these fragmentary units:

(4) M: # {no/}$_{DSS}$ # #{eso / ha sali- /}$_{XSS}$ {m- m/}$_{R}$ {((ee))}$_{TAS}$ {más ha salido de tu boca que en la televisión/}$_{DSS}$ {y-/}$_{XSS}$ {porque yo solamente te lo he visto a ti}$_{SSS}$ #
[M: # {no/}$_{DSS}$ {this/ has com-/}$_{XSS}$ {m- m/}$_{R}$ {((eeh))}$_{TAS}$ {has come more out of your mouth than out of television/}$_{SSS}$ {and-/}$_{XSS}$ {because I've only seen that in you}$_{SSS}$ #]

2.3. Statistical tests: Krippendorff's $_{u}\alpha$-family coefficients

Krippendorff (1995, 2003, 2013) and Krippendorff et al. (2016) have developed a family of statistical coefficients in order to measure agreement not only in the labelling of units by different annotators, but also in the segmentation of units in a continuum not previously pre-segmented, – i. e. in cases where there is not a total number of pre-established units for each annotator to label. This family comprises four coefficients: $_{u}\alpha$, $|_{u}\alpha$, $_{cu}\alpha$ and $_{(k)u}\alpha$. In the case of IAA, the variables taken into account by those tests are the following:

a) The *location* of the units in the continuum: this variable measures if two or more annotators have identified a same unit in the same time span.

b) The *length* of the units: this variable measures whether a unit measures the same number of milliseconds, even if not being placed in exactly the same minute and second in the conversation.

c) The total *number* of annotated units in a given span of time.

d) The *type* or label of the annotated unit.

These variables stay in close relationship with the goals of a two-fold annotation process like the one performed in this paper: on the one hand, the segmentation process involves a) placing and b) c) bounding subacts; on the other

[10] For the relationship between this category and the concept *disfluency*, see Pascual (2020).

hand, the labelling process also implies d) categorizing the types of subacts previously identified in a conversation.

Adapting the example provided by Krippendorff et. al. (2016: 2349), Figure 1 illustrates what happens when three different annotators (A, B and C) segment and annotate a conversation into subacts. The columns (1), (2), (3), (4) and (5) show the different possibilities of the analysis and, therefore, the variables taken into account by the four $_{u}\alpha$-family coefficients:

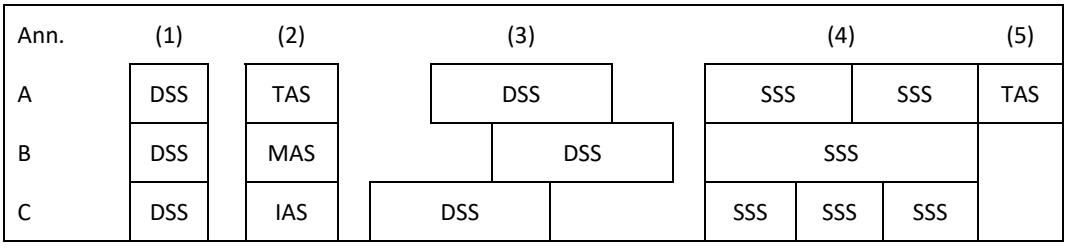

Figure 1. Possibilities for measuring agreement (adapted from Krippendorff et. al. [2016: 2349])

In column (1), all the three annotators agree in the segmentation and in the labelling of all the variables, since the units coincide in their location, length, number and type; in (2), the units show the same segmentation (location, length and number), but differ with respect to their labels (TAS, MAS and IAS); in (3), the units are not equally segmented (they are located in different time spans, albeit coinciding in length, and number) but are equally labelled (DSS in all cases); in (4), the units are equally labelled, but differ in their segmentation (they occur in the same time span, but differ in number and length); finally, in (5) there is not any agreement neither in segmentation nor in labelling (annotator A identifies a TAS while annotators B and C do not identify a linguistic unit at all).

Thus, Krippendorf's $_{u}\alpha$-family coefficients provide indicators allowing to measure agreement in both segmenting and labelling procedures. This is why Krippendorf's metrics have been chosen for measuring IAA, in contrast with other statistical tests that measure only categorical agreement in labelling such as Cohen's kappa, Fleiss' kappa or Scott's pi.[11]

The $_{u}$a, $|_{u}\alpha$, cuα and $_{(k)u}\alpha$ coefficients provide information about different aspects of the reliability of the annotation and vary in two essential points, namely in the way they compute agreement and in the type of data they take into account:

a) $_{u}\alpha$ measures overall agreement in all data, this meaning that the calculation includes both units and no-units: in our case, pauses, silences and gaps between

[11] According to Krippendorff et al. (2016: 2349), Guetzkow (1950) defined a coefficient to measure the reliability on unitizing data (i.e. identifying units on a given continuous data). However, Krippendorff et al. (2016) affirm that Guetzkow's test has several drawbacks: i) it is only applicable when a total of two annotators participate in the annotation procedure, ii) it measures disagreement of the number of units identified, but is unable to assess reliability on the agreed units and iii) the result does not provide any information about whether the identified units overlap or whether they are related in any way (i.e. have the same or a different duration).

subacts and turns); therefore, the final results contemplate data irrelevant of the annotation;

b) $_{|u}\alpha$ reduces data to a binary metric (gap vs. no-gap), and does not specify the distinction between categories; this is useful to show the agreement in the segmentation of a continuum into units; however, it does not inform about the labelling performed by each annotator;

c) $_{cu}\alpha$ shows agreement only on the units that have been assigned a value by all annotators (in our case, contemplating all types of subacts);

d) $_{(k)u}\alpha$ goes a step beyond and specifies the agreement results for each individual label in the analysis, that is, for each subact type (DSS, SSS, MAS, TAS, etc.).

In conclusion, the Krippendorff coefficients can be understood as a set of tools, leading to successive refinements of the IAA analysis: from units and no-units or gaps ($_{u}a$), to the number of units and no-units per annotator, irrespective of their labelling ($_{|u}a$); and from the labelling of all categories as a whole, excluding gaps ($_{cu}a$), to a more fine-grained account of each category in particular ($_{(k)u}a$).

3. Data and procedure

A 19 minute-long, informal conversation (4352 words) from the *Val.Es.Co. 2.0* corpus (Cabedo and Pons 2013) was segmented and labelled into different types of subacts by three expert annotators. All annotators have a degree in Linguistics, are familiar with the VAM model and have applied it previously. The annotators used the audio and the transcription files for the annotation process. They also received specific instructions and a clearly-formulated annotation scheme. The annotators carried out the segmentation and annotation of subacts independently from each other. The variables and values involved in this experiment were the following:

a) Number of annotators: annotator A, annotator B and annotator C

b) Temporal overlapping of units[12]
   i) Yes
   ii) No

c) Labels for types of subacts:
   i) SS: DSS, SSS, TopSSS, XSS
   ii) AS: TAS, MAS, IAS, XAS
   iii) XXS
   iv) Residuals

The number of possible labels for any given constituent is 10. Taking into account that agreement was measured only for the units that did not overlap in time, and that the number of annotators was three; this means that, for any constituent annotated, agreement possibilities were 1/(10*3*2).

[12] Krippendorff's *$_{u}\alpha$-family coefficients* cannot compute units overlapping in the same time span. See Section 4.1.

Once the task was completed, the annotation results were transferred to an Excel sheet, overlapped units were suppressed from the data[13] and Krippendorff's statistical $_{u}\alpha$-family coefficients were applied using the software provided by Krippendorff et al. (2016) in order to measure IAA. As the Krippendorff coefficients provide successive refinements, each test becomes informative of the fit of the analysis.

Successive rounds for calculating IAA were applied to different groupings of the same data, so as to elucidate to which extent working with a bigger number of variables had an impact on the agreement results: first, the labels were reduced to the more general categories AS and SS, in order to measure the agreement related to the procedural vs. conceptual distinction; second, taking into account all the labels representing the 10 types of subacts (DSS, SSS, TAS, MAS, etc.); and third, focusing specifically on the subtypes of procedural subacts (AS) with the aim to observe agreement on the identification of the textual, interpersonal and modal discourse functions. In each step, the analysis was performed twice in order to elucidate whether the presence or absence of the most residual subacts – undetermined subacts (XSS, XAS, XXS) and residuals (R) – influences IAA results.

4. Inter-annotation agreement results

The following sections present the results of the study. Section 4.1. displays the raw data in the quantification of the subacts and provides an insight into the performance of the three annotators. Section 4.2 shows the results of Krippendorff's coefficients in the different rounds of analysis: starting with the labels representing procedural and conceptual subacts (4.2.1 and 4.2.2), continuing with subacts conveying procedural information (4.2.3 and 4.2.4), and finishing with all the types of subacts (4.2.5 and 2.4.6). In all cases, the analysis is carried out twice, so that it can be checked out the effect of including and excluding from the calculation the most residual subact categories (XSS, XAS, XXS and R).

4.1. General results

Table 2 shows the number of units per annotator (named A, B and C). A first overview of the data shows that the total number of subacts identified by the three annotators is very similar (A n = 1331, B n = 1339, C n = 1325). This is a positive signal, especially taking into account the relatively high number of variables in the analysis.

Two additional columns indicate the number of subacts that could be computed using Krippendorff's coefficients: recall that Krippendorff's statistics cannot be applied to units overlapping in the same time-span. Due to the nature of spontaneous conversations, in this analysis overlapping affects 30.5 % of the annotated subacts, which could not be calculated and were removed from the analysis. All in all, 2776 is a relatively big number of units for measuring IAA.

[13] See previous note.

Table 2

Total of subacts annotated by each annotator

LABELS		Ann. A	Ann. B	Ann. C	Total included in Kripendorff's computation (non-overlapped subacts)	Total excluded from Kripendorff's computation (overlapped subacts)	Total
SS	DSS	662	652	635	1404 (72.04%)	545 (27.96%)	1949
	SSS	51	61	88	143 (71.50%)	57 (28.50%)	200
	TopSSS	4	14	14	27 (84.38%)	5 (15.63%)	32
	XSS	38	44	39	99 (81.82%)	22 (18.18%)	121
AS	TAS	210	193	188	445 (75.30%)	146 (24.70%)	591
	MAS	166	170	185	352 (67.56%)	169 (32.44%)	521
	IAS	87	95	75	223 (86.77%)	34 (13.23%)	257
	ASX	2	0	0	1 (50 %)	1 (50 %)	2
XXS		24	33	32	47 (52.81%)	42 (47.19%)	89
Residual		87	77	69	35 (15.02%)	198 (84.98%)	233
Total		**1331**	**1339**	**1325**	**2776 (69.49 %)**	**1219 (30.51 %)**	**3995**

Example (5) illustrates the contexts and frequency of overlapped speech — indicated by the sign "[ ]" — :

(5) S3: ellos son Dioos yy te dicen→ // [cuando tienes-]
S1: [es la-] es laa entidad de la Comunidad Valenciana↑ ¡no!
[de Europa]
S3: [de la Unión EuroPEA] que mejor [paga→ =]
S1: [que mejor paga tía] §
S3: § = a [los monitores]
S2: [((¡qué barbaridad!)) / (( )) qué- =]

[S3: they feel like God aand they tell you→ // [when you-]
S1: [it's the-] it's thee entity of the Valencian Region↑ no! [of Europe]
S3: [of the EuroPEAan union] that pays [best→ =]
S1: [that pays best dude] §
S3: § = [to instructors]
S2: [((how incredible!)) / (( )) how- =]]

In example (5), speakers (S3) and 1 (S1) are repeatedly trying to take the floor. The restart ("it's the-") and the co-construction of the collaborative intervention ("[of the European UNION] that pays [best→ to instructors] ") are illustrative of the competition to get the floor. In turn, Table 3 shows that most of the excluded subacts (represented by the sign "Ø" in Table 3) belong to sub-structural categories such as XXS (47.19 %) or R (84.9 %), as these categories are frequent in overlapped speech and are often embedded within wider-scope units (a DSS, in the case of "it's the-") (Table 3).

Table 3

Overlapping constituents removed from the analysis

Annot.	Annotation	Annotation subject to calculation
A	S1: {{[es la-]}$_{R}$ es laa entidad de la Comunidad Valenciana↑}$_{DSS}$ {¡no!}$_{TAS}$ {[de Europa↑ que mejor paga}$_{SSS}$ {tía]}$_{IAS}$ [it's the- it's thee entity of the Valencian Region↑ no! of Europe that pays best dude]	S1: Ø es laa entidad de la Comunidad Valenciana↑}$_{DSS}$ {¡no!}$_{TAS}$ Ø [it's the- it's thee entity of the Valencian Region↑ no!]
B	S1: {{[es la-]}$_{R}$ es laa entidad de la Comunidad Valenciana↑ {¡no!}$_{TAS}$ [de Europa↑ que mejor paga}$_{DSS}$ {tía]}$_{IAS}$ [it's the- it's thee entity of the Valencian Region↑ no! of Europe that pays best dude]	S1: Ø es laa entidad de la Comunidad Valenciana↑ {¡no!}$_{TAS}$ Ø [it's the- it's thee entity of the Valencian Region↑ no!]
C	S1: {{[es la-]}$_{R}$ es laa entidad de la Comunidad Valenciana↑ {¡no!}$_{TAS}$ [de Europa↑ que mejor paga}$_{DSS}$ {tía]}$_{MAS}$ [it's the- it's thee entity of the Valencian Region↑ no! of Europe that pays best dude]	S1: Ø es laa entidad de la Comunidad Valenciana↑ {¡no!}$_{MAS}$ Ø [it's the- it's thee entity of the Valencian Region↑ no!]

4.2. Inter-annotator agreement results: $_{u}\alpha$, $|_{u}\alpha$, $_{cu}\alpha$ & $_{(k)u}\alpha$

4.2.1. Conceptual versus procedural labels (SS, AS)

Table 4 shows the results based on a first distinction between constituents with conceptual or procedural meaning (SS. vs. and AS). The second row in the table shows the results of including XSS and R in the analysis. The IAA results are high in all cases.

Table 4

IAA results for conceptual and procedural labels

Categories	$_{u}\alpha$	$\|_{u}\alpha$	$_{cu}\alpha$	$_{(k)u}\alpha$
AS, SS	0.825	0.841	0.843	(AS)u-α=0.844 (SS) u-α=0.841
AS, SS, XXS, R	0.823	0.853	0.813	(AS)u-α=0.842 (SS)u-α=0.818 (XXS)u-α=0.626 (R)u-α=0.107

The positive results show that the conceptual-procedural distinction is clear-cut. In the case of $_{u}\alpha$ (= 0.825 / 0.823)[14] and $|_{u}\alpha$ (= 0.841 / 0.853), it must not be forgotten that inter- and intra-speaker pauses are taken as if they were labelled units. This means that the gaps between turns and pauses are also computed, even if they have not been labelled. Yet, the results of $_{cu}\alpha$ (0.843 / 0.813) and $_{(k)u}\alpha$ (AS = 0.844 / 0.842, SS = 0.841 / 0.818) show that once the gaps and pauses are excluded from the calculation, the agreement in the segmentation is still high, as shown by example 493:

[14] The result concerning the first row (i. e. analysis of data excluding residual units) are presented in the first place and followed by results of the second row (including residuals in the calculation).

(6) S2: ee pasé dos días bailando / mira[15] // [¡las secuelas!]
S1: [(RISAS)]
S3: [¿pero qué te ha pasao en] el ojo?
S1: pues que me caí / ((puees)) bebí un poquito de rusc→ /// de rusco (RISAS)

[S2: ee I spent two days dancing / look[16] // [the consequences!]
S1: [(LAUGH)]
S3: [but what happened] to your eye?
S2: well [que] I fell / ((well)) I drank a little bit of rusc→ /// of rusco (LAUGH)]

Example (493) is segmented by Annotator A into three SSs, whereas annotators B and C identify two SSs. All annotators agreed in considering the constituent "I spent two days dancing / look// [(at) the consequences!]" as a SS, even if its boundaries remain not as clear. Also, all three annotators identified the filler "ee" as procedural (AS) (Table 5).

Table 5

Annotation of example (493)

Annotator	Annotation
A	S2: {ee}$_{AS}$ {pasé dos días bailando}$_{SS}$ / {mira}$_{SS}$ // {[¡las secuelas!]}$_{SS}$ [S2: ee I spent two days dancing / look[17] // [the consequences!]]
B	S2: {ee}$_{AS}$ {pasé dos días bailando}$_{SS}$ / {mira // [¡las secuelas!]}$_{SS}$ [S2: ee I spent two days dancing / look[18] // [the consequences!]]
C	S2: {ee}$_{AS}$ {pasé dos días bailando}$_{SS}$ / {mira // [¡las secuelas!]}$_{SS}$ [S2: ee I spent two days dancing / look[19] // [the consequences!]]

Neither the identification of boundaries nor the distinction between conceptual and procedural content are challenged by the inclusion in the analysis of residual categories, as proven by the prevalent high results in the different scores. Although the total number of XXS (n= 47) and R (n= 35) included in the calculation only constitutes the 2.95 % of the total number of subacts (n= 2776), the $_{(k)u}\alpha$ scores are fairly good in the case of XXS (0.626), this notwithstanding the controversial nature of residuals. Indeed, residuals are sub-structural elements whose status as a pragmatic or semantic unit remains still unclear among scholars (Crible & Pascual 2019, Pascual 2020). Example 493) is a nice illustration of how residuals are well accounted for by the model:

(7) S1: [¡hombre!] yo me acuerdo que Alba para su oposición tenía soloo (1.8) veinticinco temas / y en- y en- / porque ((se hizo un)) magisterio // yy Filología tiene SETENTA Y cinco (1.1) es- es- es una diferencia notable // ¡es el triple!

[15] Pointing at her face.
[16] Pointing at her face.
[17] Pointing at her face.
[18] Pointing at her face.
[19] Pointing at her face.

> [S1: [well!] I remember that Alba had onlyy (1.8) twenty-five topics in her competiion / and in- and in- / because ((she went for a competion on)) Education // Literature has SEVENTY-five topics (1.1) it's- it's- it's a remarkable difference // it's three-times more!]

Truncations such as *y en- y en-* ("and in- and in-") or *es- es* ("it's- it's-") are correctly identified by all three annotators as residual categories (see Table 6 below). In any case are residuals annotated as AS or SS, and disagreements remain limited to choosing between the two labels in this category, that is, between XXS or R.

Table 6

Annotation of example 493)

Annotator	Annotation
A	S1:{[¡hombre!]}$_{AS}$ {yo me acuerdo que Alba para su oposición tenía soloo (1.8) veinticinco temas}$_{SS}$ / {y en- y en-}$_{XXS}$ / {porque ((se hizo un)) magisterio}$_{SS}$ // {yy filología tiene SETENTA Y cinco}$_{SS}$ (1.1) {es- es-}$_{R}$ {es una diferencia notable}$_{SS}$ // {¡es el triple!}$_{SS}$ [S1: [well!] I remember that Alba had onlyy (1.8) twenty-five topics in her competiion / and in- and in- / because ((she went for a competion on)) Education // Literature has SEVENTY-five topics (1.1) it's- it's- it's a remarkable difference // it's three-times more!]
B	S1:{[¡hombre!]}$_{AS}$ {yo me acuerdo que Alba para su oposición tenía soloo (1.8) veinticinco temas}$_{SS}$ / {y en-}$_{R}$ {y en-}$_{R}$ / {porque ((se hizo un)) magisterio}$_{SS}$ // {yy}$_{AS}$ {filología tiene SETENTA Y cinco}$_{SS}$ (1.1) {es-}$_{R}$ {es-}$_{R}$ {es una diferencia notable}$_{SS}$ // {¡es el triple!}$_{SS}$ [S1: [well!] I remember that Alba had onlyy (1.8) twenty-five topics in her competiion / and in- and in- / because ((she went for a competion on)) Education // Literature has SEVENTY-five topics (1.1) it's- it's- it's a remarkable difference // it's three-times more!]
C	S1: {[¡hombre!]}$_{AS}$ {yo me acuerdo que Alba para su oposición tenía soloo (1.8) veinticinco temas}$_{SS}$ / {y en-}$_{XXS}$ {y en-}$_{XXS}$ / {porque ((se hizo un)) magisterio}$_{SS}$ // {yy filología tiene SETENTA Y cinco}$_{SS}$ (1.1) { es- es-}$_{R}$ {es una diferencia notable}$_{SS}$ // {¡es el triple!}$_{SS}$ [S1: [well!] I remember that Alba had onlyy (1.8) twenty-five topics in her competiion / and in- and in- / because ((she went for a competion on)) Education // Literature has SEVENTY-five topics (1.1) it's- it's- it's a remarkable difference // it's three-times more!]

Disagreement in the conceptual vs. procedural distinction is limited to very specific instances of discourse markers, like *que* in *pues que me caí* ("well [que] I fell") in Table 7 or *yy* ("aand") in Table 8. In these cases, the annotators hesitate between considering them *pragmatic* discourse markers (hence, coded as an autonomous AS), or *grammatically integrated* conjunctions (hence, included into a SS):

Table 7

Disagreement among annotators (*que*)

Annotator	Annotation
A	S2: {pues}$_{AS}$ {[*que*]}$_{AS}$ {me caí}$_{SS}$ [S2: well que I fell]
B	S2: {pues}$_{AS}$ {[*que*] me caí}$_{SS}$ [S2: well que I fell]
C	S2: {pues}$_{AS}$ {[*que*] me caí}$_{SS}$ [S2: well que I fell]

Table 8

Disagreement among annotators (*y*)

Annotator	Annotation
A	S1: {yy Filología tiene SETENTA Y cinco}$_{SS}$ [S1: and Literature has SEVENTY-five topics]
B	S1:{yy}$_{AS}$ { Filología tiene SETENTA Y cinco}$_{SS}$ [S1: and Literature has SEVENTY-five topics]
C	S1:{yy Filología tiene SETENTA Y cinco}$_{SS}$ [S1: and Literature has SEVENTY-five topics]

In conclusion, as for what regards the first, basic distinction between conceptual and procedural categories, the IAA results obtained here are particularly positive.

4.2.2. Procedural labels (TAS, MAS, IAS)

After this first distinction, the IAA zooms on the three types of AS in the Val.Es.Co. model: textual, modal and interpersonal (TAS, MAS, and IAS). In a further step, the residual XAS label has been added.

To better understand this process, consider example (495):

(8) B: hmm §
C: §¡hijos de la gran puta! ¡cómo [saben!]
B: [¡ala! (RISAS)]
A: [((y aquí)) no acaba el mundo ¿eh?] §
C: § ya ((lo sé)) §
A: § [encima a la ((Laura))]
B: [(( )) (RISAS)] // (RISAS)
A: ee Miguel Nico y Lola↑ / con Laura a la Escola

[B: uhum §
C: § sons of a bitch! how well they [manage!]
B: [woah woah! (LAUGH)]
A: [((and this)) is not the end of it hein?] §
C: § yes ((I know)) §
A: § [on top of it ((Laura))]
B: [(( )) (LAUGH)] // (LAUGH)
A: ee Miguel Nico and Lola↑ / go with Laura to the nursery school]

Table 9 below shows that the performance of annotators is very similar at identifying the boundaries and categories of AS. Apart from some marginal cases, the recognition of AS boundaries and, in most cases, their categorisation as types of subacts shows a high threshold of agreement.

One of such marginal cases is the adverbial particle *encima* (Engl. *on top of it*) in example (3), which is annotated as SS by annotator B, and as SA by annotators A and C. However, A and C diverge in the type of AS assigned to *encima*: modal (MAS), for annotator A, or textual (TAS), for annotator C. A second case of disagreement is "uhum", considered as a TAS functioning as a filler by annotator A, and as an interpersonal marker (IAS) by annotators B and C.

Table 9

Annotation of example 8

Annotator	Annotation
A	B: {hmm}$_{TAS}$ C: Ø B: {¡ala!}$_{MAS}$ {(RISAS)}$_{MAS}$ A: {[((y}$_{TAS}$ Ø {¿eh?}$_{IAS}$ C: Ø A: {[encima}$_{MAS}$ Ø. B: Ø A: {ee}$_{TAS}$ Ø [B: uhum § C: Ø B: [woah woah! (LAUGH)] A: [((and Ø hein?] § C: Ø A: on top of it Ø B: Ø A: ee Ø
B	B: {hmm}$_{IAS}$ C: Ø B: {[¡ala!}$_{MAS}$ {(RISAS)]}$_{MAS}$ A: A: {((y}$_{TAS}$ Ø {¿eh?}$_{IAS}$ C: Ø A: Ø B: Ø A: {ee}$_{TAS}$ Ø [B: uhum § C: Ø B: [woah woah! (LAUGH)] A: [((and Ø hein?] § C: Ø A: on top of it Ø B: Ø A: ee Ø]
C	B: {hmm}$_{IAS}$ C: Ø B: {[¡ala!}$_{MAS}$ {(RISAS)]}$_{MAS}$ A: {[((y}$_{TAS}$ Ø{¿eh?]}$_{MAS}$ C: Ø A: {[encima}$_{TAS}$ Ø B: Ø A:{ee}$_{TAS}$ Ø [B: uhum § C: Ø B: [woah woah! (LAUGH)] A: [((and Ø hein?] § C: Ø A: on top of it Ø B: Ø A: ee Ø

The agreement levels in this new process are again high (see Table 10 below): the $_{u}\alpha$ (0.802), $|_{u}\alpha$ (0.832) and $_{cu}\alpha$ (0.846 / 0.846) metrics all exceed an IAA of 0.8. This means that not only the boundaries of units are clear, but also their categorisation. Remark also that the XAS category (with only two occurrences on 2776 subacts) does not have a negative impact on the overall good agreement results, which remain similar in both cases:

Table 10

IAA results for procedural labels

Categories	$_{u}\alpha$	$\|_{u}\alpha$	$_{cu}\alpha$	$_{(k)u}\alpha$
IAS, MAS, TAS	0.802	0.832	0.846	(IAS)u-α=0.738 (MAS)u-α=0.864 (TAS)u-α=0.870
IAS, MAS, TAS, XAS	0.802	0.832	0.844	(IAS)u-α=0.738 (MAS)u-α=0.864 (TAS)u-α=0.868 (XAS) u-α= 0.000

The fact that $_{cu}\alpha$ is higher than $_{u}\alpha$ and $|_{u}\alpha$ might suggest, as Krippenforff et al. (2016: 2358) put it, that the agreement is due mostly to the labelling of units, not to the segmentation of units and the gaps between them (since gaps are excluded from the calculation, unlike in $u\alpha$ and $|u\alpha$ computation).

As for the $_{(k)u}\alpha$ test, although the only category with a lower level of agreement is IAS (0.738), this is still a highly positive result. The $_{(k)u}\alpha$ value for TAS shows hardly any change when including residual XAS in the analysis (0-870 vs. 0.868). Overall, the model proves to be rather reliable in the segmentation of ASs.

4.2.3. All conceptual and procedural labels (DSS, SSS, TopSSS, TAS, MAS, IAS)

Finally, all the possible labels for conceptual (DSS, SSS, SSSTop) and for procedural (MAS, IAS, TAS) categories are taken into account. The results (*vid.* Table 11) are positive ($_{u}\alpha$ = 0.680 / 0.679, $|_{u}\alpha$ = 0.807 / 0.853, $_{cu}\alpha$ = 0.589 / 0.555), especially taking into account that a high number of labels (amounting to ten, with the inclusion of residual segments) on three different annotations are compared. In fact, distinguishing conceptual from procedural information does not pose a great controversy among annotators, and neither does identifying types of procedural content (see § 4.2.1 and § 4.2.2).

Table 11

IAA results for all conceptual and procedural labels

Categories	$_{u}\alpha$	$\|_{u}\alpha$	$_{cu}\alpha$	$_{(k)u}\alpha$
IAS, MAS, TAS, DSS, SSS, SSSTop	0.680	0.807	0.589	(IAS)u-α=0.620 (MAS)u-α=0.853 (TAS)u-α=0.713 (DSS)u-α=0.578 (SSS)u-α=0.286 (TopSSS)u-α=0.184

Categories	$_{u}\alpha$	$\lvert_{u}\alpha$	$_{cu}\alpha$	$_{(k)u}\alpha$
DSS, SSS, TopSSS, XSS, TAS, MAS, IAS, ASX, XXS, R	0.679	0.853	0.555	(IAS)u-α=0.620 (MAS)u-α=0.853 (TAS)u-α=0.698 (ASX)u-α=0.000 (DSS)u-α=0.554 (SSS)u-α=0.274 (TopSSS)u-α=0.184 (XSS)u-α=0.279 (XXS)u-α=0.628 (R)u-α=0.107

The $\lvert_{u}\alpha$ value being higher than the $_{u}\alpha$, taken together with a lower result of $_{cu}\alpha$, shows that the agreement among annotators arises from the identification of boundaries between units and gaps (units and pauses or silences). The segmentation of TopSSS and SSS shows lower results ((SSS)u-α = 0.286 / 0.274; (TopSSS)u-α = 0.184). In the case of TopSS, the problem may lie in the theoretical definition of the category in the model; as for SSSs, disagreements are probably due to determining how some constituents are informatively subordinated to others without making use of syntactic clues.

With respect to $_{(k)u}\alpha$, the results are still high. The high level of agreement on MAS (0.853) prevails, suggesting that this is the most reliable category among the three annotations.

To understand this last segmentation and labelling phase, consider example (9):

(9) S3: [(( ))] ¿a dónde vas? a las putas monjas [(RISAS)]
S2: [(RISAS)] (1.3) (es)pañol coloquial [(RISAS)]
S1: [sí sí coloquial total]
S2: (RISAS) // qué bueno ¿eh? el español coloquial[20] (LAUGH) (1.3) [(LAUGH)]
S1: [¡qué tía!]
S2: ((o sea)) además es- es lo primero quee que se aprende tío la- las- las cosas así

[C: [(( ))] where do you go? to the fucking nuns [(LAUGH)]
B: [(LAUGH)] (1.3) plain Spanish [(LAUGH)]
A: [yes yes fully plain]
B: (LAUGH) // how great hein? plain Spanish[21] (LAUGH) (1.3) [(LAUGH)]
A: [what a girl!]
B: ((I mean)) plus it's- it's the first thing thaat that you learn dude that- those- those things like that]

Table 12 shows how two pieces of conceptual information in example (9) (*colloquial Spanish* and *those things like that*) are labelled differently: as TopSSS

[20] Laughing.
[21] Laughing.

by annotators B and C, and as DSS by annotator A. Also, in the segment (*where do you go? to the fucking nuns*), annotators A and B identify a single DSS, whereas annotator C identifies a SSS and a DSS. The segmentation of IAS and MAS also proves to be complex, as shown in the status of "hein?" and "dude" as interpersonal cues or modalizers.

Table 12

Annotation of example 9

Annotator	Annotation
A	S3: Ø {¿a dónde vas? a las putas monjas}$_{DSS}$ {[(RISAS)]}$_{MAS}$ S2: {[(RISAS)]}$_{DSS}$ (1.3) {(es)pañol coloquial}$_{DSS}$ {[(RISAS) =]}$_{MAS}$ S1: {[sí sí}$_{DSS}$ {coloquial total]}$_{DSS}$ S2: {(RISAS)}$_{MAS}$ // {qué bueno}$_{DSS}$ {¿eh?}$_{IAS}$ {el español coloquial}$_{DSS}$ {(RISAS)}$_{MAS}$ (1.3) {[(RISAS)]}$_{DSS}$ S1: {[¡qué tíal!]}$_{DSS}$ S2: {((o sea))}$_{TAS}$ {además}$_{TAS}$ Ø {es lo primero Ø que se aprende} $_{DSS}$ {tío}$_{IAS}$ Ø {las cosas así}$_{DSS}$ [C: [(( ))] where do you go? to the fucking nuns [(LAUGH)] B: [(LAUGH)] (1.3) plain Spanish [(LAUGH)] A: [yes yes fully plain] B: (LAUGH) // how great hein? plain Spanish[22] (LAUGH) (1.3) [(LAUGH)] A: [what a girl!] B: ((I mean)) plus it's- it's the first thing thaat that you learn dude that- those- those things like that]
B	S3: Ø {¿a dónde vas? a las putas monjas}$_{DSS}$ {[(RISAS)]}$_{MAS}$ S2: {[(RISAS)]}$_{DSS}$ (1.3) {(es)pañol coloquial}$_{DSS}$ {[(RISAS) =]}$_{MAS}$ S1: {[sí sí}$_{DSS}$ {coloquial total]}$_{DSS}$ S2: {(RISAS)}$_{MAS}$ // {qué bueno}$_{DSS}$ {¿eh?}$_{IAS}$ {el español coloquial}$_{TopSSS}$ {(RISAS)}$_{MAS}$ (1.3) {[(RISAS)]}$_{DSS}$ S1: {[¡qué tíal!]}$_{DSS}$ S2: {((o sea))}$_{TAS}$ {además}$_{TAS}$ Ø {es lo primero Ø que se aprende} $_{DSS}$ {tío}$_{MAS}$ Ø {las cosas así}$_{TopSSS}$ [C: [(( ))] where do you go? to the fucking nuns [(LAUGH)] B: [(LAUGH)] (1.3) plain Spanish [(LAUGH)] A: [yes yes fully plain] B: (LAUGH) // how great hein? plain Spanish[23] (LAUGH) (1.3) [(LAUGH)] A: [what a girl!] B: ((I mean)) plus it's- it's the first thing thaat that you learn dude that- those- those things like that]
C	S3: Ø {¿a dónde vas?}$_{SSS}$ {a las putas monjas}$_{DSS}$ {[(RISAS)]}$_{MAS}$ S2: {[(RISAS)]}$_{DSS}$ (1.3) {(es)pañol coloquial}$_{DSS}$ {[(RISAS) =]}$_{MAS}$ S1: {[sí sí}$_{DSS}$ {coloquial total]}$_{DSS}$ S2: {(RISAS)}$_{MAS}$ // {qué bueno}$_{DSS}$ {¿eh?}$_{MAS}$ {el español coloquial}$_{DSS}$ {(RISAS)}$_{MAS}$ (1.3) {[(RISAS)]}$_{MAS}$ S1: {[¡qué tíal!]}$_{DSS}$ S2: {((o sea))}$_{MAS}$ {además}$_{TAS}$ Ø {es lo primero Ø que se aprende} $_{DSS}$ {tío}$_{MAS}$ Ø {las cosas así}$_{TopSSS}$

[22] Laughing.
[23] Laughing.

Annotator	Annotation
	[C: [(( ))] where do you go? to the fucking nuns [(LAUGH)] B: [(LAUGH)] (1.3) plain Spanish [(LAUGH)] A: [yes yes fully plain] B: (LAUGH) // how great hein? plain Spanish[24] (LAUGH) (1.3) [(LAUGH)] A: [what a girl!] B: ((I mean)) plus it's- it's the first thing thaat that you learn dude that- those- those things like that]

5. Discussion

The results obtained in the different rounds of IAA analysis show a high level of agreement. In most cases, the coefficient values reach a threshold of 0.800; otherwise, the rates are superior to 0.500, with the exception of the most residual subact units. Despite the lack of scientific consensus on what an "acceptable" level of IAA should be (Arstein & Poesio 2008; van Enschot et al. in press; Kripendorff et al. 2016), the application of the Val.Es.Co. annotation protocol for segmenting conversations into subacts yields a very positive outcome, especially taking into consideration the fact that the annotation procedure is complex and involves two tasks: segmenting and labelling units in a conversational continuum.

The successive groupings of categories in the different rounds of analysis lead to differences on IAA results: needless to say, the greater the number of labels in the calculation, the lower IAA rates. The main results in each round of analysis can be summed up as follows:

– The comprehensive distinction between substantive and adjacent subacts (SS vs. AS), shows a noticeable high level of agreement among annotators, even when the most fragmentary units (XXS and R) are included in the model.

– Procedural labels (AS) offer a robust IAA result that reach over 0.800 (see 4.2.2), even when including the most residual AS unit: the XAS. Including XAS ($_{(k)u}a$ = 0.000) in the calculation does not entail a general decrease on agreement, nor significantly affects the overall IAA for AS. This shows that agreement on AS categories is prevalently high.

– As for all subacts labels, the overall IAA results are high. MAS are the subacts that show higher agreement rates, also in correspondence to the general trend shown by AS, whose IAA results are higher than in the case of conceptual subacts (SS). SSS and TopSSS are the labels showing the lowest IAA rates, which suggests that these categories call for a more thorough definition in the model. They outline the difficulty of analyzing the hierarchical organisation of conceptual information in spoken language, a genre that precisely stands out for a non-prototypical distribution of information and a non-prototypical syntactical organisation, in comparison to more formal or written uses of language.

– Finally, the inclusion of the most residual units do not lead to an increase in the rate of disagreement. SXX and R are sub-structural constituents that bring to

[24] Laughing.

light the difficulties underlying the analysis of spontaneous speech. This study shows that the VAM is able to account for these residual segments by offering labels for their analysis, unlike other models of discourse segmentation (Pascual 2020).

6. Conclusions

IAA emerges as a method for testing the reliability and replicability of corpus annotation protocols. This paper tested the performance of three annotators following the VAM annotation protocol, which in turn, allows to assess the validity of this model. This is also the first time when Krippendorf's coefficients are applied to the whole process of discourse segmentation, setting thus new standards for validation within the field.

The present study has followed a two-fold procedure for segmentation: first the conversational continuum has been divided into discourse units; and second, each unit has been classified as a type of subact. The complexity of this annotation procedure contrasts with most IAA studies, where the measurement of agreement relies only in the categorization of pre-defined units whose boundaries have been set in advance (see for example Crible & Degand 2019a, Scholman et al. 2016).

Krippenforff's ua-family coefficients were applied to measuring IAA in several rounds of analysis of the same data. As outlined in section 5, the results of the experiment are very positive, since high levels of IAA were obtained in most analyses. Agreement has proven to reach positive results — yielding coefficients over 0.8 — when it comes to distinguishing conceptual and procedural content (4.2.1) and the different procedural functions conveyed by AS (4.2.2). The few shortcomings of the protocol are explained by the fact that it is hard to define constituents (SSS and TopSSS), thereby calling for a better account of such units (4.2.3).

In conclusion, Krippendorf's coefficients, applied for the first time to test a model of discourse segmentation, validate the Val.Es.Co. model as an optimal method to fully analyze a conversation into pragmatically-based discourse units.

REFERENCES

Albelda Marco, Marta & Pedro Gras Manzano. 2011. La partícula escalar ni en español coloquial. In González Ruiz, Ramón & Carmen Llamas Saíz (eds.). *Gramática y discurso. Nuevas aportaciones sobre partículas discursivas del español*, 11–31. Pamplona: Eunsa.

Albelda Marco, Marta. 2007. *La intensificación como categoría pragmática: revisión y propuesta*. Bern: Peter Lang.

Artstein, Ron & Poesio, Massimo. 2005. Bias decreases in proportion to the number of annotators. In Rogers, James (ed.), *Proceedings of FG-MoL 2005: The 10th conference*

on Formal Grammar and The 9th Meeting on Mathematics of Language Edinburgh, 139–148. Stanford: CSLI Publications [online version: http://web.stanford.edu/group/cslipublications/cslipublications/FG/2005/FGMoL05.pdf (accessed December 2020)].

Artstein, Ron & Poesio, Massimo. 2008. Inter-coder agreement for Computational Linguistics. *Computational Linguistics* 34 (4), 556–596.

Artstein, Ron. 2017. Inter-annotator agreement. In Ide, Nancy & Pustejovsky, James (eds.), *Handbook of Linguistic Annotation*, 297–313. Dordrecht: Springer.

Bello, Andrés. 1847. *Gramática de la lengua castellana destinada al uso de los americanos*. Madrid: Arco Libros.

Blanche-Benveniste, Claude. & Jeanjean, Colette. 1987. *Le français parlé*. Didier Erudition: Paris.

Briz, A. & Val.Es.Co. Group. 2003. Un sistema de unidades para el estudio del lenguaje coloquial. *Oralia* 6. 7–61.

Briz, A. 1998. *El español coloquial en la conversación. Esbozo de pragmagramática*. Barcelona: Ariel

Briz, Antonio & Pons Bordería, Salvador. 2010. Unidades, marcadores discursivos y posición. In Loureda Lamas, Óscar & Acín Villa, Esperanza (eds.), *Los Estudios Sobre Marcadores del Discurso en Español, Hoy*, 327–358. Madrid: Arco Libros.

Briz, Antonio. 2011. La subordinación sintáctica desde una teoría de unidades del discurso: el caso de las llamadas causales de la enunciación. In Bustos, J. et al. (coord.): *Sintaxis y análisis del discurso hablado en español. Homenaje a Antonio Narbona.* Sevilla: Universidad de Sevilla (I). 137–154.

Cabedo Nebot, Adrián & Salvador Pons Bordería. 2013. *Corpus Val.Es.Co. 2.0*. http://www.valesco.es/?q=corpus (accessed December 2020).

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic, *Computational Linguistics* 22 (2), 249–254.

Carlson, Lynn, Marcu, Daniel & Okurowski, Mary Ellen. 2003a. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory, In van Kuppevelt, Jan & Smith, Ronnie W. (eds.), *Current and New Directions in Discourse and Dialogue*, Springer, Dordrecht, 85–112.

Carlson, Lynn, Marcu, Daniel & Okurowski, Mary Ellen. 2003b. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Proceedings of the Second SIGdial Workshop on Discourse and Dialogue. https://www.aclweb.org/anthology/W01-1605.pdf (accessed December 2020).

CGuetzkow, Harold. 1950. Unitizing and categorizing problems in coding qualitative data. *Journal of Clinical Psychology* 6 (1). 47–58.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1). 37–46.

Cohen, Jacob. 1968. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70 (4). 213–220.

Crible, Ludivine & Degand, Liesbeth 2019a. Domains and Functions: A Two-Dimensional Account of Discourse Markers, *Discours*, 24. http://journals.openedition.org/discours/9997. (accessed December 2020).

Crible, Ludivine & Degand, Liesbeth 2019b. Reliability vs. granularity in discourse annotation: What is the trade-off? *Corpus Linguistics and Linguistic Theory* 15 (1). 71–99.

Crible, Ludivine & Pascual, Elena. 2020. Combinations of discourse markers with repairs and repetitions in English, French and Spanish. *Journal of Pragmatics* 156. 54–67. DOI: https://doi.org/10.1016/j.pragma.2019.05.002. (accessed December 2020).

Degand, Liesbeth & Simon, Anne-Catherine. 2009a. Minimal discourse units in spoken French: On the role of syntactic and prosodic units in discourse segmentation. *Discours* 4. DOI: http://discours.revues.org/5852 (accessed December 2020).

Degand, Liesbeth & Simon, Anne-Catherine. 2009b. Mapping prosody and syntax as discourse strategies: How Basic Discourse Units vary across genres. In Barth-Weingarten, Dagmar, Dehé, Nicole & Wichmann, Anne (eds.), *Where prosody meets pragmatics: research at the interface*, 79–105. Bingley: Emerald.

Degand, Liesbeth & Simon, Anne-Catherine. 2011. L'analyse en unités discursives de base: pourquoi et comment? *Langue française* 170. 45–59.

Estellés Arguedas, Maria. 2011. *Gramaticalización y paradigmas: un estudio a partir de los denominados marcadores de digresión en español*. Bern: Peter Lang.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5). 378–382.

Grisot, Cristina. 2015. *Temporal Reference: Empirical and Theoretical Perspectives. Converging Evidence from English and Romance*. Geneva: University of Geneva. PhD Dissertation.

Grisot, Cristina. 2017. A quantitative approach to conceptual, procedural and pragmatic meaning: Evidence from inter-annotator agreement. *Journal of Pragmatics* 117. 245–263.

Groupe de Fribourg (A. Berrendonner, dir.) 2012. *Grammaire de la période*, Berne: Peter Lang.

Grupo Val.Es.Co. 2014. Las unidades del discurso oral. La propuesta Val.Es.Co. de segmentación de la conversación (coloquial). *Estudios de Lingüística del Español* 35 (1). 11–71. http://infoling.org/elies/35/elies35.1-2.pdf. (accessed December 2020).

Krippendorff, Klaus, Mathet, Yann, Bouvry, Stéphane & Widlöcher, Antoine. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity: International Journal of Methodology* 50. 2347–2364.

Krippendorff, Klaus. 1970. Bivariate agreement coefficients for reliability of data. In Borgatta, Edith R. and Bohrnstedt, George W. (eds.). *Sociological Methodology*, vol. 2, Jossey-Bass Inc., San Francisco, 139–150.

Krippendorff, Klaus. 1995. On the Reliability of Unitizing Continuous Data. *Sociological Methodology* 25. 47–76.

Krippendorff, Klaus. 2013 [1980]. *Content Analysis: An Introduction to Its Methodology*. 3rd. edition. Thousand Oaks (California): ASGE Publications Inc.

Latorre, Lidia. 2017. *La unidad mínima en la conversación coloquial: delimitación y cuantificación*. Valencia: Universidad de Valencia. Master's dissertation, unpublished.

Marcu, Daniel, Amorrortu, Estíbaliz, & Romera, Magdalena. 1999. Experiments in constructing a corpus of discourse trees. In Walker, Marilyn. (ed.), *Towards Standards and Tools for Discourse Tagging* (Proceedings of the ACL'99 Workshop, College Park, Maryland). New Brunswick: Association for Computational Linguistics 48-57.

Miltsakaki, Eleni, Prasad, Rashmi, Joshi, Aravind & Webber, Bonnie. 2004. Annotating discourse connectives and their arguments. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL Boston, Massachusetts*. 9–16. https://www.aclweb.org/anthology/W04-2703/ (accessed December 2020).

Mírovský, Jiri, Mladová, Lucie & Zikánová, Sárka. 2010. Connective-based measuring of the inter-annotator agreement in the annotation of discourse in PDT. In Huang, Chu Ren & Jurafsky, Dan. (eds.), *Proceedings of the 23rd International Conference on Computational Linguistics: Posters Volume (COLING '10)*. Beijin: *Chinese* Information Processing Society of China and Association for Computational Linguistics. 775–781. https://dl.acm.org/doi/10.5555/1944566.1944655 (accessed December 2020).

Morel, Mary-Annick & Danon-Boileau, Laurent. 1998. *Grammaire de l'intonation. L'exemple du français*. Paris: Ophrys.

Narbona, Antonio. 1986. Problemas de sintaxis coloquial andaluza. *Revista Española de Lingüística* 16 (2). 229–276.

Narbona, Antonio. 1992. Hacia una sintaxis del español coloquial. In *Congreso de la Lengua Española (1992, Sevilla)*, *Instituto Cervantes*, 721–740. https://idus.us.es/xmlui/handle/11441/29504. (accessed December 2020).

Narbona, Antonio. 2012. Los estudios sobre el español coloquial y la lingüística. *Revista Española de Lingüística* 42 (2). 5–32.

Pascual Aliaga, Elena. 2018. Análisis prosódico de las estructuras truncadas en la conversación coloquial española: funciones de formulación y atenuación. In García Ramón, Amparo & Soler Bonafont, María Amparo (eds.). *ELUA: Estudios de antenuación en el discurso, Anexo IV*, 57–84.

Pascual, Elena. 2015a. Aproximaciones a la caracterización prosódica de los subactos, la unidad discursiva mínima del sistema Val.Es.Co. In Cabedo, A. (ed.), *Perspectivas actuales en el análisis fónico del habla. Tradición y avances en la fonética experimental.* Annex 7 of *Normas. Revista de Estudios Lingüísticos Hispánicos*. 137–150.

Pascual, Elena. 2015b. Aproximación a la segmentación del subacto en la conversación coloquial española. In Henter, Sara, Izquierdo, Silvia and Muñoz, Rebeca (eds.), *Estudios de pragmática y traducción*. Murcia: EDITUM. 73–102.

Pascual, Elena. 2020. *Los truncamientos en la conversación coloquial. Estudio de las huellas de formulación discursiva desde un modelo de unidades de lo oral*. Valencia: Universidad de Valencia. PhD Dissertation.

Pons Bordería, Salvador & Maria Estellés Arguedas. 2009. Expressing digression linguistically: Do digressive markers exist? *Journal of Pragmatics* 41 (5). 921–993.

Pons Bordería, Salvador (ed.). 2014. *Discourse Segmentation in Romance Languages*. Amsterdam/Philadelphia: John Benjamins.

Pons Bordería, Salvador. 2008. Gramaticalización por tradiciones discursivas: El caso de 'esto es'. In Kabatek, Johannes (ed.). *Sintaxis histórica del español y cambio lingüístico: Nuevas perspectivas desde las Tradiciones Discursivas*, 249–274. Madrid: Iberoamericana.

Pons Bordería, Salvador. 2016. Cómo dividir una conversación en actos y subactos. In Bañón Hernández, Antonio Miguel, Espejo Muriel, María del Mar, Herrero Muñoz-Cobo, Bárbara & López Cruces, Luis. *Oralidad y análisis del discurso: homenaje a Luis Cortés Rodríguez*, 545–566. Almería: Universidad de Almería.

Prasad, Rashni, Dinesh, Nikil, Lee, Alan, Miltsakaki, Elena, Robaldo, Livio, Joshi, Aravind & Webber, Bonnie. 2008. *The Penn Discourse Treebank 2.0*. In Calzolari, Nicoletta, Choukri, Khalid, Maegaard, Bente, Mariani, Joseph, Odijk, Jan & Tapias, Daniel. (eds.), P*roceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco*. 2961–2968. http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf. (accessed December 2020).

Prasad, Rashni, Webber, Bonnie, Lee, Alan & Joshi, Aravind. 2019. *The Penn Discourse Treebank 3.0*. LDC2019T05. Philadelphia: Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2019T05#. (accessed December 2020).

Riou, M. 2015. A methodology for the identification of topic transitions in interaction. *Discours*, 16. http://journals.openedition.org/discours/8997. (accessed December 2020).

Roulet, Eddy et al. 1985. *L'articulation du discours en français contemporain*, Berne: Peter Lang.

Roulet, Eddy, Fillietaz, Laurent and Grobet, Anne. 2001. *Un modèle et un instrument d'analyse de l'organisation du discours*. Berne: Peter Lang.

Roulet, Eddy. 1991. Vers une approche modulaire de l'analyse du discours. *Cahiers de Linguistique Française* 12. 53–81.

Rysová, Magdaléna, Pavlína Synková, Jiří Mírovský, Eva Hajičová, Anna Nedoluzhko, Radek Ocelák, Jiří Pergler, Lucie Poláková, Veronika Scheller, Jana Zdeňková & Šárka

Zikánová. 2016. *Prague Discourse Treebank 2.0.* Data/software, ÚFAL MFF UK, Prague, Czech Republic. (http://hdl.handle.net/11234/1-1905, accessed December 2020).
Sacks, Harvey, Schegloff, Emanuel A. & Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* 50 (4). 696–635.
Salameh Jiménez, Shima, Estellés Arguedes, Maria & Pons Bordería, Salvador. 2018. Beyond the notion of periphery: An account of polyfunctional discourse markers within the Val.Es.Co. model of discourse segmentation. In Beeching, Kate, Ghezzi, Chiara & Molinelli, Piera (eds.). *Positioning the Self and Others. Linguistic perspectives*. Amsterdam/Philadelphia: John Benjamins, 105–125.
Salameh Jiménez, Shima. 2021. *Reframing Reformulation: A Theoretical-Experimental Approach Evidence from the Spanish Discourse Marker "o sea"*. Bern: Peter Lang.
Sanders, Ted, Spooren, Wilbert & Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15. 1–35.
Sanders, Ted, Spooren, Wilbert & Leo Noordman. 1993. Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics* 4 (2). 93–133.
Scholman, Merel, Jacqueline Evers-Vermeul & Ted Sanders. 2016. A step-wise approach to discourse annotation: towards a reliable categorization of coherence relations. *Dialogue & Discourse* 7 (2). 1–28.
Scott, William A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly* 19 (3). 321–325.
Sinclair, John McHardy & Malcom Coulthard. 1975. *Toward an Analysis of Discourse: The English used by Teachers and Pupils.* Oxford: Oxford University Press.
Sornicola, Rosana. 1981. *Sul parlato*. Bologna: Il mulino.
Spooren, W. & Degand, L. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory* 6 (2). 241–266.
Stati, Sorin. 1990. *Le transphrastique*. Paris: Presses Universitaires de France.
Van Dijk, Teun A. 1977. *Text and context: Explorations in the semantics and pragmatics of discourse*. London: Logman.
van Enschot, Renske, Spooren, Wilbert, van den Bosch, Antal, Burgers, Christian, Degand, Liesbeth, Evers-Vermeul, Jacqueline, Kunneman, Florian, Liebrecht, Christine, Linders, Yvette & Maes, Alfons. In press. Taming our wild data: On intercoder reliability in discourse research. *Dialogue & Discourse*.
Wilson, Deirdre. 2011. The conceptual–procedural distinction: Past, present and future. In: Escandell-Vidal, Victoria, Leonetti, Manuel & Ahern, Aoife (eds.). *Procedural Meaning: Problems and Perspectives*, 1–31. Leiden, The Netherlands: Brill.
Zufferey, Sandrine & Andrea Popescu-Belis. 2004. Towards Automatic Identification of discourse markers in dialogs: The case of *like*. In Strube, Michael & Candy Sidner (eds.). *5th SIGdial Workshop on Discourse and Dialogue.* Proceedings of the Workshop, Cambridge, Massachusetts. East Stroudsbur: Association for Computational Linguistics. https://www.aclweb.org/anthology/W04-2313.pdf (accessed December 2020).

Article history:
Received: 10 November 2020
Accepted: 12 February 2021

История статьи:
Дата поступления в редакцию: 10 ноября 2020
Дата принятия к печати: 12 февраля 2021

Bionotes:
Salvador PONS BORDERÍA is Professor of Spanish Linguistics at the University of Valencia, Spain. He is a member of the Val.Es.Co. Research Group and his research interests include spoken language, approximatives, and the synchronic and diachronic description of discourse markers.
***Contact information**:*
Universidad de Valencia
Valencia, Spain
e-mail: salvador.pons@uv.es
ORCID: 0000-0001-5788-5506

Elena PASCUAL ALIAGA holds a Phd in Spanish Linguistics and is a member of the Val.Es.Co. Research Group. Her research interests include sub-structural elements and disfluencies in spoken conversations.
Contact information:
Universidad de Valencia
Valencia, Spain
e-mail: elena.pascual@uv.es
ORCID: 0000-0002-1912-4957

Сведения об авторах:
Сальвадор ПОНС БОРДЕРИА – профессор испанской лингвистики Валенсийского университета (Испания), член исследовательской группы Val.Es.Co. Его научные интересы включают разговорную речь, апроксимативы, а также синхроническое и диахроническое описание дискурсивных маркеров.
***Контактная информация**:*
Universidad de Valencia
Valencia, Spain
e-mail: salvador.pons@uv.es
ORCID: 0000-0001-5788-5506

Елена ПАСКУАЛЬ АЛИАГА – доктор испанской лингвистики, член исследовательской группы Val.Es.Co. В сферу ее научных интересов входят субструктурные элементы, а также факторы, мешающие плавности устной речи.
Контактная информация:
Universidad de Valencia
Valencia, Spain
e-mail: elena.pascual@uv.es
ORCID: 0000-0002-1912-4957