

Event and entity coreference across five languages: Effects of context and referring expression

Luca Bevacqua*

*School of Philosophy, Psychology & Language Sciences
University of Edinburgh*

LUCA.BEV@HOTMAIL.IT

Sharid Loáiciga*

*Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg*

SHARID.LOAICIGA@GU.SE

Hannah Rohde

*School of Philosophy, Psychology & Language Sciences
University of Edinburgh*

HANNAH.ROHDE@ED.AC.UK

Christian Hardmeier

*Department of Computer Science, IT University of Copenhagen
Department of Linguistics and Philology, Uppsala University*

CHRHA@ITU.DK

Editor: Massimo Poesio

Submitted 08/2021; Accepted 11/2021; Published online 12/2021

Abstract

Current work on coreference focuses primarily on entities, often leaving unanalysed the use of anaphors to corefer with antecedents such as events and textual segments. Moreover, the anaphoric forms that speakers use for entity and event coreference are not mutually exclusive. This ambiguity has been the subject of work in English, with evidence of a split between comprehenders' preferential interpretation of personal versus demonstrative pronouns. In addition, comprehenders are shown to be sensitive to antecedent complexity and aspectual status, two verb-driven cues that signal how an event is being portrayed. Here we extend this work via a comparison across five languages (English, French, German, Italian, and Spanish). With a story-continuation experiment, we test how different referring expressions corefer with entity and event antecedents and whether verbal features such as argument structure and aspect influence this choice. Our results show widely consistent, not categorical biases across languages: entity coreference is favoured for personal pronouns and event coreference for demonstratives. Antecedent complexity increases the rate at which anaphors are taken to corefer with an event antecedent, as does portraying an event as completed though the latter does not reach significance. Lastly, we report a comparison of the same referring expressions to refer to entity and event antecedents in a trilingual parallel corpus annotated with coreference. Together, the results provide a first crosslingual picture of coreference preferences beyond the restricted entity-only patterns targeted by most existing work on coreference. The five languages are all shown to allow gradable use of pronouns for entity and event coreference, with biases that align with existing generalizations about the link between prominence and the use of reduced referring expressions. The studies also show the feasibility of manipulating targeted verb-driven cues across multiple languages to support crosslingual comparisons.

Keywords: Event anaphora, coreference resolution, pronouns, multilingual.

*. Shared first authorship.

1. Introduction

Work in coreference has focussed primarily on entity coreference, typically between third person personal pronouns and human antecedents. The use of anaphors to refer to non-human entities has been less studied, especially when the reference is to antecedents that are not entities, such as events or full clauses. Non-entity coreference relations, including not only events but also previous portions of a text in discourse-deixis or metalinguistic description, present both formal and practical challenges. If the antecedent can be any of the available previously mentioned entities, events, propositions or even entire rhetorical arguments, the number of candidate antecedents becomes potentially very large. In addition, the anaphoric forms that speakers use for different kinds of coreference are not mutually exclusive: sentences (1) and (2) show the demonstrative *this* referring back to respectively an entity (the wine) and an event (the successful aging of the wine).¹

1. The Northern Italian wine aged well. This was one of the best wines I'd ever tasted.
2. The Northern Italian wine aged well. This made the wine more marketable.

This ambiguity has been the subject of work on the coreference system in English, with evidence of a split between comprehenders' preferential interpretation of personal versus demonstrative pronouns (Çokal et al., 2018; Loáiciga et al., 2018). Specifically, personal pronouns are preferred when referring to concrete entities, and demonstratives when referring to other more complex or abstract antecedents (see §2.1). Demonstratives have in fact been identified as the expressions speakers use to signal a prominence shift, particularly to change the discourse topic (Givón, 1983) and to reject the most prominent antecedent as a possible antecedent (Comrie, 1997).

Different languages use different referring expressions, ordering the elements in a given language's pronominal system along a prominence scale: for example, while in English personal pronouns encode antecedents that are highly accessible, topical and activated in short-term memory, in languages with widespread use of zero anaphors such as Italian or Spanish, personal pronouns will preferentially be used to refer to less prominent referents. Moreover, less prominent antecedents will need more semantically rich types of referring expressions in order to be retrieved (such as overt pronouns, which encode number and gender, or even nouns/names), whereas more prominent antecedents can be referenced with a wider variety of referring expressions (von Heusinger and Schumacher, 2019). It is often the case that these richer referring expressions will also be phonologically richer. For our purposes, we talk about heavy (and light) referring expressions to indicate forms that encode both more semantic and phonological content.

In addition, when processing coreference, comprehenders are shown to be sensitive to a wide range of features related to structural and meaning-driven properties of the passage (see reviews such as Greene et al., 1992; Carlson, 2003; Rohde, 2019). Those findings include a large body of work in psycholinguistics, much of which has emphasized English (with notable exceptions of course, e.g., Kaiser and Trueswell, 2004; Mayol, 2018). Here we pick up on two English findings from a psycholinguistic study: antecedent complexity (number of arguments) and aspect, two verb-driven cues that signal how an event is being portrayed.

While the definition of "accessibility" (that we use as defined in Accessibility Theory: Ariel 1988, 2004) or "prominence" is much discussed (see von Heusinger and Schumacher, 2019, for a

1. All examples are taken from the experiments described in §3.

definition of application of the concept), some predictions are recurrent across theories. For example, the accessibility of a target antecedent is assumed to decrease with a higher number of possible competitors (see e.g. Centering Theory: Grosz et al. 1995). Beyond entity coreference, one can consider the potential antecedents and competitors made available by the description of an event. An event antecedent is more complex than an entity antecedent in that it not only includes more than one entity, but also the relationships between the entities.

One way to operationalise antecedent count is to consider the category of a verb within the causative-inchoative alternation (Haspelmath, 1993), which can then be used as a proxy for antecedent choice complexity. Alternating verbs admit both intransitive and transitive uses (e.g., *The snow melted / The sun melted the snow*), corresponding to inchoative and causative interpretations, respectively. This alternation is not present in all verbs (e.g., *The battery died / *Frequent use died the battery*). Jespersen (1927) classes verbs that can undergo this alternation as “move and change verbs”, as they often describe changes in state or movement: the transitive use, then, makes available the implicit entity that brought on the change. Alternating verbs, by offering this additional implicit entity to the pool of coreference possibilities even when the verb appears in the intransitive, have been shown to increase event coreference in English (Loáiciga et al., 2018); for non-alternating verbs, the sole entity available in the sentence seemingly becomes relatively more salient by not having competition from other implicit entities. However, this is a gradable phenomenon, in that some verbs position in between categories: even verbs which are most firmly in the non-alternating category can have rare or idiomatic transitive uses (e.g. internal objects, *to die a cruel death*). Languages may differ in how clean-cut the divide between alternating and non-alternating verbs is, and in how sensitive coreference is to this distinction. In the study reported here, we compare the availability of event and entity coreference across alternating vs non-alternating verb classes.

Another factor known to influence discourse representations and coreference processing is verb aspect. Verb aspect is used as a cue to the portrayal of an event. Our study also tests whether an aspectual manipulation may influence a comprehender’s perception of the concreteness of the event, and thereby its availability as an antecedent. Events which are portrayed as completed might have a salient end-state and, thus, be more encapsulated and retrievable (Moens and Steedman, 1988) or, on the contrary, the ongoing state of an event may activate it and, with it, its properties and world knowledge (Ferretti et al., 2001, 2007, 2009).

The relevance of the studies we present here is underscored not only by the lack of work on abstract anaphora and the possibility of making crosslingual extensions to the limited prior work on this topic, but also by the similar questions raised in coreference-annotated corpora that inform computational systems for coreference resolution. Existing resources in the field of coreference resolution also reflect the difficulty in the treatment of pronouns. For example, given the complexity involved in identifying non-nominal anaphora, and its relatively low frequency, most annotation efforts and hence most coreference systems focus on anaphora to nouns. The OntoNotes corpus (Pradhan and Xue, 2009), for instance, although the largest in the community, does not include annotations of event entities coreferential with pronouns. Corpora including events exist, but they are smaller in size and less frequently used. The ParCorFull corpus (Lapshinova-Koltunski et al., 2018), for instance, includes the annotation of event anaphora in its guidelines, but event references only constitute 12.5% of the total annotated markables in English. ParCorFull is a parallel corpus that also includes annotations in German and French. In German, event annotations amount to 11%; in French the proportion is even lower, however it should be noted that the lower number of documents for this part of the corpus makes the estimate less reliable.

Improving our understanding of the interpretation of event anaphora and non-nominal anaphora in general is a step towards better anaphora and coreference resolution systems (Poesio et al., 2015). Current systems struggle to identify and resolve this type of reference simply because they have not been trained to do so (Heinzerling et al., 2017), as annotated corpora focus on nominal anaphora. Annotated corpora are crucial for system development and large scale linguistic analysis, but equally important is a sound theoretical understanding of the phenomenon one wishes to annotate. Finally, a more detailed and most importantly multilingual description of these coreference patterns can be useful for studies on multilingualism, which could refer to figures in a native speakers population as a point of comparison, and for language teaching.

This paper presents work on these different types of coreference via a comparison across five languages (English, French, German, Italian, and Spanish), which are broadly related typologically but which differ in their use of grammatical gender and case and their pronominal systems, most notably the availability of a null pronoun. We ask when and to what degree event instances serve as antecedents when a competing entity referent is also available. The goal is to model human choices in five different language settings to investigate the similarities and differences that arise as a product of the structural differences in the languages, and to shed light on the relationship between event and nominal anaphora in order to inform future coreference annotation efforts and coreference systems.

We report five story continuation studies targeting participants' resolution of pronominal forms *It vs This* in English, *Il vs Cela vs C'est* in French, *Es vs Das vs Dies* in German, *Ciò vs Questo vs zero pronoun* in Italian, and *Esto vs Este vs zero pronoun* in Spanish.

Regarding antecedent complexity and aspectual status, we test the impact of (i) the availability of an additional implicit argument to the verb, namely the complexity created via the causative-inchoative alternation and (ii) the aspectual difference of portraying of an event as completed or ongoing. Regarding anaphoric forms, we test (iii) whether event coreference is inferred more for demonstratives over personal pronouns and (iv) whether this distinction is categorical, with a clear division of labour, or gradable.

Our results show consistent and gradable biases: participants encountering a personal pronoun are more likely to continue the sentence writing about the entity, while demonstratives prompt more event continuations across all five languages. The complexity of the antecedent, manipulated through the proxy of alternation, increases event coreference, though this is significant in only four of the five languages studied. The portrayal of an event as completed, on the other hand, seems to have a much smaller role in influencing coreference. We speculate how the varying aspectual systems of the five languages may undermine this effect. Lastly, we report corpus frequencies comparing the use of the same referring expressions to refer to entity and event antecedents, along with their coreference chains. This comparison confirmed the preference for events to be referred to mostly with demonstrative pronouns and entities with personal pronouns, but with a more gradable pattern than in the experiments with human participants. In contrast, the corpus study failed to find a similar pattern for the proxy of verb alternation due to a lack of annotated resources with alternation status or thematic roles.

Together, the results provide the first crosslingual picture of coreference preferences beyond the restricted entity-only patterns that the target of most existing work on coreference. The five languages are all shown to allow gradable use of personal and demonstrative pronouns for entity and event coreference, with systematic biases that align with existing generalizations about the use of lighter referential forms for less complex and more concrete referents.

2. Related Work

Referring expressions play a central role in human communication and are at the heart of multiple linguistics theories and applications involving natural language understanding. For both, a lot of attention has been paid to the most standard cases represented by nominal expressions and personal pronouns while more atypical cases have been understudied. Among these is the phenomenon of abstract anaphora – anaphora that involve reference to abstract entities such as events or states (Asher, 1993) whose complexity is increased because the form and interpretation of the referring expressions involved have a high degree of variability. This complexity is also behind the lack of resources annotated with this type of anaphora, simultaneously hindering its study.

2.1 Linguistic descriptions

Linguistic studies on the subject have employed both corpora and psycholinguistic methods. Corpus-based studies are a source of insights about language use, since the written texts they are based on are natural passages after all. They offer better estimates for building coreference resolution systems that will be used on those texts. On the other hand, corpus-based studies do not provide any explanation as to why a particular item follows a certain distribution, and they grant little control over the confounding variables responsible for that distribution. In this respect, psycholinguistic studies are more suitable for capturing the cognitive processes behind naturally occurring phenomena. Psycholinguistic research has focused on using theoretical constructs of complexity, salience, and information status to capture coreference patterns.

In English, the anaphoric use of pronouns to refer to entities or abstract events has been evolving in the last centuries (Azuma, 2008): while *it* has consistently been the preferred form to corefer to pronominal, highly accessible antecedents, the distal demonstrative *that* has grown in use to refer to NP or clausal antecedents (even overtaking the personal pronoun for non-nominal antecedents), positioning itself between *it* and the proximal demonstrative *this*, the most rarely used and marked of the three expressions. In an estimate of the current relative frequency of nominal coreference, Gundel et al. (2005) report that while about 84% of personal pronouns had nominal antecedents, only 28% of demonstratives did.

In other studies, the demonstratives *this* and *that* have been grouped together, assuming that they pattern together and behave differently from *it* (as confirmed in the context of textual deixis: see Çokal et al. 2014). Hedberg et al. (2007), for example, show how the two demonstratives tend to have non-nominal antecedents, unlike *it*: the personal pronoun requires the antecedent to be more highly activated, to be in the focus of attention, while demonstratives can retrieve antecedents that are less activated because of their complexity or because they were not directly introduced, such as events. This is confirmed in the study by Çokal et al. (2018), where *it* was found to corefer to concrete entities both in production and interpretation, while *this* was biased towards non-NP, less salient antecedents. Likewise, in our previous study (Loáiciga et al., 2018) we find a preference for *it* to refer to entities and of *this* to refer to events. Similar results were also found by Wittenberg et al. (2021) looking at *it* versus *that*, with the demonstrative being preferred for event coreference: this is interpreted as a “conceptual bundling” action of *that*, which would make event antecedents accessible by wrapping their complex structure in a simpler referring expression.

In a comparison of demonstratives across English, German, Italian and Spanish (Dipper et al., 2011), English and German were shown to prefer the use of demonstratives for retrieving abstract anaphors, while Italian and Spanish were not shown to display a preference. The German findings

were further confirmed in a sentence rating study by Bentzen and Anderssen (2019), in which speakers of German were shown to prefer *das* to refer to non-nominal antecedents, with *es* being a less common but acceptable option when the antecedent is used as a “continuing topic”. A corpus study comparing English, Danish and Italian (Navarretta, 2007) confirms this pattern, finding that in Italian the difference between personal and demonstrative pronouns with regard to their type of coreference is not significant, unlike in English, where a contrast in use individuates different kinds of anaphors. In the Italian data, however, null pronouns were used for abstract referents in only 30% to 42% of instances (respectively in the original-Italian and parallel translations corpora). This is in keeping with antecedent accessibility hierarchies that position zero anaphora as requiring the highest level of accessibility (cf. Givón, 1983; Ariel, 1988, 2004). Assuming entities to be more stereotypical antecedents than events, as shown by crude annotation distributions (see §1) the Italian and Spanish null pronouns should be employed less for abstract anaphora.

More generally, the possibility of referencing an event with *it* or *this* (or *that*) is captured in models of discourse deixis. In line with Bentzen and Anderssen’s study, Webber (1991) posits that propositions, particularly those that are focussed, are referred to first with either *this* or *that* in immediately following sentences. In later instances of coreference with those same events, the pronoun *it* can then be used. The difference between the two demonstratives is tested by Çokal et al. (2014), finding that both *this* and *that* refer preferably to the most recent clause. Building on Centering Theory (Grosz et al., 1995), Passonneau (1989) analyzes intra-sentential instances of *it* vs *that* with an explicit NP antecedent, finding that *it* is used to refer to the center (most often the subject), whereas *that* favors the less prominent non-centers.

For French, Cornish (2015) shows how the demonstratives *cela* and *ça*, along with the neuter clitics *le*, *y* and *en*, can retrieve a sentential or verb-phrasal antecedent, and this relation is not simply formal-syntactic. A further claim is made comparing French with English, and saying that the anaphoric distinction between nominal and non-nominal antecedents has a correspondence to the forms: personal pronouns such as *she/he* and *elle/il* can only be used for nominal antecedents, while other expressions such as *do it/this/that* in English and *faire cela/ça* in French form a contrasting set used exclusively for non-nominal antecedents (such as predicates). Cornish goes on to observe that in French the nominative neuter pronoun *il* is rarely used as a propositional anaphor (which will use *ce/cela/ça*), because it is often used as an expletive.

In French corpus studies on the matter, Tutin (2002) shows the presence of demonstratives used as anaphors, although much less frequently than personal pronouns and clitics, but the type of anaphora is not considered. Vieira et al. (2005), comparing French and Portuguese, classify antecedents in the two categories of “concrete” and “abstract”, where the first includes existing entities (such as people, things, places) and the second includes notions, actions and states. Their results show that, in both languages, demonstratives are used for abstract antecedents in more than three quarters of cases. However, their study targets demonstrative noun phrases, that is demonstrative adjectives rather than pronouns, so the results may not be comparable.

Here, we target these same previously studied languages, but we adopt a story continuation methodology to probe the availability of entity and event antecedents for demonstrative, personal and null pronouns. Where psycholinguistic approaches have been used previously to analyze *it* vs *that*, the emphasis has been on English. For example, Brown-Schmidt et al. (2005) report that comprehenders show a preference to interpret *that* to refer to a complex, composite antecedent (e.g., *I’ll have the hamburger and fries. I’ll have that, too.*), independent of other metrics of the salience

of the referent. These results leave an open question regarding the crosslinguistic generalisability of such claims.

2.2 Computational approaches

Reference to non-nominal antecedents has largely been a niche area in computational linguistics research (see review by Kolhatkar et al., 2018). The most extensive annotation efforts in the field of coreference resolution have focused on English nominal coreference. OntoNotes (Pradhan et al., 2013), the largest and most frequently used corpus for training coreference resolution systems, for instance, only includes verbs if “they can be coreferenced with an existing noun phrase” according to its guidelines. Corpora with a richer annotation of event pronouns exist, but are much smaller. One such resource is the ARRAU corpus (Uryupina et al., 2020), whose size amounts to about 20% of version 5 of OntoNotes. ParCorFull (Lapshinova-Koltunski et al., 2018) also contains annotations of event pronouns in English, German and French.

The scarcity of manually annotated resources has led to the use of artificial training data for coreference resolution systems of English non-nominal anaphora. Kolhatkar et al. (2013) study the resolution of anaphoric shell nouns such as ‘this issue’ or ‘this fact’ by exploiting instances such as ‘the fact that. . .’. Marasovic et al. (2017) construct training examples based on specific patterns of verbs governing embedded sentences. Also on this front, Loáiciga et al. (2020) use parallel machine translation data in many languages for automatic data creation of event pronouns.

Before the breakthrough of neural end-to-end systems in coreference resolution (Lee et al., 2017), coreference resolvers needed to do explicit mention classification in order to exclude non-referential mentions before any resolution was attempted. In this context, the pronoun *it* has been targeted, as many of its uses are non-referential (e.g. *it rains*, *it’s Tuesday*, *it seems like*). Evans (2001) proposes the classification of the pronoun *it* into seven classes using contextual features. Boyd et al. (2005) report similar results of around 80% accuracy on the classification of instances of *it* using more complex syntactic patterns. Bergsma and Yarowsky (2011) describe a system for identifying non-referential pronouns using statistics of word sequences (*n*-grams) from the world wide web, however without accounting explicitly for event reference.

The many uses of *it* are also particularly relevant in dialogue, where event reference is much more common than in text data. However, there has been little recent research dedicated to coreference for general purpose dialogue systems, even though current text coreference resolution systems are not trained to manage dialogue data. In this context, Müller (2007) proposes a disambiguation of *it* together with the deictic pronouns *this* and *that*. Both Eckert and Strube (2000) and Byron (2002) also report on systems able to resolve both nominal and non-nominal anaphora. Working in a domain close to dialogue, Lee et al. (2016) create a corpus for *it*-disambiguation in question answering.

More recently, Loáiciga et al. (2017) proposed a semi-supervised approach based on a combination of syntactic and semantic features for the classification of *it*. Yaneva et al. (2018), on the other hand, report on experiments using features from eye gaze that prove to be more effective than any of the other types of features reported in previous works. This prior work points to the interest in understanding the behavior of pronouns like *it*, whose range of referential uses (entity, event, etc.) and non-referential uses presents a persistent challenge for computational systems. The approach we take here is to bring a technique from psycholinguistics to better model entity/event coreference.

3. Experiments

The work we present here tests entity/event coreference across five languages. We use a story continuation paradigm in which we present a context sentence and then assess how a subsequent referential form is interpreted. We collect story continuations in English, French, German, Italian and Spanish, measuring the rate of entity-versus-event coreference. We manipulate the heaviness of referential forms (null or personal pronoun versus demonstrative pronoun) along with two previously studied properties of the context sentence: aspect (perfective versus imperfective) and antecedent complexity (sentences with versus without an additional implicit argument, as determined by the alternating/non-alternating status of the verb). The goal is to use the same methodological probe and systematic context manipulations to test coreference biases across an inventory of coreferential forms in English, French, German, Italian and Spanish.

3.1 Choice of target pronouns

Our chosen target pronouns can be seen in Table 1. While the English pronouns are the same as those used in our previous study (Loáiciga et al., 2018), where we targeted the personal pronoun *it* and the demonstrative *this*, other factors guided the choice in the other four languages.

Among our set of 5 languages, only Italian and Spanish allow a null subject in their inventory of possible referential forms. Given the existence of theories pointing to a division of labour between null and overt pronouns in the two languages (see e.g. Carminati, 2002, for Italian, applied to Spanish in e.g. Alonso-Ovalle et al., 2002), and given hierarchies positioning null-anaphors at the highest point on accessibility/prototypicality scales (see Givón, 1983, and Ariel, 1988), our study allows us to check whether a division of labour also applies to event vs entity coreference: for example, if entity antecedents are more available/accessible, a null pronoun may corefer more easily to an entity, while an event could be picked up by an overt form.

Other than the null, in Spanish we targeted the masculine proximal demonstrative adjective *este* and the neuter demonstrative pronoun *esto*. While *esto* is a singular-only pronoun, *este* can also be used as a determiner and its status is sometimes controversial (RAE, 2010, §17.2.2b). Moreover, the Real Academia Española states that *esto* (and the other neuter demonstratives *eso* and *aquello*) are used to refer to inanimate entities or propositional content (while their use for animals is uncommon and for people, offensive) (§17.2.5b); however, the function of referring to propositional content is not mentioned for *este* (nor for the analogues *ese* and *aquel*).

In Italian we targeted the masculine proximal demonstratives *questo* and *ciò*: the two demonstratives were chosen following up from Navarretta's (2007) corpus study, in which *questo* had a slight preference for entities and *ciò* had a preference for events. Specifically, Serianni (1997, p. 198) describes *ciò* as a pronoun of neuter value (= this thing, that thing), and states it is of very common use, especially in written language, while spoken language would more commonly use *questo* (this) or *quello* (that).

For French, we chose the two pronouns *il* and *cela* and the expression *c'est*, composed of the particle *ce* and the copula in the third person singular of the indicative present tense. Our use of *c'est* reflected our concern that the bare form *ce* would be interpreted as a determiner (*ce chien*), limiting our ability to observe entity-vs-event coreference with the demonstrative NP *ce* itself. Using *c'est* prompts also allowed us to avoid the problem that the bare form *ce* would not have been compatible with a continuation with the verb *est* because of the obligatory vowel elision, while using the letter-punctuation combination *c'* in isolation posed a risk of confusion for the participants. *Il* is a third-

English	it, this
French	il, cela, c'est
German	es, das, dies
Italian	<i>null</i> , <i>ciò</i> , <i>questo</i>
Spanish	<i>null</i> , <i>esto</i> , <i>este</i>

Table 1: Pronominal prompts in the five languages.

person personal pronoun, it is the dummy subject in impersonal forms² and is used anaphorically to retrieve an infinitive or subjunctive clause. The demonstrative *cela* is a neutral form which is used for non-gendered antecedents like propositions. The Grevisse grammar (Grevisse and Goosse, 2008, §691ff.) includes in this neuter category, along with the distal pronoun *cela* and its proximal counterpart *ceci*, also *ça* and *ce*. *Cela* is used more in written language and can only refer to people in an informal register, while being normally used to refer to something “that we cannot name with precision” (§698.c). *Cela* and *ça* have replaced *ce* in most of its uses, but *ce* is still used for inanimate objects or to refer back to a sentence, and followed by a copula it refers to “what comes before or the situation” (§702).

For German, the Duden grammar (Dudenredaktion, 2009) says that *dies* is a shortened form of *dieses* which is predominantly used as a pronoun rather than a determiner, while *diese* can be used both ways (§372). In this account, phoric and deictic reference are distinguished such that the first links to referents without pointing explicitly while the second explicitly points to an object of discourse (§1818): textual content can be referred to phorically with *es* and deictically with *das* (§1821). Finally, while anaphoric personal pronouns can refer to nouns in distant sentences, the grammar explains that a demonstrative like *dieser* links partly anaphorically and partly deictically to the closest nominal candidate for reference (§1827), without addressing the potential of *dies* and *dieses* to refer to non-nominal antecedents.

3.2 Design and materials

The structure of the experiments in the five languages was the same. In each language, the 24 experimental items consisted of a sentence describing a situation followed, after a full stop and a line break, by a pronominal prompt and a space to type in a continuation. The pronominal prompts were manipulated within items. For the null pronoun condition in Italian and Spanish, the continuation box was simply not introduced by a pronoun. The pronominal prompts used in the different languages are reported in Table 1, and examples of items in different conditions of alternation and aspect for each language are given in (1)–(5):

- (1) **English:**
 - a. **NoAlt, Imperf:** The colonial building was collapsing slowly. It/This ...
 - b. **Alt, Perf:** The cake for the guests cooked poorly. It/This ...
- (2) **French:**
 - a. **NoAlt, Perf:** Le bâtiment colonial a coulé sous la neige. Il/Cela/C'est ...
The building colonial has collapsed under the snow.

2. Interestingly, this function could historically be taken by *il* and *cela* alike, a use surviving in some expressions (e.g. *il/cela est vrai*, “it/that’s true”).

- b. **Alt, Imperf:** Le pain que j'avais mis au four cuisait encore. Il/Cela/C'est
The bread that I-had put in-the oven was-cooking still.

...

(3) **German:**

- a. **NoAlt, Perf:** Das prachtvolle Gebäude zerfiel über die Jahre. Es/Das/Dies ...
The magnificent building decayed over the years.
- b. **Alt, Imperf:** Das Geräusch, das ich hörte, erstarb unversehens. Es/Das/Dies ...
The noise that I heard died suddenly.

(4) **Italian:**

- a. **NoAlt, Perf:** Il palazzo coloniale è collassato improvvisamente. Questo/Ciò/null ...
The palace colonial is collapsed suddenly.
- b. **Alt, Imperf:** Il vino piemontese invecchiava bene. Questo/Ciò/null ...
The wine Piedmontese was-aging well.

(5) **Spanish:**

- a. **NoAlt, Perf:** El problema burocrático aparecía de nuevo. Esto/Este/null ...
The problem bureaucratic appeared once more.
- b. **Alt, Imperf:** El dolor que me molestaba se pasó de repente.
The pain that me was-bothering REFL disappeared suddenly.
Esto/Este/null ...

Participants of every group saw all 24 experimental items, presented with an even distribution of the language's pronoun prompts, interleaved with 42 fillers. Of these, 18 were items of an unrelated experiment involving named entities, 20 were real fillers including a context sentence concerning two referents followed by an adverbial prompt in a new sentence (e.g., *However, Because of this*), and four were control items with unambiguous obvious responses (e.g., *Wilma played a Led Zep-pelin guitar solo in front of the crowd. It was Stairway to _____*). The stimuli were as similar as possible across languages, deviating from a literal translation when it would not have sounded natural.

The stimuli also included between-item manipulations of the status of the event as completed or ongoing. Half of the stimuli appeared in the perfective and half in the imperfective. The aspectual forms used were the present perfect and past continuous for English and their corresponding forms in the other languages: the passé composé and the imperfect in French, the passato prossimo and imperfect in Italian, and the preterite and imperfect in Spanish. In German, where aspect is not encoded in tenses, the präteritum was used; the aspect could in this case be inferred either through adverbials or contextually.

Moreover, half of the stimuli had an alternating verb and half had a non-alternating verb (in a pattern that did not correspond to the perfective/imperfective split). The examples in (1)-(5) all include a non-alternating verb, and they exemplify different combinations of the two between-item manipulations of aspect and alternation.

Note that verbs in an inchoative form can be syntactically different in Italian and Spanish from the other three languages. In particular, a reflexive particle can enter the construction. While in Italian this depends on the verb and can be avoided (as in (4-b)), in Spanish these verbs require a

	N (N_m)	Age: range	mean	σ
English	42 (36)	22–70	37.1	11.3
French	42 (22)	18–55	30.9	10.0
German	31 (25)	18–55	31.9	10.9
Italian	43 (31)	18–48	29.7	8.3
Spanish	45 (27)	18–67	33.0	9.7
	203 (141)	18–70	32.4	10.2

Table 2: Summary statistics regarding the participants. The first column indicates the total number of participants as well as the number of monolingual participants (in brackets).

reflexive particle, as in (5-b), which literally translates to “The pain that annoyed me passed itself suddenly”. In Italian we avoided all verbs where a reflexive was necessary to ensure grammaticality. In Spanish the only possibility was including the reflexive in all alternating verbs. While it could be hypothesised that the reflexive, which corefers with the entity antecedent by necessity, could prime an entity reading, our results show the opposite direction, with these verbs prompting more event continuations (see §3.5.5): since no priming can be noted and the inclusion of the reflexive is mandatory in Spanish, we consider these stimuli comparable to those in the other four languages.

3.3 Participants and procedure

Participants were recruited on Amazon Mechanical Turk, targeting users from specific countries by IP address (i.e. USA, France, Germany, Italy and Spain). They received \$8/€7 for an estimated 45-60 minutes task. We excluded non-native speakers from all analyses. Summary demographics for the resulting datasets by language are reported in Table 2. Participants were not controlled for class or education level. To decide whether to include bilinguals since birth or only monolingual speakers of the target language, this factor was added to the models and a model comparison was run between the two models: in none of the languages the addition of this information improved the model’s fit (with p ranging between 0.12 and 0.91 for the bilingualism factor). Bilingual participants were thus included in the data set.

Continuations were collected via a web-based interface that participants could access from their own computer through MTurk. The website displayed a background questionnaire, a consent form and an instructions page, then each item was presented on a page by itself with a text box for participants to use for writing their continuation.

3.4 Annotation

The continuations were double-annotated for event or entity coreference. The annotators, who included the authors and were either native speakers or very competent speakers with a background in linguistics, based their decision on the semantics of the continuation. The pronominal prompt was hidden during the annotation process to avoid influencing the interpretation. Continuations which did not include a subject-position reference to the event or entity included in the prompt were excluded from analysis. This included a total of 1345 continuations, equivalent to 23.9% of the

initially gathered data, such as pleonastic uses (e.g. *It was still foggy*), adjectival uses (e.g. *This wine was great*), or reference to another unrelated entity (e.g. *This is a beautiful morning*). Each remaining continuation was annotated with a *strict* and a *liberal* interpretation in parallel. In the strict annotation, an example was annotated as ambiguous if there was the slightest doubt about the correct reading. In the liberal annotation, annotators were allowed to use their intuitive judgements when both readings were possible, but one of them seemed overwhelmingly more likely.

Note that this means that some cases deemed ambiguous in the strict annotation are disambiguated in the liberal annotation. We chose to annotate the data both strictly and liberally to be able to evaluate the trade-off between quantity and noise of the data: the strict annotation necessitates excluding more data points from the analysis, but the liberal annotation leads to more noisy (although more numerous) data. The comparison between models computed on the strict and liberal annotation is described in Section §3.6.

To reconcile the double annotations, the following rules were applied:

- The labels used for the valid continuations were: Event, Entity, Ambiguous;
- If both annotators agreed on a label in the strict annotation, the same label was also assigned for the liberal annotation;
- When the two annotators disagreed (that is, one annotation was “Event” and the other was “Entity”), the continuation was excluded from the analyses (as were invalid continuations);³
- In the liberal annotation scheme only, if one annotator assigned “Event” or “Entity” and the other labelled it as “Ambiguous”, the Event/Entity label was chosen (i.e. if one annotator did not resolve an ambiguous reading, the opinion of the other prevailed): these continuations are only included in the liberal analyses in §3.6.

3.5 Results with the strict annotation

The strict annotation was used as main analysis: comparing the same models computed with data annotated strictly and liberally, the models with strict data showed better fits (with, e.g., BICs being up to halved). The results from the liberal analyses are reported in §3.6 as a comparison with the strict analyses and in Appendix A. Given that the analysis on the strict data was later repeated with a different model (as reported below in §3.7), a Bonferroni correction was applied to the significance thresholds dividing them by two (the number of total analyses). These Bonferroni-corrected thresholds as well as the results of the models are reported in Table 3.

In all languages, non-native participants were excluded from the analysis, and the data was subset to continuations annotated as either event or entities (in the strict annotation scheme). Participants who described themselves as native but not monolingual were included in the analysis: model comparison showed that the addition of this further variable to the models did not improve the fit. The number of native participants and of monolinguals is reported in Table 2, the total number of data points for each language is reported in the respective sections. Alternation and aspect were deviation-coded in all languages: this means that, when reporting the results, only one estimate (the effects of perfective aspect and alternation) is reported; the significance and estimate for the opposite value of each factor (imperfective and non-alternating) are the same, with opposite signs. The

3. The percentages of disagreements for the strict annotation were: English, 17.7%; French, 21.9%; German, 13.7%; Italian, 18.1%; Spanish, 4%.

coding of the referring expression factor is three-way in all languages but English, and it is detailed in each language's section.

All models used generalised mixed-effects logistic regression (Bates et al., 2015b) and were computed with the `lme4` package (Bates et al., 2015b) in R (R Development Core Team, 2008).

3.5.1 ENGLISH

The analysis for English should replicate the results of Loáiciga et al. (2018), wherein we found a bias for *It* to corefer with entities and *This* with events, along with an effect of verb type, with verbs permitting alternation yielding more event coreference. The current analysis adds verbal aspect as a further predictor. The data included a total of 539 observations.

The model predicted the `isEvent` binary outcome using referential Form (*it* vs *this*), Aspect, Alternation, and their interactions as binary predictors. All predictors were deviation-coded. To select the best fitting model, models with different fixed effects interactions were compared. The interactions of Alternation and Aspect, Form and Aspect or Form and Alternation did not significantly improve the fit ($p = 0.52$, $p = 0.31$ and $p = 0.75$, respectively). The chosen model thus includes fixed effects for Form, Aspect, and Alternation, with no interactions.

The maximal random effect structure was used when supported by the data (Barr et al., 2013). Where a model did not converge, the random effects were successively removed, chosen by lowest variance. The maximal converging model includes random intercepts and slopes for Alternation by participant and for referential Form by item. Following the recommendation of Bates et al. (2015a), we ran a principal components analysis of the random effects structure, which did not diagnose any overspecification. The model results are reported in Table 3.

The model shows that Form has a significant effect where the use of *This* increases event coreference ($p < 0.0005$, see Figure 1). Verb Alternation and Aspect did not show a significant effect (respectively $p = 0.42$ and $p = 0.93$).

3.5.2 FRENCH

The French analysis followed a similar procedure (cf. §3.5.1). The total number of observations was 582, verb Alternation and Aspect were deviation-coded and the three-way Form was coded as the differences “*C’est – Cela*” and “*Il – C’est*”.

Model comparison showed that the best model was one with no interactions: adding an interaction of Alternation and Aspect, Form and Aspect or Form and Alternation did not improve the model fit (respectively, $p = 0.25$, $p = 0.25$, and $p = 0.49$). The model included a random intercept by participant and random intercept and slopes for the Form by item. The results of the model are reported in Table 3.

The fixed effects show a significant difference between *Il* and *C’est* ($p < 0.0005$), with *Il* yielding fewer event continuations than either of the other two variants. The difference between *C’est* and *Cela*, on the other hand, is not significant ($p = 0.48$). Verb Alternation did not reach significance, with alternating verbs prompting more event continuations ($p = 0.05$), nor did aspect ($p = 0.82$).

Given the three-way nature of Form, its effects were confirmed by subsetting the data to two of the three variants. In the model with *Il* and *C’est* only, as well as in the that with *Il* and *Cela* only, condition proved significant (both $p < 0.0005$ and with the same direction of effect as in the full model); in the model with *C’est* and *Cela* only, condition was not significant ($p = 0.25$),

confirming that *C'est* and *Cela* do not pattern significantly differently in how they bias event or entity coreference (cf. Figure 1). Alternation was significant only in the model with *Il* and *C'est* ($p = 0.04$), where alternating verbs prompted more event continuations.

3.5.3 GERMAN

The German data included a total of 405 observations. As Form is a three-way factor (*Es* vs *Das* vs *Dies*), it was centred as the difference between pairs of its values, i.e. “*Dies – Das*” and “*Es – Dies*”.

Following a similar procedure as that outlined in §3.5.1 and §3.5.2, models were compared to select significant interactions. However, no interaction significantly improved the model fit (Alternation and Aspect: $p = 0.7$, Form and Aspect: $p = 0.28$, Form and Alternation: $p = 0.18$), so the chosen model includes predictors for the Form, Aspect, and Alternation with no interactions. The maximally converging model includes random intercepts by participant and by item. The results of the model are reported in Table 3.

A significant difference is confirmed between *Es* and *Dies*, wherein *Dies* also yields more event coreference than *Es* ($p < 0.0005$). *Dies* and *Das* are only borderline different in their influence on coreference ($p = 0.06$, cf. Figure 1). Alternation and Aspect did not reach significance (both $p = 0.07$), however the direction of effect for Alternation is the same one as in our previous English study and in the French data (alternating verbs being biased towards events).

The effects of Form were confirmed subsetting the data into pairs of the three possible pronouns and fitting new models. Form was significant in all three models (with largest $p = 0.04$), with the directions expected from the full model. Verb Aspect showed a significant effect, with perfective verbs prompting more event continuations, in all three models (largest $p = 0.04$), while Alternation showed a significant effect in the model with *Es* and *Dies* and in that with *Es* and *Das*, with alternating verbs yielding more event continuations (both $p = 0.04$).

3.5.4 ITALIAN

The Italian analysis also followed a similar procedure (cf. §3.5.1-3.5.3). The three-way Form was coded as the differences “*Ciò – Null*” and “*Questo – Ciò*”. The total number of observations was 343.

Model comparison showed that the interaction of Form and verb Alternation significantly improved the model fit over a model with no interactions ($p = 0.04$), while the other two interactions did not (Alternation and Aspect: $p = 0.48$, Condition and Aspect: $p = 0.66$). The random effects included random intercepts by participant and item. The model results are reported in Table 3.

The model output shows a general tendency for event continuations in the intercept, due to our choice of referring expressions in the design ($p < 0.0005$), and a significant difference between the null pronoun and *Ciò* ($p < 0.0005$, see Figure 1), whereby the null pronoun yields fewer event continuations, while the difference between the two overt pronouns is not significant ($p = 0.36$). Neither verb Alternation nor Aspect reached significance (respectively, $p = 0.92$ and $p = 0.3$). Neither interaction reached significance.

Again, the effects of Form were confirmed subsetting the data to two of the three variants. Form was found to be significant in the model with the null pronoun and *Ciò* and in that with the null pronoun and *Questo* (both $p < 0.001$), but not in the model with the two demonstratives ($p = 0.79$), confirming that the pronoun with a different pattern with regard to event vs entity coreference is

the null pronoun, which is biased towards entity continuations. In the model including the null pronoun and *Questo* only and the one including *Questo* and *Ciò*, the interaction between Form and Alternation was significant: specifically, both null pronouns and *Ciò* were used proportionally less than *Questo* to refer to events when an alternating verb was present (respectively $p = 0.04$ and $p < 0.001$).

3.5.5 SPANISH

The Spanish analysis followed a similar procedure as the other languages (cf. §3.5.1-3.5.4). The three-way Form was coded as the differences “Esto – Este” and “Null – Esto”. The number of observations was 453.

Like for Italian (§3.5.4), the model with an interaction between Form and Alternation improved the fit ($p = 0.01$), while the other interactions did not (Alternation and Aspect: $p = 0.3$, Form and Aspect: $p = 0.18$). The model thus included this interaction alongside the main effects, as well as random intercepts by participant and item. The results of the model are reported in Table 3.

The model shows a significant difference between *Esto* and *Este*, and between the null pronoun and *Esto* (both $p < 0.0005$): *Esto* yields more event coreference than either *Este* or the null pronoun. While Aspect did not seem to have an effect on event vs entity coreference ($p = 0.51$), there was a main effect of Alternation whereby verbs allowing alternation were biased towards event coreference ($p = 0.001$). Alternation was also significant in interaction with “Null – Esto”: null pronouns are used more than *Esto* to refer to events in the presence of alternating verbs ($p = 0.007$), modulating the direction of the main effect.

The effects of Form were confirmed subsetting the data to two of the three pronouns. Form was significant in all models: in the model with the null pronoun and *Esto* only, as well as in the that with *Este* and *Esto* only, with the same direction of effect as in the full model (respectively $p = 0.03$ and $p < 0.001$); in the model with the null pronoun and *Este*, with *Este* yielding fewer events than the null pronoun ($p = 0.002$) (cf. Figure 1). Alternation was significant in the two models including null pronouns (in the model with null and *Esto*, $p = 0.009$; in that with null and *Este*, $p < 0.001$); moreover, in these two models Alternation was also significant in interaction with Form: with alternating verbs, null pronouns produce proportionally even more event continuations than either *Este* or *Esto* (both $p = 0.02$).

3.6 Results with the liberal annotation

As previously mentioned, the models using the liberal annotation of event or entity score worse in measures of the relative model quality like the BIC, AIC and log-likelihood. We therefore chose to base our analysis on the models with strict annotation. Although the results of the two types of annotation are very similar, it is worth pointing out some of the differences.

In Italian, the interaction between conditions and alternation no longer improved the model ($p = 0.43$), and was as such not included. On the other hand, with the liberally annotated data the interaction between alternation and aspect improved the model fit in English and German (respectively $p = 0.03$ and $p < 0.001$). In English this interaction reached significance: alternating verbs in the perfective aspect had a smaller effect than in the imperfective in eliciting event continuations ($p = 0.02$).

As a main effect, alternation often reached significance, with alternating verbs yielding more event coreference in both English and French (respectively $p = 0.04$ and $p = 0.03$), as well as in

EVENT VS ENTITY INTERPRETATION

Effect	Estimate	Std.Error	z value	$Pr(> z)$	
English					
(Intercept)	0.21	0.52	0.41	0.68	
Condition (this)	6.71	0.98	6.87	< 0.0005	***
Aspect (perf)	-0.07	0.80	-0.08	0.93	
Alternation (alt)	0.87	1.07	0.81	0.42	
French					
(Intercept)	0.10	0.53	0.85	0.85	
C'est – Cela	-0.42	0.59	-0.71	0.48	
Il – C'est	-7.02	1.21	-5.81	< 0.0005	***
Aspect (perf)	0.19	0.84	0.22	0.82	
Alternation (alt)	1.70	0.87	1.95	0.05	
German					
(Intercept)	1.06	0.49	2.15	0.03	
Dies – Das	-1.30	0.69	-1.88	0.06	
Es – Dies	-6.79	0.88	-7.71	< 0.0005	***
Aspect (perf)	1.47	0.82	1.80	0.07	
Alternation (alt)	1.47	0.80	1.84	0.07	
Italian					
(Intercept)	2.26	0.46	4.92	< 0.0005	***
Ciò – Null	5.99	1.08	5.54	< 0.0005	***
Questo – Ciò	-0.78	0.85	-0.92	0.36	
Aspect (perf)	0.72	0.69	1.04	0.30	
Alternation (alt)	-0.07	0.79	-0.10	0.92	
(Ciò – Null):Altern	-0.27	1.52	-0.18	0.86	
(Questo – Ciò):Altern	3.04	1.69	1.80	0.07	
Spanish					
(Intercept)	0.41	0.40	1.03	0.31	
Esto – Este	7.16	0.86	8.37	< 0.0005	***
Null – Esto	-5.83	0.88	-6.60	< 0.0005	***
Aspect (perf)	0.43	0.65	0.66	0.51	
Alternation (alt)	2.36	0.72	3.27	0.001	**
(Esto – Este):Altern	-1.05	1.13	-0.93	0.35	
(Null – Esto):Altern	3.85	1.43	2.69	0.007	*

Bonferroni-corrected significance: *** $p < 0.0005$, ** $p < 0.005$, * $p < 0.025$

Table 3: Estimated models fixed effects (strict annotations).

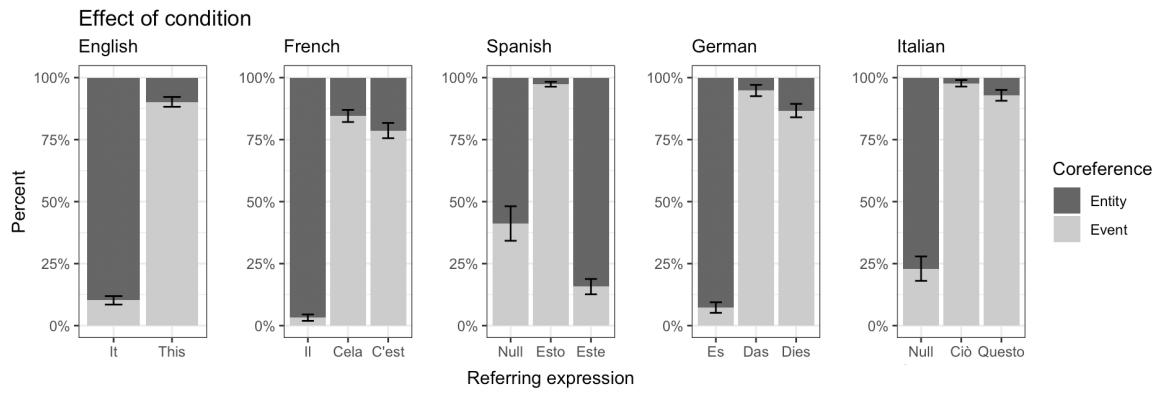


Figure 1: Event and entity coreference by pronominal prompt in the five languages.

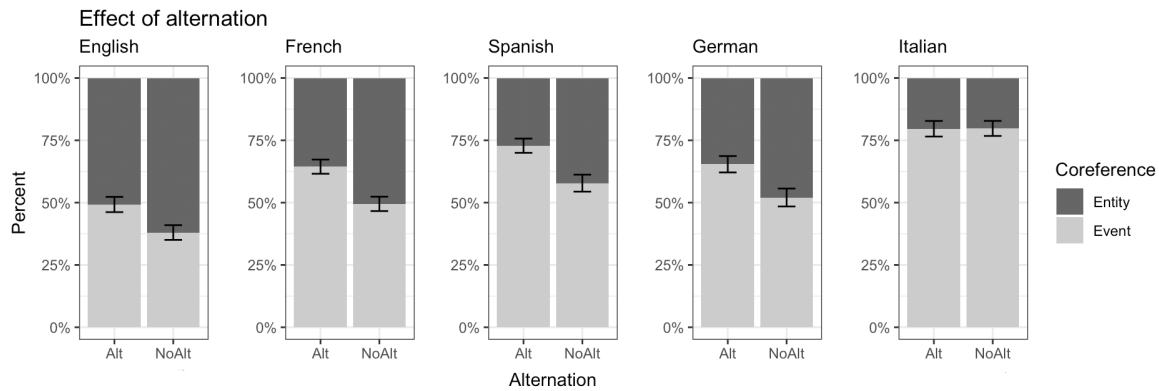


Figure 2: Event and entity coreference by verb type in the five languages.

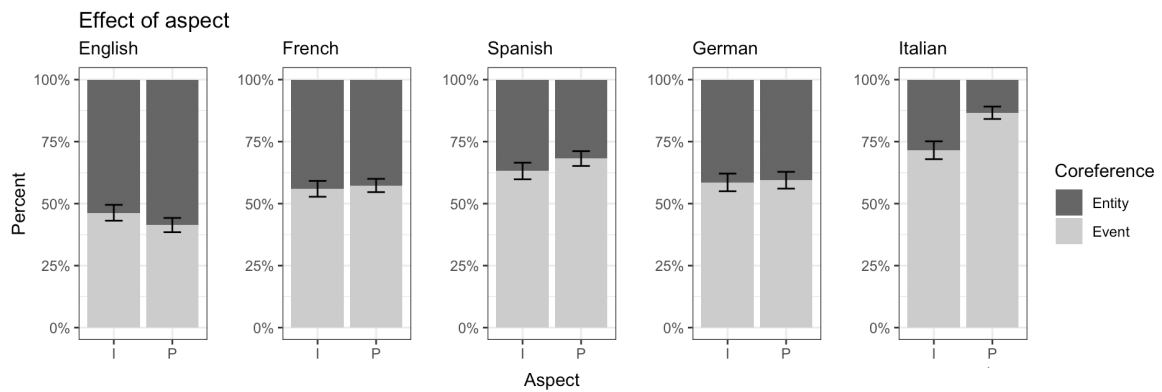


Figure 3: Event and entity coreference by verb aspect in the five languages.

Spanish, where it also showed significance in the strict annotation ($p = 0.001$). In the Italian model, alternation did not reach significance ($p = 0.81$), but the perfective aspect significantly yielded more event continuations, unlike in the strict model ($p = 0.04$).

Finally, in the Spanish model the interaction between alternation and “Null–Esto” reached significance ($p = 0.03$): while generally *Esto* yields more event continuations than the null pronoun, this effect is less pronounced in alternating verbs than in non-alternating verbs.

3.7 Cross-language model

Given the p values of alternation often being borderline (particularly before the Bonferroni correction), as in French ($p = 0.05$) and German ($p = 0.07$), a model was run on the strict annotation including data from all five languagees. This model was only run on the strict annotation of the data. To diminish the risk of Type I error, brought on by repeating the analysis a second time, a Bonferroni (1936) correction was applied to the significance levels, dividing them by two (the number of languages and, accordingly, previous models). The Bonferroni correction was also retro-actively applied to the model described in §3.5 and reported in Table 3. The model results are reported in Table 4.

While adding interactions between aspect and alternation or aspect and condition did not improve the model’s fit over a model with no interactions (respectively $p = 0.32$ and $p = 0.79$), the interaction of alternation and condition did ($p = 0.005$), and was thus included in the model specification. A model with the three-way interaction did not converge.

The selected model modelled event coreference as a binary, with verb aspect, alternation, referring expression and the interaction of alternation and referring expression as predictors, as well as random intercepts for participant and item, both nested within language. All the referring expressions in Table 1 were included, distinguishing between null pronouns in Italian and in Spanish. All predictors were sum-coded: the model estimates are then to be read as a deviation from the intercept, i.e. as a bias of the referring expression towards events or entities (respectively positive and negative β).

While aspect did not show a significant crosslinguistic effect ($p = 0.24$), alternation did ($p < 0.001$), with alternating verbs yielding more event continuations. All referring expressions deviated significantly from the intercept, or, in different terms, their yield of event or entity continuations was not at chance level (all at $p < 0.001$). More specifically, personal pronouns were biased towards entity coreference and demonstratives were biased towards events, with the exception of *Este* in Spanish, which yields more entity coreference (as already seen in the language-internal model, §3.5.5).

Finally, the interactions between the Italian and Spanish null pronoun conditions and alternation reached significance: while in Italian the null pronoun reduces the effect of alternating verbs ($p = 0.01$), in Spanish it enhances it ($p = 0.003$). This echoes findings on the distribution and behaviour of null pronouns in the two languages, which seem to differ in similar contexts (e.g. Russo et al., 2012; Filiaci, 2010).

4. Discussion

The results of the five studies display similarities between our target languages. In all languages, heavier referring expressions (specifically, demonstratives) bias coreference towards event continuations (as shown in Figure 1). This is in keeping with theories positing that richer, more uniquely-

Effect	Estimate	Std.Error	z value	$Pr(> z)$	
(Intercept)	0.67	0.17	3.92	< 0.0005	***
Aspect (perf)	0.17	0.14	1.17	0.24	
Alternation (alt)	0.59	0.16	3.74	< 0.0005	***
English					
Condition (it)	-3.36	0.38	-8.75	< 0.0005	***
Condition (this)	2.04	0.39	5.25	< 0.0005	***
Altern:it	0.09	0.34	0.27	0.78	
Altern:this	-0.52	0.35	-1.47	0.14	
French					
Condition (il)	-5.36	0.65	-8.26	< 0.0005	***
Condition (cela)	1.71	0.37	4.57	< 0.0005	***
Condition (c'est)	1.31	0.38	3.48	< 0.0005	***
Altern:il	0.64	0.61	1.04	0.30	
Altern:cela	-0.35	0.34	-1.01	0.31	
Altern:c'est	0.19	0.35	0.55	0.59	
German					
Condition (es)	-3.51	0.38	-9.13	< 0.0005	***
Condition (das)	2.12	0.42	5.03	< 0.0005	***
Condition (dies)	2.01	0.37	5.46	< 0.0005	***
Altern:es	0.14	0.35	0.41	0.68	
Altern:das	-0.53	0.40	-1.34	0.18	
Altern:dies	-0.18	0.34	-0.54	0.59	
Italian					
Condition (null)	-2.44	0.45	-5.49	< 0.0005	***
Condition (ciò)	3.24	0.52	6.28	< 0.0005	***
Condition (questo)	3.09	0.60	5.16	< 0.0005	***
Altern:null	-1.07	0.41	-2.58	0.01	*
Altern:ciò	-0.60	0.48	-1.25	0.21	
Altern:questo	0.52	0.58	0.90	0.37	
Spanish					
Condition (null)	-1.69	0.56	-3.00	0.003	**
Condition (este)	-3.04	0.43	-7.06	< 0.0005	***
Condition (esto)	3.88	0.51	7.57	< 0.0005	***
Altern:null	1.62	0.55	2.94	0.003	**
Altern:este	0.28	0.39	1.71	0.48	
Altern:esto	-0.23	0.48	-0.48	0.63	

Bonferroni-corrected significance: *** $p < 0.0005$, ** $p < 0.005$, * $p < 0.025$

Table 4: Estimated models fixed effects for the cross-language model.

referential expressions (like demonstratives, that add a distal trait to those of personal pronouns and null forms) will be used to retrieve less “stereotypical” material (e.g. Ariel, 1988): given that event coreference is much rarer than entity coreference (cf. §1), it is likely that events will be implicitly treated as a marked case, granting the use of expressions used for less accessible antecedents. A notable exception to this pattern is the Spanish demonstrative *este*, which was not only biased towards entity continuations, but even more so than the null pronoun. One possibility is that, unlike its neuter counterpart *esto*, the masculine form *este* could be used preferably for (masculine) entities, for analogy with the feminine form *esta*. Note that the masculine is the default gender of the entities used in the experiment.

The results of the studies often agree with the (rough) prescriptions of grammars. The RAE (2010) predicted the use of Spanish *esto* to refer to antecedents other than entities, and similarly in French the Grevisse and Goosse (2008) grammar predicted the use of *cela* and *ce* for non-entities. The Italian data agrees with Serianni (1997): both *ciò* and *questo* are dispreferred for entity antecedents. For German, the Duden (2009) grammar does not provide a very clear indication of use, in that it says that both *es* and *das* can be used to refer to non-entities (however differently) – yet, our data shows a marked tendency for *es* to refer to entities. However, our studies differ from these prescriptions in that they offer an estimation of the degree to which the different forms are non-categorical.

Comparing the results with those of corpus and psycholinguistic studies also yields some interesting observations. The English results confirm a contrast in use between the personal and demonstrative pronouns already found in Loáiciga et al. (2018), Navarretta (2007) and Dipper et al. (2011), which also agrees with our results on German, in which *es* continues referential chains of topical entities while *das* refers to non-nominal antecedents. Dipper et al. (2011) does not find a clear distinction in use between demonstrative and personal pronouns in Spanish, and our results add some detail to the claim showing how *este* does not pattern as expected. In Italian, Navarretta (2007) finds that while *ciò* is used for non-entities, *questo* has a light preference for entities: this is in contrast with the very clear tendency we found for the demonstrative to retrieve events. The French results are in line with Vieira et al. (2005) in that demonstratives are used for abstract antecedents more often.

Finally, our studies targeted two verbal features: alternation and aspect. Alternating verbs yielded more event continuations in more languages, even if the effect did not reach significance in Italian. In Italian, the two demonstratives show ceiling selection of the events as antecedents: this near-categorical behaviour may have obscured the effect. The crosslinguistic effect of alternation is visualised in Figure 2.

On the other hand, aspect did not show a significant effect across languages; nonetheless, the direction is quite consistently the same in which the representation of an event as completed through a perfective aspect yields more events (except for English: see Figure 3). This, however, was only significant in the Italian model run with the liberal annotation of our data and in interaction with alternation in English, whereby non-alternating verbs yield more event continuations in the perfective aspect, whereas alternating verbs do so in the imperfective.

5. Comparison with annotated coreference

In order to contextualise our findings in the coreference resolution scenario, we extract coreference relations from a multilingual parallel corpus to check whether the same patterns observed in the previous experiments are also apparent in the annotated corpus data.

We work with the ParCorFull corpus, which includes coreference annotations for the English, German and French languages. ParCorFull includes texts from TED talks transcripts and newswire data originally in English, and comprises approximately 160,000 tokens. The underlying coreference scheme was designed for uniform annotations across the languages (see Lapshinova-Koltunski and Hardmeier, 2018, for details). The annotated elements (markables) in this corpus include pronouns, nouns, nominal phrases or elliptical constructions that are parts of a coreference pair (antecedent-anaphora), as well as verb phrases or clauses being antecedents of event anaphora. The annotated antecedents are of two different types: entities and events. Entities can be either pronouns or noun phrases, whereas events include verb phrases, clauses, or a set of clauses.

5.1 Event vs entity antecedent proportions

To reproduce the parameters of the experiments described above, we first extract all coreference chains headed by lexical entity or event antecedents, thus excluding cataphora. We then retain all mentions of the pronouns of interest (EN: *this, it*; DE: *es, das, dies*; FR: *c', il, cela*) in a subject position and order them according to their appearance in the text.

Table 5 summarises the distribution of the mentions retained after filtering. The index $i = n$ reflects the order of re-mention of the antecedent by a pronoun of interest. $i = 1$ corresponds to equivalent cases to those produced in the experiments with human participants, where a particular type of antecedent is re-mentioned for the first time with one of the different pronouns. $i = 2$ corresponds to the second re-mention of an antecedent, $i = 3$ to the third, and so forth. Note that while this is a parallel corpus, the number of times there is a re-mention of an antecedent varies in each of the languages, with French being the language with the most re-mentions. Additionally, for all the languages, it can be seen that there is a clearly preferred form for re-mention pervasive through time as the text increases in length, regardless of the type of antecedent. These are light forms: *it* for English, *es* for German, and *il* for French.

In order to assess Webber's 1991 proposal that demonstrative pronouns are used to refer for the first time to parts of the context which are in focus, making them available to serve as antecedents for personal pronouns later, we looked closely at the English examples corresponding to the columns $i = 1$ and $i = 2$ in Table 5. From the 54 events first introduced with *this*, we found 4 examples rementioned with *it*, as illustrated in (6). These are too few cases to draw any clear conclusion, leaving the question open of whether this pattern is upheld more frequently with the demonstrative *that*.⁴

- (6) But just *if you take malaria out, deaths from everything else go down*. And the economist Jeff Sachs has actually quantified what *this* means for a society. What *it* means is, if you have malaria in your society, your economic growth is depressed by 1.3 percent every year, year after year after year, just this one disease alone.

4. As for cases where a demonstrative pronoun is also rementioned with a demonstrative pronoun we found five cases with *this*; 16 with *das*, 2 with *dies*; and no cases in French.

EVENT VS ENTITY INTERPRETATION

Antecedent	Anaphor	Re-mention index											
English		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$	$i = 10$	$i = 11$	$i \geq 12$
Entity	<i>this</i>	24	21	5	1	0	1	1	0	0	0	0	0
	<i>it</i>	220	113	63	34	22	15	4	0	0	0	2	9
Event	<i>this</i>	54	10	4	1	0	1	0	0	1	1	0	0
	<i>it</i>	60	42	18	8	5	1	2	1	0	0	0	0
German		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$	$i = 10$	$i = 11$	$i \geq 12$
Entity	<i>es</i>	87	48	24	9	5	1	2	0	2	0	1	1
	<i>das</i>	118	14	9	0	1	3	0	1	0	0	0	2
	<i>dies</i>	4	3	1	0	1	0	0	0	0	0	0	0
Event	<i>es</i>	20	17	8	3	1	0	0	0	0	0	0	0
	<i>das</i>	187	26	5	0	0	0	0	0	0	0	0	0
	<i>dies</i>	10	3	1	0	0	0	0	0	0	0	0	0
French		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$	$i = 10$	$i = 11$	$i \geq 12$
Entity	<i>c'</i>	74	38	19	12	9	2	1	1	1	0	2	2
	<i>il</i>	62	47	26	21	20	13	11	5	6	4	5	56
	<i>cela</i>	16	2	3	4	0	1	0	0	1	0	0	2
	<i>c'</i>	90	25	7	3	0	0	1	0	0	0	1	0
Event	<i>il</i>	1	3	0	0	0	0	0	0	0	0	0	0
	<i>cela</i>	30	7	4	0	0	0	0	0	0	0	0	0

Table 5: Re-mention frequency of entity and event antecedents in English, German and French with the anaphors of interest as annotated in the ParCorFull corpus. The index $i = n$ represents the subsequent order in which the pronoun appears after an antecedent has been introduced.

	Antecedent	English		German			French		
		<i>this</i>	<i>it</i>	<i>es</i>	<i>das</i>	<i>dies</i>	<i>c'</i>	<i>il</i>	<i>cela</i>
Human responses	Entity	22 4%	282 52%	140 35%	5 1%	21 5%	37 6%	180 31%	33 6%
	Event	203 38%	32 6%	11 3%	91 22%	137 34%	140 24%	6 1%	186 32%
Corpus annotation	Entity	24 7%	220 61%	87 20%	118 28%	4 1%	74 27%	62 23%	16 6%
	Event	54 15%	60 17%	20 5%	187 44%	10 2%	90 33%	1 0%	30 11%

Table 6: Comparison between the proportions of event and entity antecedent interpretations in the human experiments with the annotations in ParCorFull of antecedents and their first re-mention ($i = 1$). Note that the percentages are computed using the total counts per language.

A direct comparison between the experiments with human participants and the corpus data is summarised in Table 6. As it is observed in the human continuations, events are mostly referred to with demonstrative pronouns and entities with personal pronouns, but this preference is not categorical. In English, for example, events in the corpus have a similar proportion of *it* vs *this*. In German, the clear preference is *das*, while in French is *c'* and not *cela* as one might expect. While the human event coreference rates are more balanced across languages, as a result of controlling the parameters for the experiment, the proportion of events in the corpus varies per language. For English, there is around 32% events in contrast to 68% entities, while for German the proportion is evenly balanced with 51% events. The difference is worth noting since the documents in the corpus are complete parallel versions of each other, suggesting a difference in conceptualisation at the time of translation. French, on the other hand, is in the middle with 44% events. However, this portion of the corpus contains fewer documents than its English and German counterparts.

5.2 Governing verb alternation status

In order to draw a similar comparison with respect to the alternation status of the verbs in the corpus, we also extracted the verbs relevant for the entities and events reported in Table 6. In the case of entity antecedents, we extracted the verb to which the head of the noun phrase is attached. For event antecedents, we extracted the verb from the antecedent itself.

Studies on the causative alternation propose that verbs can be ranked on a universal scale of likelihood of spontaneous occurrence (Haspelmath, 1993). In this scale, verbs are ranked according to the degree to which they are non-agentive (Samardžić, 2014, p. 180). This means that verbs ranking higher in spontaneous occurrence can occur without an explicit agent causing the event, e.g., *break*, *open*, and are hence more likely to participate in the causative alternation. Following Haspelmath and using large scale corpus data, as opposed to typological observations, Samardžić

estimates the degree of spontaneity for the 354 English verbs permitting the causative alternation reported by Levin (1993). The underlying assumption for the spontaneity score is that causative and anticausative uses of a verb have a correlation with transitive and intransitive examples in corpora (reported to be $r = 0.67$, $p < 0.01$ using the Spearman test).

We then compare the verbs extracted from the ParCorFull antecedents against the list of scored verbs reported in Samardžić (2014). Unfortunately, being restricted to Levin's verbs, this list is very small, and contains only a few of our verbs. The extracted English verbs which matched the list are listed in Appendices B and C. We could not find a similar resource for French or German.

It is very difficult to estimate a similar spontaneity score for our verbs since, unlike Samardžić who worked with Levin's list and the Proposition Bank (Palmer et al., 2005) which is annotated for semantic roles, we have many verbs for which we do not have any sort of gold standard resource that help us estimate the accuracy of our calculations. In our case, equating transitive uses with alternating verbs (causatives) and intransitive uses with non-alternating verbs (anti-causatives) is a dangerous simplification since many verbs in the corpus do not necessarily have the proper frame of thematic roles playing a part in the causative alternation.

Interpreting and generating coreference involves many levels of linguistic processing, including verb semantics (for the effect of thematic roles, see e.g. Stevenson et al., 1994; Arnold, 2000). Between our experimental items, we manipulated the alternation status of the verb because the experimental framework grants us complete control over the stimuli. Estimating semantic predictors from a corpus is much harder, partly because of the lack of resources annotated with this type of information. With our work, we highlight the value in considering semantic predictors for the processing of coreference.

6. Conclusions

Our study shows a crosslinguistic bias whereby the coreference to entities or events is biased by multiple factors: the referring expressions used and features of the verb constituting the argument structure involving the possible entity antecedents. Moreover, our data gives a clear distribution of event and entity coreference across referring expressions and languages, with controlled manipulations achieving less noise than in a corpus analysis.

Confirming the predictions based on hierarchies describing the use of referring expressions based on their antecedent's accessibility (e.g. Ariel 1988), our results show that lighter referring expressions are biased towards entity antecedents and, conversely, heavier expressions are biased towards events. The lesser accessibility of events can be explained in multiple ways: they are a less common type of antecedent (as corpus measures show: see §1, 5), they are more complex and less easily introduced directly (Hedberg et al., 2007).

The comparisons with a coreference-annotated corpus generally replicate the human results, but they also suggest that these patterns might be further influenced by the time at which an antecedent is re-mentioned in a coreference chain. The corpus results also show slight differences in the referring expressions' proportions, which may be due to stylistic variation. In addition, the fact that the different forms appear non-categorically with either entities or events points towards an interplay between both types of antecedents in the discourse, indicating that studying one without the other might result in an incomplete picture.

We also investigated verbal features and their possible influence on coreference. On the question of verb aspect, this study saw a general tendency for completed events to be taken up as an

antecedent more than ongoing events in all languages but English, but this effect did not reach statistical significance. Further research with fewer confounding factors is needed to confirm whether the mostly uniform direction it showed in our data is upheld.

On the other hand, the event structure of verbs was also shown to clearly influence the coreference patterns. Specifically, further increasing the complexity of an event by having an implicit agent in alternating verbs creates more competition for the explicit entity: since the accessibility of a referent decreases with the increase in the total number of possible referents (Grosz et al., 1995), the event antecedent becomes relatively more accessible and is chosen more as a likely continuation.

Last, the cross crosslingual effect of the verb alternation shows that, from a cognitive point of view, adding one more competitor to a pool of possible antecedents affects the relative accessibility of the other antecedents, even when the newly added competitor is not made explicit. This raises questions about whether introducing implicit and explicit elements has the same effect on coreference (both qualitatively and, especially, quantitatively), and whether an antecedent's conceptual availability as given by world knowledge matters as much as its explicit presence in a discourse. Such differences in the activation of antecedents could be explored via “on-line” psycholinguistic methods (e.g. measurement of reaction times or eye-tracking). We must leave these questions for future research.

Acknowledgements

Hannah Rohde was supported by a Leverhulme Trust Prize in Languages & Literatures. Christian Hardmeier was supported by the Swedish Research Council under grant 2017-930. The authors gratefully acknowledge the funding. We also thank Sebastiano Gigliobianco for his help optimizing our extraction scripts, and Elina Lartaud, Ludovica Inserra and Paola Jalili for helping with data annotation. Last, are also grateful to the anonymous reviewers for their insightful comments that helped improve this paper.

References

- Luis Alonso-Ovalle, Susana Fernández-Solera, Lyn Frazier, and Charles Jr. Clifton. Null vs. overt pronouns and the topic-focus articulation in Spanish. *Rivista di Linguistica*, 14(2):151–170, 2002.
- Mira Ariel. Referring and accessibility. *Journal of Linguistics*, 24(1):65–87, 1988.
- Mira Ariel. Accessibility marking: Discourse functions, discourse profiles, and processing cues. *Discourse Processes*, 37(2):91–116, 2004.
- J. E. Arnold. The effect of thematic roles on pronoun use and frequency of reference continuation. *University of Pennsylvania Working Papers in Linguistics*, 6(3):209–235, 2000.
- Nicholas Asher. *Reference to Abstract Objects in Discourse*. Springer Netherlands, Dordrecht, 1993.
- Hiroimi Azuma. A diachronic view of pronominal reference in English. In Christer Johansson, editor, *Second Workshop on Anaphora Resolution, WAR II*, volume 2, pages 1–9, Bergen, 2008.

- Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3): 255–278, 2013.
- Douglas Bates, Reinhold Kliegl, Shrvan Vasishth, and Harald Baayen. Parsimonious mixed models, 2015a. URL <https://arxiv.org/abs/1506.04967>.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015b.
- Kristine Bentzen and Merete Anderssen. The form and position of pronominal objects with non-nominal antecedents in Scandinavian and German. *The Journal of Comparative Germanic Linguistics*, 22(2):169–188, 2019.
- Shane Bergsma and David Yarowsky. NADA: A robust system for non - referential pronoun detection. In Iris Hendrickx, Sobha Lalitha Devi, António Branco, and Ruslan Mitkov, editors, *Anaphora Processing and Applications: 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, Lecture Notes in Artificial Intelligence, pages 12–23. Springer, Faro, 2011.
- Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.
- Adriane Boyd, Whitney Gegg-Harrison, and Donna K. Byron. Identifying non-referential *it*: a machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, pages 40–47, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.
- Sarah Brown-Schmidt, Donna K. Byron, and Michael K. Tanenhaus. Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language*, 53(2):292–313, 2005.
- Donna K. Byron. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL 2002, pages 80–87, Philadelphia, 2002. Association for Computational Linguistics.
- Gregory Carlson. Reference. In L. R. Horn and G. Ward, editors, *The Handbook of Pragmatics*. Wiley-Blackwell, Oxford, 2003.
- Maria Nella Carminati. *The Processing of Italian Subject Pronouns*. Doctoral dissertation, University of Massachusetts Amherst, 2002. URL <https://scholarworks.umass.edu/dissertations/AAI3039345>.
- Derya Çokal, Patrick Sturt, and Fernanda Ferreira. Deixis: *This* and *That* in written narrative discourse. *Discourse Processes*, 51(3):201–229, 2014.
- Derya Çokal, Patrick Sturt, and Fernanda Ferreira. Processing of *it* and *this* in written narrative discourse. *Discourse Processes*, 55(3):272–289, 2018. doi: 10.1080/0163853X.2016.1236231. URL <https://doi.org/10.1080/0163853X.2016.1236231>.

- Bernard Comrie. Pragmatic binding: Demonstratives as anaphors in Dutch. In *Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure*, pages 50–61, Berkeley, 1997.
- Francis Cornish. *Anaphoric Relations in English and French: A Discourse Perspective*. Taylor & Francis, Abingdon-on-Thames (UK), 2015.
- Stefanie Dipper, Christine Rieger, Melanie Seiss, and Heike Zinsmeister. Abstract anaphors in German and English. In Iris Hendrickx, Sobha Lalitha Devi, António Branco, and Ruslan Mitkov, editors, *Anaphora Processing and Applications. 8th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2011*, pages 96–107. Springer, Faro, Portugal, 2011.
- Dudenredaktion. *Duden. Die Grammatik: Unentbehrlich für richtiges Deutsch*. Dudenverlag, Mannheim–Wien–Zürich, 2009.
- Miriam Eckert and Michael Strube. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89, 2000.
- Richard Evans. Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing*, 16(1):45–57, 2001.
- Todd R Ferretti, Ken McRae, and Andrea Hatherell. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547, 2001.
- Todd R. Ferretti, Marta Kutas, and Ken McRae. Verb aspect and the activation of event knowledge. *Journal of experimental psychology. Learning, memory, and cognition*, 3(1):182–196, 2007.
- Todd R. Ferretti, Hannah Rohde, Andrew Kehler, and Melanie Crutchley. Verb aspect, event structure, and coreferential processing. *Journal of memory and language*, 61(2):191–205, 2009.
- Francesca Filiaci. Null and overt subject biases in Spanish and Italian: A cross-linguistic comparison. In *Selected Proceedings of the 12th Hispanic Linguistics Symposium*, pages 171–182, Somerville (MA), 2010. Cascadilla Proceedings Project. URL <https://www.lingref.com/cpp/hls/12/paper2415.pdf>.
- Thomas Givón. *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. John Benjamin, Amsterdam, 1983.
- S. B. Greene, G. McKoon, and R. Ratcliff. Pronoun resolution and discourse models. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18:266–283, 1992.
- Maurice Grevisse and André Goosse. *Le bon usage: Grammaire française*. Duculot, Louvain-la-Neuve (Belgium), 14th edition, 2008.
- Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.
- Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. Pronouns without np antecedents: How do we know when a pronoun is referential? In António Branco, Tony McEnery, and Ruslan Mitkov, editors, *Anaphora Processing: Linguistic, Cognitive and Computational Modelling*, pages 351–364. John Benjamins, Amsterdam, 2005.

- Martin Haspelmath. More on the typology of inchoative/causative verb alternations. In Bernard Comrie and Maria Polinsky, editors, *Causatives and transitivity*, pages 87–120. John Benjamins, Amsterdam, 1993.
- Nancy Hedberg, Jeanette K. Gundel, and Ron Zacharski. Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. In *Proceedings DAARC 2007 (Discourse Anaphora and Anaphora Resolution Colloquium)*, Lagos (PT), 2007.
- Benjamin Heinzerling, Nafise Sadat Moosavi, and Michael Strube. Revisiting selectional preferences for coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1332–1339, Copenhagen, 2017. Association for Computational Linguistics.
- Otto Jespersen. *Modern English grammar on historical principles, Part III: Syntax (Second Volume)*. Allen and Unwin, London, 1927.
- Elsi Kaiser and John C. Trueswell. The role of discourse context in the processing of a flexible word-order language. *Cognition*, 94:113–147, 2004.
- Varada Kolhatkar, Heike Zinsmeister, and Graeme Hirst. Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 300–310, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/D13-1030>.
- Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. Anaphora with non-nominal antecedents in computational linguistics: a Survey. *Computational Linguistics*, 44(3): 547–612, September 2018. URL <https://www.aclweb.org/anthology/J18-3007>.
- Ekaterina Lapshinova-Koltunski and Christian Hardmeier. Coreference corpus annotation guidelines, 2018. URL <https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/11372/LRT-2614/Guidelines.pdf>.
- Ekaterina Lapshinova-Koltunski, Christian Hardmeier, and Pauline Krielke. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of 11th Language Resources and Evaluation Conference*, pages 00–00, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). to appear.
- Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- Timothy Lee, Alex Lutz, and Jinho D. Choi. QA-It: classifying non-referential it for question answer pairs. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 132–137, Berlin, 2016. Association for Computational Linguistics.
- Beth Levin. *English verb classes and alternations: a preliminary investigation*. The University of Chicago Press, Chicago, 1993.

- Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. What is it? disambiguating the different readings of the pronoun ‘it’. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2017, pages 1336–1342, Copenhagen, 2017. Association for Computational Linguistics.
- Sharid Loáiciga, Luca Bevacqua, Hannah Rohde, and Christian Hardmeier. Event versus entity co-reference: Effects of context and form of referring expression. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–103, New Orleans, June 2018. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W18-0711>.
- Sharid Loáiciga, Christian Hardmeier, and Asad Sayeed. Exploiting cross-lingual hints to discover event pronouns. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 99–103, Marseille, France, May 2020. European Language Resources Association. URL <https://www.aclweb.org/anthology/2020.lrec-1.12>.
- Ana Marasovic, Leo Born, Juri Opitz, and Anette Frank. A mention-ranking model for abstract anaphora resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 221–232, Copenhagen, Denmark, 2017. Association for Computational Linguistics.
- Laia Mayol. Asymmetries between interpretation and production in Catalan pronouns. *Dialogue and Discourse*, 9:1–34, 2018.
- Marc Moens and Mark Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28, 1988.
- Christoph Müller. Resolving it, this, and that in unrestricted multi-party dialog. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 816–823, Prague, 2007. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/P07-1103>.
- Costanza Navarretta. A contrastive analysis of abstract anaphora in Danish, English and Italian. In António Branco, Tony McEnery, Ruslan Mitkov, and Fátima Silva, editors, *Proceedings of DAARC 2007*, pages 103–109. Centro de Linguística da Universidade do Porto, 2007.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005. doi: 10.1162/0891201053630264. URL <https://www.aclweb.org/anthology/J05-1004>.
- Rebecca J. Passonneau. Getting at discourse referents. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 51–59, Vancouver, 1989. Association for Computational Linguistics.
- Massimo Poesio, Roland Stuckardt, and Yannick Versley. Challenges and directions of further research. In Massimo Poesio, Roland Stuckardt, and Yannick Versley, editors, *Anaphora Resolution: Algorithms, Resources and Application*, pages 487–500. Springer-Verlag, Berlin–Heidelberg, 2015.

- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, August 2013. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W13-3516>.
- Sameer S. Pradhan and Nianwen Xue. OntoNotes: The 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 11–12, Boulder, Colorado, May 2009. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/N09-4006>.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, 2008. URL <http://www.R-project.org>.
- Real Academia Española (RAE). *Nueva gramática de la lengua española. Manual*. Real Academia Española y Asociación de Academias de la Lengua Española, Madrid, 2010.
- Hannah Rohde. Pronouns. In C. Cummins and N. Katsos, editors, *Oxford Handbook of Experimental Semantics and Pragmatics*. Oxford University Press, Oxford, 2019.
- Lorenza Russo, Sharid Loáiciga, and Asheesh Gulati. Italian and Spanish null subjects. a case study evaluation in an MT perspective. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1779–1784, Istanbul, Turkey, 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/813_Paper.pdf.
- Tanja Samardžić. *Dynamics, causation, duration in the predicate-argument structure of verbs: A computational approach based on parallel corpora*. PhD thesis, University of Geneva, Geneva, Switzerland, 2014.
- Luca Serianni. *Italiano. Grammatica, Sintassi, Dubbi*. Garzanti, Turin, 1997.
- Rosemary J. Stevenson, Rosalind A. Crawley, and David Kleinman. Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548, 1994.
- Agnès Tutin. A corpus-based study of pronominal anaphoric expressions in french. In *Proceedings of DAARC 2002*, Lisbon, 2002.
- Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa Rodriguez, and Massimo Poesio. Annotating a broad range of anaphoric phenomena, in multiple genres: the ARRAU corpus. *Natural Language Engineering*, 26:95–128, 2020.
- Renata Vieira, Susanne Salmon-Alt, Caroline Gasperin, Emmanuel Schang, and Gabriel Othéro. Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. *Anaphora Processing: linguistic, cognitive and computational modeling*, pages 385–403, 2005.
- Klaus von Heusinger and Petra B. Schumacher. Discourse prominence: Definition and application. *Journal of Pragmatics*, 154:117–127, 2019.

- Bonnie Webber. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135, 1991.
- Eva Wittenberg, Shota Momma, and Elsi Kaiser. Demonstratives as bundlers of conceptual structure. *Glossa: A Journal of General Linguistics*, 6(1):33, 2021.
- Victoria Yaneva, Le An Ha, Richard Evans, and Ruslan Mitkov. Classifying referential and non-referential it using gaze. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4896–4901, Brussels, 2018. Association for Computational Linguistics.

Appendix A. Model results with the liberal annotations

In this appendix, we report the results of the models run with the liberal annotation (Table 7). The models were chosen following the same procedure described in §3.5. Other than the fixed effects as shown on the table, the models included the following random effects: a random intercept and slope for condition by participant and for aspect by item in English; random intercepts by participant and intercept and slopes for aspect by item in French; random intercepts by participant and item in German; random intercepts by item only in Italian; and random intercept and slopes for conditions by participant and random intercepts by item in Spanish.

Effect	Estimate	Std.Error	z value	$Pr(> z)$	
English					
(Intercept)	0.41	0.22	1.88	0.06	
Condition (this)	3.96	0.37	10.84	< 0.001	***
Aspect (perf)	-0.27	0.39	-0.69	0.49	
Alternation (alt)	0.80	0.39	2.06	0.04	*
Aspect:Alternation	-1.89	0.80	-2.35	0.02	*
French					
(Intercept)	0.45	0.35	1.27	0.21	
C'est – Cela	-0.12	0.27	-0.42	0.67	
Il – C'est	-4.80	0.43	-11.07	< 0.001	***
Aspect (perf)	-0.21	0.68	-0.30	0.76	
Alternation (alt)	1.45	0.66	2.21	0.03	*
German					
(Intercept)	1.14	0.39	2.91	0.004	**
Dies – Das	-0.77	0.48	-1.61	0.11	
Es – Dies	-5.53	0.57	-9.67	< 0.001	***
Aspect (perf)	0.37	0.64	0.58	0.56	
Alternation (alt)	0.85	0.64	1.34	0.18	
Aspect:Alternation	0.07	1.27	0.05	0.96	
Italian					
(Intercept)	1.67	0.30	5.63	< 0.001	***
Ciò – Null	4.18	0.45	9.34	< 0.001	***
Questo – Ciò	-0.62	0.35	-1.75	0.08	
Aspect (alt)	1.14	0.56	2.04	0.04	*
Alternation (perf)	0.13	0.55	0.24	0.81	
Spanish					
(Intercept)	0.47	0.35	1.34	0.18	
Esto – Este	5.70	0.76	7.53	< 0.001	***
Null – Esto	-4.50	0.65	-6.94	< 0.001	***
Aspect (perf)	0.06	0.48	0.12	0.90	
Alternation (alt)	1.65	0.52	3.20	0.001	**
(Esto – Este):Altern	0.47	0.77	0.61	0.54	
(Null – Esto):Altern	2.10	0.96	2.20	0.03	*

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table 7: Estimated models fixed effects (liberal analysis).

Appendix B. Verbs from ParCorFull and their spontaneity scores

Here we present the list of verbs used with an entity antecedent in ParcorFull. The counts column comes from our extraction, while the causative-rate, anticausative-rate and spontaneity scores are taken from Samardžić (2014).

Verb	Counts	Causative-rate	Anticausative-rate	Spontaneity-score
balance	1	0.18	0.05	1.34
grow	1	0.14	0.78	-1.68
open	1	0.54	0.14	1.33
run	1	0.3	0.56	-0.64

Appendix C. Verbs from ParCorFull and their spontaneity scores

Here we present the list of verbs used as an event antecedent in ParcorFull. The counts column comes from our extraction, while the causative-rate, anticausative-rate and spontaneity scores are taken from Samardžić (2014).

Verb	Counts	Causative-rate	Anticausative-rate	Spontaneity-score
break	1	0.33	0.3	0.09
burn	1	0.23	0.18	0.22
close	1	0.2	0.14	0.39
collect	1	0.27	0.06	1.55
drive	1	0.36	0.11	1.2
fly	1	0.22	0.78	-1.27
move	1	0.11	0.8	-1.97
run	3	0.3	0.56	-0.64
stand	1	0.15	0.85	-1.76