



Maes, Alfons & Krahmer, Emiel & Peeters, David. 2022. Explaining variance in writers' use of demonstratives: A corpus study demonstrating the importance of discourse genre. *Glossa: a journal of general linguistics* 7(1). pp. 1–36. DOI: <https://doi.org/10.16995/glossa.5826>



Open Library of Humanities

Explaining variance in writers' use of demonstratives: A corpus study demonstrating the importance of discourse genre

Alfons Maes, Tilburg University, Department of Communication and Cognition, TiCC, Tilburg, the Netherlands, maes@tilburguniversity.edu

Emiel Krahmer, Tilburg University, Department of Communication and Cognition, TiCC, Tilburg, the Netherlands, e.j.krahmer@tilburguniversity.edu

David Peeters, Tilburg University, Department of Communication and Cognition, TiCC, Tilburg, the Netherlands, d.g.t.peeters@tilburguniversity.edu

Demonstratives such as *this* and *that* are among the most frequently used words in texts. But what are the factors that determine whether a writer uses one demonstrative form (proximal *this*) or another (distal *that*)? Here we report a large-scale corpus analysis in three written genres to empirically contrast theories based on differences in referent activation and prominence with a recent proposal suggesting that genre is the main driver of written demonstrative variance. We consistently observe that discourse genre is indeed the main predictor of writers' demonstrative variation in English text. More specifically, a clear preference for distal demonstratives is found when the addressee is considered more prominent in the given discourse setting (as in news reports), whereas an overall preference for proximal demonstratives is observed when the knowledgeable writer feels more responsibility for the produced discourse themselves, as in an expository context (e.g. wikipedia texts). In such expository contexts, proximal demonstratives hence indicate that the referent is psychologically situated near the writer, whereas in interactional and narrative discourse the writer uses distal demonstratives to reach out to the addressee. These findings shed new theoretical light on some of the most frequently used and studied words in human language.

Glossa: a journal of general linguistics is a peer-reviewed open access journal published by the Open Library of Humanities. © 2022 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

OPEN ACCESS



1 Introduction

Demonstratives such as *this* and *that* are among the most frequently used words in the languages of the world (Levinson et al. 2018). In both spoken and written forms of human communication, they help the language user to refer to something that is deemed relevant to the addressee of the conveyed message. Traditionally, a clear distinction is made between the exophoric and endophoric use of demonstratives, with exophoric demonstratives referring to actual entities in the speech situation, predominantly studied in spoken interaction (“Look at *that* toucan over there!”), and endophoric demonstratives covering all other uses (e.g., Halliday & Hasan 1976; Diessel 1999). Here we focus on the *endophoric* use of demonstratives in written texts.

Interestingly, all spoken languages offer their writers more than one type of demonstrative, such as English *this* versus *that* or Turkish *bu*, *şu*, and *o* (Diessel 1999; Peeters & Özyürek 2016). As a consequence of this variability within and across languages in the availability of different demonstrative devices, a longstanding theoretical question is what determines whether writers include one demonstrative type (e.g. proximal *this*) versus another (e.g. distal *that*) in their referential expression. Dominant theories in this domain propose that a writer’s choice of demonstrative type is driven by the cognitive status that a referent is presumed to have in the mental model of the reader (e.g. Prince 1981; Ariel 1990; Gundel et al. 1993). The relative degree of activation of a discourse referent, or whether it is in focus or not, has indeed been suggested to distinguish a writer’s use of proximal versus distal forms (e.g. Sidner 1983; Ariel 1990; Gundel et al. 1993). Related proposals see different degrees of prominence reflected in proximal versus distal forms, and stress the freedom a writer has to assign a higher (proximal) or lower (distal) degree of prominence to a discourse referent through the demonstrative (e.g. Kirsner 1979; Strauss 2002).

In contrast with these traditional theories that focus on discourse-structural factors, Peeters et al. (2021) recently introduced a framework of demonstrative reference that proposes *discourse genre* as an essential predictor of a writer’s choice of endophoric demonstrative type. Specifically, it was suggested that the presumed relation between writer, addressee, and referent in the mental model of the writer will largely explain a writer’s choice of demonstrative in a text. They hence do not deem this presumed relation a result of local discourse considerations (e.g. variables resulting in a particular cognitive status of the referent, as in the aforementioned approaches), but rather consider it part of the global sociocultural knowledge writers possess about specific text or discourse genre. They specifically hypothesize a dominant preference for distal demonstratives in text genres where the role of the addressee is relatively more prominent, as in conversational and narrative discourse, and a dominant preference for proximal demonstratives when writers feel more responsible for produced discourse on topics they consider themselves knowledgeable about, as in expository texts (Peeters et al. 2021). Available studies or existing corpora that would allow for testing this recent proposal against traditional theories are, however, scarce.

The aim of this paper is therefore to advance our understanding of what factors drive a writer's choice of demonstrative by contrasting and testing the different theoretical views outlined above. Our focus is on demonstrative variance in written 'one-to-many' discourse, featuring writers and addressees who cannot rely on direct interactional feedback, nor on 'private' common ground. Typical examples are newspaper articles, wikipedia texts, and written product reviews.

In the remainder of this paper, we will first clarify our position with respect to the definition of demonstratives and classes of demonstratives that we will use (Section 2). We will then in more detail review existing theoretical proposals that identify various factors that are suggested to drive a writer's choice of endophoric demonstrative type (Section 3). In Section 4, we will evaluate existing corpus work and argue that empirical support in favor of earlier theoretical proposals has remained largely inconclusive. Next, we will discuss what specific predictions can be derived from the different theoretical views and how these can be tested using a corpus analysis (Section 5). In the method section we will then discuss the set-up and coding of the corpus that we used to contrast the different theoretical views (Section 6) and Section 7 reports the results of this corpus study. In Section 8 we will discuss our results in light of the existing literature, while Section 9 offers theoretical conclusions and an outlook.

2 Demonstratives and demonstrative classes

In line with recent theoretical work, we make a basic distinction between text-based and situation-based endophoric demonstratives in written discourse (cf. Maes et al. 2022; see **Figure 1** below). Text-based demonstratives take explicit linguistic elements as their primary interpretation cue. They include anaphoric and cataphoric demonstratives with nominal as well as non-nominal antecedents (the latter traditionally being termed 'discourse deictic', e.g., Himmelmann 1996; Diessel 1999). These text-based classes contain not only regular but also borderline cases with antecedent and anaphor having (slightly) different denotations or semantic interpretations, like nominal cases of bridging, deferred, or generic reference (e.g. Lücking 2018; Doran & Ward 2019), as well as non-nominal cases of referent coercion (e.g. Webber 1988; Kolhatkar et al. 2018). Text-based demonstratives further include first mention demonstratives, which introduce new referents on the basis of triggers provided in the demonstrative NP itself (as in recognitional demonstratives). Situation-based demonstratives, on the other hand, find their interpretation outside the written text itself, but in the writing situation, for example in the communicative situation (or *origo*) of the text (e.g. 'this week') or in the container of the text (e.g. 'this chapter'). As situation-based demonstratives are by definition proximal in English, and thus do not show any demonstrative variance, we do not empirically analyze them further in this paper.

Although these two main classes roughly represent what traditionally has been termed anaphoric vs. deictic respectively (e.g., Diessel 1999; Levinson 2004), we will not use these notions in this sense, as we consider all demonstratives (situation-based and text-based demonstratives,

demonstrative pronouns and demonstrative NPs) deictic (cf. Maes et al. 2022). As we see it, all text-based demonstratives need the deictic or marked force of a demonstrative (as compared to a regular pronoun or definite NP), either to guarantee successful referent identification, or to create additional inferences. For example, anaphoric demonstrative pronouns with nominal antecedents more often have a non-subject antecedent than regular pronouns (e.g. Brown-Schmidt et al. 2005; Kaiser & Trueswell 2008; Fossard et al. 2012; Çokal et al. 2014; 2018), and typically bring less accessible entities into focus (e.g. Linde 1979; Hauenschild 1982; Ariel 1990; Gundel et al. 1993; 2004). When they are used to access highly activated discourse referents, the markedness of demonstrative pronouns and determiners can be used for functions other than just referent identification. For example, determiners in a modified demonstrative NP are known to create additional inferences compared to definite determiners (e.g. ‘predicating’ demonstrative NPs, Maes & Noordman 1995; Schnedecker 2006; Doran & Ward 2019). Likewise, in some languages demonstrative pronouns that are used to refer to highly accessible human referents create a pejorative effect with respect to the referent (Sichel & Wiltschko 2021). This effect typically applies only when referent identification is guaranteed (and thus a regular pronoun could also have been used), and it is explained by the markedness of a demonstrative pronoun lacking the feature ‘person’, which is considered part of the content of personal pronouns in these languages.

In case of non-nominal antecedents and higher-order abstract referents, demonstrative pronouns are found to be used more frequently than personal pronouns (e.g. Maes 1997; Gundel et al. 2004; Kolhatkar et al. 2018) and in some languages, like Hebrew, demonstratives are required to access non-nominal antecedents (Sichel & Wiltschko 2021: 57). Demonstratives also enable access to new referents (e.g., recognitional *that* or indefinite *this*) more easily than definite NPs, and more easily allow for cataphoric relations than pronouns. Finally, they are also more powerful than regular pronouns in borderline cases in which mental representations of referents have to be created based on indirect cues: both nominal cues (e.g., in deferred/bridging or generic reference) and non-nominal cues (e.g., in cases of referent coercion).

3 Theoretical proposals on endophoric demonstrative variation

Broadly speaking, three different types of theories have been proposed to answer the question of what the main factors are that drive a writer’s choice of endophoric demonstrative. We will first discuss in Section 3.1 activation-based theories, which state that discourse-internal dynamics lead to referents with variable cognitive statuses, such that a referent’s presumed relative accessibility, givenness, or prominence drives demonstrative variation in text (e.g. Sidner 1983; Ariel 1990; Gundel et al. 1993). In Section 3.2, we will review existing theories that suggest that demonstrative variation is best explained by a set of subtle interactional and attitudinal inferences based on the referent’s assumed psychological proximity or distance from the writer (e.g. Fraser & Joly 1979; 1980; Chen 1990; Niimura & Hayashi 1994; Glover 2000; Jackson 2013). In Section

3.3, we will summarize our proposal that hypothesizes that the assumed interaction between writer, referent, and addressee in different discourse genres is the most important predictor of a writer's choice of endophoric demonstrative type (Peeters et al. 2021).

3.1 Discourse-internal factors: accessibility, givenness, prominence

In activation-based theories, proximal and distal demonstratives are considered part of a larger set of referential expression types that speakers and writers may use to activate or reactivate a discourse referent in the mind of one's addressee. Each type can be seen as expressing a particular degree of mental activation or prominence of the underlying referent. This has resulted in proposals claiming a difference in higher (for proximal demonstratives, indefinite *this* excluded) or lower (for distal demonstratives) degree of assumed activation of a discourse referent, hence considering the speaker's or writer's choice of demonstrative type as the straightforward outcome of given discourse-internal properties.

The two most influential theories in the domain of endophoric reference are Ariel's accessibility hierarchy (Ariel 1990) and the givenness hierarchy introduced by Gundel and colleagues (Gundel et al. 1993). In these theories, the type of referring expression used (e.g. *the apple* vs. *this apple* vs. *it*) is argued to correspond to the *cognitive status* that a referent is presumed to have in the mental model of the reader or listener. Demonstratives are thus interpreted as referring expressions that compete for production with alternative referring expressions (e.g. pronouns or definite NPs). Both the accessibility hierarchy and the givenness hierarchy assign demonstratives an intermediate cognitive status at a position between personal pronouns and definite noun phrases. The two hierarchies differ as to the cognitive status attributed within the closed set of demonstratives. The accessibility hierarchy (Ariel 1990) assumes that proximal demonstrative forms (compared to distal demonstrative forms) refer to relatively more accessible entities, and that demonstrative pronouns refer to entities that are more accessible than those referred to by demonstrative NPs. In the givenness hierarchy (Gundel et al. 1993), it is distal demonstrative NPs (*thatN* or *familiar that*, e.g. *that book*) that have a special status as they refer to entities that are currently less activated compared to entities referred to with proximal or distal demonstrative pronouns, or with proximal demonstrative NPs (*thisN*, e.g. *this book*).

Several other theoretical proposals use similar pragmatic notions to express the idea of proximal demonstratives accessing more prominent entities than distal ones. Some of them align well with the idea of demonstratives expressing different activation statuses of referents, in that proximal demonstratives are said to move an entity into the new focus of discourse, while distal demonstratives would refer to a currently non-central discourse entity (e.g. Linde 1979; Sidner 1983; McCarthy 2002; Swierzbis 2010). Related proposals associate proximal and distal demonstratives with high versus low focus or deictic force respectively (e.g. Kirsner 1979; Oh 2001; Strauss 2002; Wu 2004; Oh 2009), thereby considering observed variation in a writer's

choice of demonstrative type as a rhetorical tool enabling them to not only express “the force with which the hearer is instructed to seek the referent” (Strauss 2002: 135), but also present referents as relatively more or less important, noteworthy, or foregrounded (Kirsner 1979). Human, singular, and named entities (i.e. entities previously referred to using a proper name) would possess a high degree of prominence, making such referents ideal candidates for proximal demonstratives (Kirsner 1979). A strong asset of such proposals is that they are typically claimed to apply to all occurrences of text-based demonstratives.

3.2 Subtle interactional factors: the referent’s psychological distance from the writer

The *exophoric* use of demonstratives, in which demonstratives are used in reference to an aspect of the physical surroundings of a speech event, has commonly been considered the ontogenetic, phylogenetic, and grammatical basis from which other types of use (e.g. endophoric) have derived (e.g. Bühler 1934; Lyons 1977; Diessel 1999; Tomasello 2008). As such, influential theoretical descriptions of exophoric demonstrative use have implicitly been taken as a starting point for theories of endophoric demonstrative variance. Such exophoric theories have often taken a referent’s *relative* distance from the speaker (e.g. Diessel 1999) as the main explanandum of observed variation in demonstrative use. Not surprisingly then, the notion of distance has also been assumed to explain endophoric demonstrative variation in text. Because a referent’s physical distance is largely irrelevant in written discourse, distance is typically reconceptualized in a metaphorical sense. This has resulted in fine-grained analyses of demonstrative variation in terms of a referent’s ‘mental distance’ to the speaker or writer.

These accounts of demonstrative variance adhere to the view that demonstratives “do not necessarily or only guide the hearer to the intended referent, but may in some cases contribute to what is implicitly communicated as well” (Scott 2013: 56) such that they “can be used as a resource for fostering a sense of common ground and shared perspective between interlocutors” (Acton & Potts 2014: 4). As such, demonstratives can be seen as triggering different types of pragmatic inferences about the interaction between writer, addressee, and referent. Proximal demonstratives are seen as expressing the writer’s involvement and (mostly positive) attitude towards a referent, while distal demonstratives basically carry two types of inferences: either they represent the negation of the writer’s involvement and attitudes, often resulting in negatively framed inferences about referents, or they can be seen as playing a part in creating a common interactional space with the addressee.

The idea of proximal demonstratives expressing the (positive) involvement of the writer vs. distal demonstratives expressing the (negative) backside can be found in many proposals. For example, Wolter (2006) argues that all proximal demonstratives are subject to a proximity condition, while distal demonstratives are considered unmarked for proximity. She considers

proximity as physical nearness to the speaker, similar to other studies (Roberts 2002; Elbourne 2008), but her proposal extends to the endophoric usage of demonstratives, in that she reinterprets proximity as “speaker control over the identification of the referent” (Wolter 2006: 108). A large number of studies use similar but slightly different notions to express the same idea of psychological distance. Referents of proximal demonstratives are for instance argued to be inside the speaker’s sphere or personally involved, as opposed to outside the speaker’s sphere or subjectively dissociated (Cornish 2001; 2017), to be part of the present topic or concern as opposed to the past (Fraser & Joly 1980; Glover 2000), to represent interest, relevance, focusing, as opposed to rejection and distance (Fraser & Joly 1980; Chen 1990), to remain involved in the subject, as opposed to express distance from it (Lakoff 1974), to be found in the world of the speaker, not in the world shared by hearer and speaker (MacLaran 1980), to become increasingly situated in the subjective belief state of the speaker or in their attitude towards a corresponding proposition (Smith 1995), to be relatively close to the author’s own arguments and position, and positive, as opposed to less desirable and further removed from the author’s position in terms of distance (Zhang 2015), to be in the referential center, as opposed to at a neutral vantage point (Laczkó 2010), to be part of the speaker’s situation, as opposed to disengaged from the speaker (Danon-Boileau 1984), to be in the domain of the speaker’s direct experience (Niimura & Hayashi 1994; 1996), or to be associated with affection, interest, and pride, as opposed to contempt, disapproval, dislike, and mental remoteness (Petch-Tyson 2000).

When demonstratives are said to express inferences about the *interaction between writer and addressee*, distal demonstratives no longer play the bad cop. They are seen as an appeal or an invitation to the addressee (Lakoff 1974; Auer 1981), as coordinating the involvement of speakers and addressees (Cheshire 1996), as expressing the addressee’s (as opposed to the writer’s) responsibility in relation to the referent (Fraser & Joly 1980), as reflecting “consensual” (as opposed to “discordance”) deixis where the discourse common to both interlocutors contains the referent (Danon-Boileau 1984), or as “what you have just mentioned” versus “what I have just mentioned” (Halliday & Hasan 1976), the latter being congruent with the observation that in interactive communication, typically distal demonstratives are used to comment upon the remarks of another speaker (A: “*I am the best*”. B: “*That/#This is a lie*”, e.g. Lakoff 1974; Gundel et al. 1988; Chen 1990).

In sum, what these slightly different proposals have in common is that proximal and distal demonstrative forms are assumed to express subtle inferences about the relation of the referent vis-à-vis the speaker or writer and sometimes their intended addressee. It should be noted that these qualitative proposals referring to different instantiations of a referent’s psychological distance are typically based on the analysis of endophoric demonstratives in specific written or spoken contexts, and offer valid explanations for these examples, which however do not necessarily generalize to all demonstratives used in discourse. This raises the question to what extent the proposed underlying variables may explain endophoric demonstrative variation in

general. Furthermore, subtle inferences of psychological distance can be easily illustrated and substantiated in individual occurrences of endophoric demonstratives, but they are less easy to predict or put to an empirical test.

3.3 Discourse genre

As a theoretical alternative to the dominant activation- and prominence-based accounts, we recently introduced a framework of demonstrative reference that proposes *discourse genre* as the main predictor of endophoric demonstrative variation (Peeters et al. 2021). It stresses that writers have pre-existing assumptions on how writers, addressees, and referents interact in the context of specific discourse genres, in particular given that different discourse genres come with different discourse goals (e.g., to narrate, expose, or evaluate; e.g., Weaver & Kintsch 1991; Mar et al. 2021). These assumptions are based on the experiences we built up as writers and readers with regard to how discourse is shaped depending on genre goals. Under this account, demonstratives are considered as ways of expressing default interaction assumptions connected to particular discourse genres, thereby exploiting the typical values of different demonstrative variants in exophoric and endophoric context: speaker-proximity for proximal and addressee-orientation for distal demonstratives. Thus, it is proposed that an increasing preference for distal demonstrative anaphors is observed when the role of the addressee becomes more prominent in the discourse setting at hand (as in narrative discourse), while an increasing preference for proximal demonstrative anaphors is found when speakers or writers are focused more on the topic, as in an expository context (Peeters et al. 2021).

In existing studies on genre, discourse goals are typically not found to be explicitly connected to ‘specific assumptions on the interaction between writer, addressee, and referent’ in the way we conceptualize them here. Yet, interaction assumptions can easily be seen as a common denominator of the major characteristics differentiating genres. For example, in two recent review articles on two genres studied extensively in the learning sciences (Clinton et al. 2020; Mar et al. 2021), the beneficial effect of narrative over expository texts on comprehension and memory has been explained by alluding to typical characteristics of stories (over essays), such as topic familiarity, resemblance with everyday experiences, structure predictability, and the evocation of emotion. All these aspects can be considered addressee-oriented as they highlight the connection between writer and addressee(s). Conversely, expository discourse is typically characterized much more in terms of the relation between the writer and their topic, as expository texts commonly communicate the writer’s knowledge and ideas about a specific topic, for example by introducing and elaborating on a theme, thereby employing structures “depending on their purpose, making them less familiar and less predictable” (see Mar et al. 2021: 733,

and references therein). Lifelong experiences with such text characteristics can be considered as breeding ground for the assumed pre-existing assumptions argued for in our proposal.

Note that this theory combines two ideas on demonstrative variance widely acknowledged in existing literature. First, the importance of *referent proximity* in the choice between different (exophoric) demonstrative variants is not cancelled out in written discourse, but transferred into differences in the assumed psychological relation between speaker, referent, and addressee, as explained in Section 3.2. Second, the use of demonstrative variants “may differ systematically according to the social fields and genres in which speech occurs” (Hanks 2005: 194). The latter idea is often applied to broad genre notions such as spoken vs. written, and formal/higher vs. informal/lower discourse. For example, recognitional *that* is often linked to spoken discourse as the distal demonstrative is considered as a signal to the addressee that a given referential expression may be elaborated on if necessary (Auer 1981; Himmelmann 1996; Schlegloff 1996). In addition, restrictive *that* is typically considered characteristic of ‘higher’ registers (Wolter 2006), while distal demonstratives are often associated with conversational and informal registers (e.g., Cheshire 1996; Acton & Potts 2014). In addition, as we will see in Section 4, many corpus-based studies on demonstratives implicitly offer preliminary evidence for the potential relevance of genre in explaining demonstrative variance.

Demonstrative variants can be considered as perfectly equipped to express writer-addressee-referent assumptions in different genres. Yet, specific text properties, such as the presence of more or fewer proximal vs. distal demonstratives, do not in itself define genre. The connection between demonstratives, specific discourse goals, and genres is indeed defeasible. Put simply, expository texts are known to typically contain more complex words than narrative texts, but using complex words in itself does not make a text expository, only the higher-order goal does. The relation between discourse goals and text variables is hence fluid, as is indeed widely observed in the genre literature (Clinton et al. 2020; Mar et al. 2021). For example, for methodological reasons, studies comparing the effect of different genres may match text variables known to differentiate the two genres, such as the topic of the text (Cunningham & Gall 1990; Wolfe & Mienko 2007) or its complexity (Best et al. 2008; McNamara et al. 2011), thus showing that a text may remain expository or narrative in nature even if typical characteristics do not differ.¹ Likewise, readers are known to process the same text differently depending on the genre instructions (and thus genre expectations) they received (e.g. Zwaan 1994), suggesting that the same text can meet different discourse goals. One can even think of achieving a specific reading goal using the format of another genre, for example when stories are used to teach learners about a specific topic.

¹ Out of 37 studies reviewed in Mar et al. (2021), 8 control for topic/content and 15 for text complexity.

Therefore, in line with the proposal we put forward earlier (Peeters et al. 2021), we do not expect to find *only* proximal or distal demonstratives in expository or narrative text. Rather, we expect local discourse conditions to now and then overrule the genre default, in particular when there is special reason to do so, either based on referent identification, in line with the proposals described in Section 3.1, or based on specific inferences associated with referents or addressees, as described in Section 3.2. Yet, this theoretical account does predict genre to have the strongest effect, and thus claims to explain demonstrative variance for the bulk of demonstrative variants used in written discourse that can “be replaced by the other with very little effect on the meaning” (Stirling & Huddleston 2002: 1506). We consider the focus on this silent majority, while at the same time leaving enough room for types of demonstrative use that come with a strong proximal or distal preference, an appealing feature of this framework. Accordingly, an adequate test of the framework needs to exceed the level of individual clear case examples. A suitable method to do so is to analyze demonstratives in natural corpora of written text.

4 Corpus research into endophoric demonstrative variation

Theoretical proposals aiming to explain the main predictors of endophoric demonstrative variation, such as those sketched in Section 3, can be tested in an ecologically valid way using text corpora. Complementing approaches that rely on meta-linguistic judgments or experimental manipulations, a corpus analysis offers the opportunity to study the use of demonstratives by a large number of different writers in situations in which they naturally used demonstratives as part of a larger text and did not explicitly rely on their meta-linguistic intuitions. Here we are particularly interested in contrasting theories that claim that discourse-internal factors (e.g. the cognitive status or prominence of the referent) mainly explain demonstrative variation with the recent framework that proposes that discourse genre is the main predictor of a writer’s choice of demonstrative type.

Over the years, different types of corpora have indeed been analyzed for referential expressions in general (e.g. Toole 1996; Uryupina et al. 2020) and demonstratives in particular (e.g. Maes 1996; Byron & Allen 1998; Petch-Tyson 2000; Botley & McEnery 2001a; 2001b). Although these studies offer valuable distributional information, no existing corpus presents an in-depth analysis of demonstrative variance in a sizeable and balanced set of well-defined classes of demonstratives occurring in different genres. Early proposals often tested their claims on demonstrative variation in small-scale or genre-unspecific corpora (Kirsner 1979; Ariel 1988; Gundel et al. 1988; Himmelmann 1996; Oh 2001). Other studies restricted themselves either to demonstrative NPs (Maes 1996), proximal (Poesio & Modjeska 2005) or distal (Passonneau 1989; Byron & Allen 1998) demonstratives, used overlapping classes of demonstratives (Botley &

McEnery 2001b), or focused on only a single genre (Botley & McEnery 2001b; Potts & Schwarz 2010; Acton & Potts 2014).

To our knowledge and surprise, only a single previous corpus study directly tested the activation-based theories of endophoric demonstrative variation (Botley & McEnery 2001b). In a 100.000-word news corpus, the authors measured the referential distance between demonstrative and antecedent. When distance was measured in words, partial support for activation-based theories was found. It was also confirmed that pronominal demonstratives referred to antecedents at a smaller distance than demonstrative NPs did. When distance was measured in sentences, however, a pattern of results that was opposite to predictions made by activation-based theories was observed. Specifically, the referential distance for distal pronouns turned out to be smaller than for proximal pronouns, a finding that was also observed elsewhere (Gundel et al. 1988; Maes 1996). Based on these mixed outcomes, it remains unclear to what extent endophoric demonstrative variation is indeed explained by the degree of presumed activation or cognitive accessibility of the referent.

In contrast, several existing corpora do provide preliminary evidence for the proposal that discourse genre drives a writer's choice of demonstrative type, but do so in an indirect way. Corpora of interactional spoken discourse (Passonneau 1989; Byron & Allen 1998) and written news reports (Botley & McEnery 2001a) have for instance been shown to include substantially more distal than proximal demonstratives. Expository corpora have shown the opposite preference. For example, Poesio and Modjeska (2005) hardly find distal demonstratives in their corpora of museum object descriptions and medical leaflets. The relevance of genre is also suggested in cross-linguistic studies of demonstratives based on parallel corpora that compare translation patterns of demonstratives in original and translated literary (e.g. Wu 2004; Goethals 2007; 2013; Ribera & Cuenca 2013; Bartkute 2020) or non-literary (e.g. Vanderbauwhede et al. 2011; Pavesi 2013) text genres. Their corpora show large differences in the proportion of proximal and distal demonstratives in different genres and languages, which however remain largely unexplained as the focus of these studies is on discovering subtle differences in translation strategies and demonstrative systems in different languages, rather than explaining demonstrative variance itself.

Undoubtedly the largest relevant corpus of formal written discourse so far has been collected by researchers studying the use of English as a second language. Expository academic essays were gathered from students in a wide variety of different countries, after which their demonstrative use was compared to the use of demonstratives in similar essays that were written in the students' native language (e.g. Petch-Tyson 2000; Blagoeva 2004; Lenko-Szymanska 2004; Oh 2009; Labrador 2011; Zhang 2015; Jin 2019). In these corpora, an average of about 70% of all demonstratives is found to be proximal, a regularity also observed in the expository genre of scientific articles (Gray 2010).

In sum, it seems that expository texts elicit mainly proximal demonstratives, whereas narrative and interactional discourse - in which the role of the addressee is relatively more prominent - lead to a preference for distal demonstratives. A specific study analyzing demonstrative variation across different genres is however missing.

5 The current study

The current corpus study aims at identifying the variables responsible for the writer's choice between proximal and distal endophoric demonstratives in English. As our corpus, we use all text-based demonstratives ($n = 2232$) from the corpus selected and coded in Maes et al. (2022), see also **Figure 1**. This corpus consists of three different, common genres of written discourse: a corpus of written news reports, an encyclopedic wikipedia corpus, and a corpus of written product reviews. As such, it will allow for contrasting and testing predictions derived from the different theoretical viewpoints discussed above.

Accessibility theory makes clear predictions with regards to which factors should be considered proxies of a referent's degree of accessibility in a text corpus. Specifically, four variables are proposed to define the presumed differences in accessibility between pronominal (e.g. *this*) and nominal (e.g. *this book*) demonstratives: referential distance, saliency, competition, and unity (Ariel 1988). *Referential distance* measured in sentences or words is typically considered most (easily) applicable to corpus research (Ariel 1988; Botley & McEnery 2001b). One should consider distance measured in sentences rather than words as the most reliable proxy of a referent's presumed accessibility, as the structural (sentential or syntactic) position of antecedent and anaphoric NPs, rather than their absolute distance in words, is widely acknowledged to affect the activation of entities and the form of all types of referring expressions (e.g. Grosz et al. 1995; Kaiser & Trueswell 2008; Fukumura & van Gompel 2015). Moreover, demonstratives often have non-nominal antecedents that are located on their near left, but can at the same time be quite lengthy, which renders it problematic to decide what the distance in words between demonstrative and antecedent actually is.

The notion of *saliency* can be put on a par with the variables said to determine the focal value of referents. Two reliable and easily measurable variables present themselves here as syntactic proxy of the saliency of discourse relevance, namely the syntactic function (subject or not) and the sentence position (sentence-initial or not) of the demonstrative. These characteristics are indeed used in many experimental studies on the variation of referential expressions, in particular nominal versus pronominal expressions (e.g. Kaiser & Trueswell 2008; Fukumura & van Gompel 2015).

Competition is less relevant as a coding variable here, as endophoric demonstratives hardly occur in situations (typical for exophoric demonstrative use) in which there are two or more

competing referents (Maes 1996; Maes et al. 2022). *Unity* between referent and antecedent (i.e. whether these belong to the same unit, such as a paragraph) is also less relevant, as most endophoric demonstratives find their antecedent in the immediately preceding sentence (e.g. Passonneau 1989; Maes 1996).

In conclusion, on the basis of accessibility theory, it is predicted that proximal demonstratives refer more often to antecedents in the same sentence (vs. in a previous sentence) compared to distal demonstratives. In addition, proximal demonstratives should appear more often in subject and sentence-initial position compared to distal demonstratives. In addition, accessibility theory predicts pronominal demonstratives to have more often a subject and sentence-initial position and near antecedents than demonstrative NPs.

Similar to accessibility theory, distance, syntactic function, and sentence position can be used in the exact same way to test the givenness hierarchy (Gundel et al. 1993). Furthermore, in the givenness hierarchy, the special status of distal demonstrative NPs is based on a combination of two characteristics: their familiarity assumption and their ability to introduce new referents. Only the latter can be coded reliably in written corpora, such that significantly more first mention (familiar, recognitional) *that* than first mention (indefinite) *this* demonstratives should be observed to support the theory. Otherwise, the givenness hierarchy makes the same predictions as the accessibility hierarchy. Importantly, both the accessibility hierarchy and the givenness hierarchy do not make specific predictions for an effect of discourse genre.

Referent *prominence* may be tested by analyzing lexical characteristics of referents, as proposed by Kirsner (1979). He considers human, singular, and named entities as most prominent, and thus candidates for proximal demonstratives. Also, he would predict proximal demonstratives to more often come with new (as opposed to repeated) nouns, compared to their antecedents. Similarly, one may test for the influence of other semantic characteristics of referents on demonstrative variation. For example, Rocca and colleagues asked respondents to select a demonstrative for a variety of singular nouns, without any further context, and found that proximal demonstratives were more tightly associated with manipulable entities than distal ones (Rocca et al. 2019). In sum, although a large portion of demonstratives are by definition excluded from such analyses because they have non-nominal or abstract NP antecedents, a higher proportion of proximal, as opposed to distal demonstratives with human, singular, named, manipulable, or new noun entities consistent over genres would provide support for the effect of such referent characteristics.

Detecting subtle inferences based on *psychological distance* can best be done by inspecting selected individual examples in their context, as has been done extensively in the studies we discussed in Section 3.2. These studies convincingly show the relevance of such inferences in

demonstrative variance. Yet, to our knowledge, objective variables, enabling analysts to reliably code such inferences in large corpora, are not available. Therefore, we decided to focus on a corpus consisting of carefully selected genres and test whether distributional preferences of demonstratives show up that can be explained in terms of the attitudinal and interactional inferences ascribed to the attested examples in the endophoric proposals.

Finally, as outlined in Section 3.3, our recent framework of demonstrative reference hypothesizes strong effects of *discourse genre* in explaining demonstrative variation (Peeters et al. 2021). On the basis of this account, written news reports are expected to have a distal preference given their narrative and addressee-oriented character. Encyclopedic wikipedia entries are hypothesized to display an overall preference for proximal demonstratives, given the expository nature of these texts. Finally, evaluative product reviews are expected to show both aspects of an orientation towards the addressee and more writer-oriented attitudinal inferences. Therefore, for this specific corpus, a relatively balanced mix of proximal and distal demonstratives can be predicted.

6 Method

To contrast and test the different predictions outlined above, we selected all text-based demonstratives from the corpus described in Maes et al. (2022). These demonstratives come from three text genres: narrative (news articles), expository (wikipedia texts), and evaluative (book reviews) texts. The news texts consisted of the first 3000 paragraphs of the AQUAINT-2 Information-Retrieval Text Research Collection (Voorhees & Graff 2008), thus containing a total of 2021 Associated Press news articles on national and international news from the period 2004 to 2006. Wikipedia entries consisted of the complete GREC corpus, hence containing 1755 Wikipedia entries on persons, mountains, rivers, cities, and countries (Belz et al. 2010). The book reviews consisted of the first 3000 paragraphs of the Amazon product data corpus (He & McAuley 2016; <http://jmcauley.ucsd.edu/data/amazon/>), thus yielding 1904 reviews of a dozen books written in a relatively informal style by assumedly less professional writers. Using the Stanford CoreNLP system (<https://stanfordnlp.github.io>), in these corpora all sentences were automatically selected that contained at least one demonstrative, plus the three sentences that preceded the retrieved sentence (or as many preceding sentences as available when the retrieved sentence was the first, second, or third in the text). This resulted in a comparable number of text-based demonstratives for the three genres (News: $n = 825$; Wikipedia: $n = 609$; Reviews: $n = 798$).

In Maes et al. (2022), this corpus was used to develop a new taxonomy of endophoric demonstratives that distinguishes different types of text-based and situation-based demonstratives, as summarized in **Figure 1** and discussed in Section 2 above.

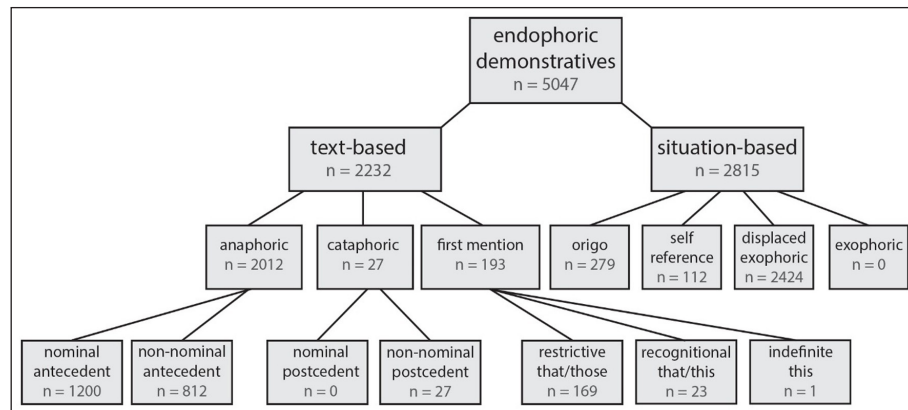


Figure 1: The taxonomy of endophoric demonstrative reference, as introduced by Maes et al. (2022), including the number of demonstratives observed per class in the corpus. For the current study, only text-based demonstratives were included, as situation-based demonstrative are, in English, all proximal by definition and thus do not show any demonstrative variation.

For the current study, we selected all text-based ($n = 2232$) demonstratives from this corpus, consisting of anaphoric ($n = 2012$), cataphoric ($n = 27$), and first mention ($n = 193$) demonstratives. We excluded situation-based demonstratives, such as references to the origo of discourse (e.g. ‘this week’) or self-references (e.g. ‘this chapter’), as they are proximal by definition, and thus lack any analyzable demonstrative variance.

For all text-based demonstratives ($n = 2232$), we coded all variables needed to test the theoretical proposals: *demonstrative type* (proximal *this/these* vs. distal *that/those*), *demonstrative number* (singular *this/that* vs. plural *these/those*), *demonstrative form* (pronominal vs. unmodified NP vs. modified NP vs. modified elliptic NP), *syntactic function* (subject vs. non-subject), *sentence position* (initial vs. non-initial), *type of referent* (abstract vs. concrete vs. human vs. named human). In addition, for both anaphoric and cataphoric demonstratives ($n = 2039$) we coded their *antecedent type* (NP vs. non-nominal) and *referential distance* [same sentence vs. previous sentence(s)]. Finally, for unmodified anaphoric NPs with a nominal antecedent ($n = 692$), we coded the *lexical relation* between the anaphor noun and the head noun of their antecedent (same vs. different noun). Descriptive results based on our coding are presented in **Table 1**.

Although the coding of the demonstratives across the mentioned variables are all clear and well-defined, about 20% of the data ($n = 467$) was separately coded by an independent second coder. For each of the coding variables, agreement between the two coders was between 96,4% and 98,7%. In total, 62 coding differences were found in 53 out of the 467 fragments (14 demonstrative form, 17 antecedent type; 14 distance, 11 syntactic function, 6 sentence position). Most of these ($n = 41$) were simple errors made by one of the coders and quickly resolved. The other cases were resolved after discussion, in particular concerning the exact extension of a non-nominal antecedent, or the difference between a nominal or non-nominal antecedent.

The fragments, first and second coding, and the dataset used in the analysis are available as online supplementary materials via the Open Science Framework (see Data Availability Statement below).

7 Results

7.1 Descriptive statistics

In **Table 1**, proportions of proximal and distal demonstratives are given for all coded variables. Overall, we observed a proximal preference for some variables previously related to activation and prominence, such as when demonstratives were singular (77,9% vs. 63,9%), in subject function (55,5% vs. 40,9%), or in sentence-initial position (55,0% vs. 44,3). Conversely, a distal preference was observed for plural referents, non-subject, and non-initial demonstratives. In addition, distal demonstratives were found to often have their antecedent in the same sentence (45,1% vs. 19,2%). But many of the observed proportions differed widely and wildly across variables, as shown in **Table 1**. Some of this variation seems intuitively plausible, such as the higher productivity of distal pronouns in the (more informal) reviews corpus, as also found in interactional corpora (e.g. Passonneau 1989; Byron & Allen 1998). In testing our hypotheses and discussing the results below, we will come back to some meaningful (combinations of) preferences numerically observed here.

7.2 Quantitative analysis

7.2.1 Overall Analysis

The descriptive statistics in **Table 1** numerically confirm the proposed demonstrative preferences in writer-oriented (expository) versus reader-oriented (narrative) discourse. In order to test the predictive power of genre in relation to local discourse-structural variables sensitive to the accessibility or prominence of referents (i.e. distance, syntactic function, and sentence position), we carried out a first, overall binary logistic regression analysis. The binary dependent variable in this analysis was the use of a proximal (*this*, *these*, coded as 0) or distal (*that*, *those*, coded as 1) demonstrative. Because for cataphoric and first mention demonstratives by definition not all variables could be coded (see above), only instances of anaphoric demonstrative use ($n = 2012$) were included in the analysis. In light of our theoretical predictions, we opted for a hierarchical regression approach to data analysis (forced entry), comparing a model (Model 1) that included three discourse-structural activation-based factors (referential distance, syntactic function, sentence position) to a model (Model 2) that additionally included genre (three levels: news, wiki, reviews) as a categorical predictor.

Table 2 presents the coefficients of the models. Model 1 explained significantly more variance in the data compared to a baseline, null model, $\chi^2(3) = 188.16, p < .001, R^2 = .09$ (Cox-Snell), .12 (Nagelkerke). As such, the three activation-based predictors together explained about 9%

	News n = 825; 37%		Wiki n = 609; 27,3%		Reviews n = 798; 35,7%		All n = 2232	
	prox	dist	prox	dist	prox	dist	prox	dist
text-based demonstratives % proximal vs. % distal	n = 166 20,1%	n = 659 79,9%	n = 474 77,8%	n = 135 22,2%	n = 322 40,4%	n = 476 59,6%	n = 962 43,1%	n = 1270 56,9%
demonstrative number	73,5 26,5	56,9 43,1	81,6 18,4	72,6 27,4	74,5 25,5	71,2 28,8	77,9 22,1	63,9 36,1
demonstrative form	27,1 44,0 27,1 1,8	29,7 41,7 4,1 24,5	27,0 61,4 11,6 0,0	10,4 43,7 3,7 42,2	38,5 40,1 20,5 0,9	45,2 25,6 15,3 13,9	30,9 51,2 17,3 0,6	33,5 35,9 8,3 22,3
syntactic function	50,0 50,0	48,3 51,7	64,3 35,7	14,8 85,2	45,3 54,7	38,0 62,0	55,5 44,5	40,9 59,1
sentence position	47,0 53,0	51,6 48,4	68,1 31,9	28,1 71,9	39,8 60,2	38,7 61,3	55,0 45,0	44,3 55,7
type of referent	91,6 4,8 2,4 1,2	75,3 5,2 19,3 0,2	78,1 21,7 0,2 0,0	74,8 11,9 12,6 0,7	82,6 4,7 1,6 11,2	77,7 9,7 11,3 1,3	81,9 13,1 1,0 4,0	76,1 7,6 15,7 0,6

(Contd.)

	News <i>n</i> = 825; 37%		Wiki <i>n</i> = 609; 27,3%		Reviews <i>n</i> = 798; 35,7%		All <i>n</i> = 2232	
%	<i>prox</i>	<i>dist</i>	<i>prox</i>	<i>dist</i>	<i>prox</i>	<i>dist</i>	<i>prox</i>	<i>dist</i>
<i>anaphoric - cataphoric demonstratives (n = 2012)</i>	<i>n</i> = 166	<i>n</i> = 583	<i>n</i> = 474	<i>n</i> = 122	<i>n</i> = 314	<i>n</i> = 380	<i>n</i> = 954	<i>n</i> = 1085
<i>antecedent type</i>	54,2 45,8	61,1 38,9	70,0 30,0	86,9 13,1	50,0 50,0	41,8 58,2	60,7 39,3	57,2 42,8
<i>referential distance</i>	29,5 70,5	36,2 63,8	13,7 86,3	71,3 28,7	22,0 78,0	50,3 49,7	19,2 80,8	45,1 54,9
<i>unmodified NP anaphors with nominal antecedent (n = 692)</i>	<i>n</i> = 49	<i>n</i> = 190	<i>n</i> = 233	<i>n</i> = 53	<i>n</i> = 84	<i>n</i> = 83	<i>n</i> = 366	<i>n</i> = 326
<i>lexical relation</i>	42,9 57,1	33,2 66,8	40,8 59,2	24,5 75,5	31,0 69,0	31,3 68,7	38,8 61,2	31,3 68,7
<i>anaphor-antecedent N</i>								

Table 1: Per corpus (News, Wiki, Reviews) and across the three corpora (All), (i) the proportion of proximal and distal text-based demonstratives is provided for the variables demonstrative number (singular vs. plural), form (pronouns, unmodified NPs, modified NPs, modified elliptic NPs), syntactic function (subject vs. non-subject), sentence position (initial vs. non-initial), and type of referent (abstract, concrete, human vs. named human); (ii) the proportion of proximal and distal anaphoric and cataphoric demonstratives is provided for the variables antecedent type (nominal vs. non-nominal) and referential distance (same sentence vs. earlier sentence); (iii) the proportion of proximal and distal unmodified NP anaphora with nominal antecedent is provided as a function of whether it has either a lexically same vs. different head noun.

(Cox-Snell) to 12% (Nagelkerke) of variance in the dependent variable. As shown in **Table 2**, all three predictors contributed significantly to the model. Model 2, in which genre was added as an additional predictor, explained significantly more variance in the data compared to Model 1, $\chi^2(5) = 679.45, p < .001, R^2 = .29$ (Cox-Snell), .38 (Nagelkerke). As such, adding genre to the model led to 20% (Cox-Snell) to 26% (Nagelkerke) of *additional* variance in the writers' choice of demonstrative type being explained. All predictors contributed significantly to the final model (all p 's < .003). Based on these two models, it can hence be concluded that all four predictors (distance, syntactic function, sentence position, genre) significantly contributed to explaining variation in demonstrative type in the data. Genre, however, clearly explained most variance.

When the same analysis was carried out on only demonstratives considered anaphoric in traditional taxonomies, that is, demonstrative anaphors with a nominal antecedent ($n = 1200$), thus excluding cases with a non-nominal antecedent (*discourse deixis*, $n = 812$), a highly similar pattern of results was observed. A model that included genre as an additional predictor [Model 2: $\chi^2(5) = 464.52, p < .001, R^2 = .32$ (Cox-Snell), .43 (Nagelkerke)] explained significantly more variance (18–24% more) in the data compared to a model that included only distance, syntactic function, and sentence position [Model 1: $\chi^2(3) = 179.23, p < .001, R^2 = .14$ (Cox-Snell), .19 (Nagelkerke)]. All predictors contributed significantly to the final model (all p 's < .008).

	b	Sig. (2-tailed)	Lower 95% CI	Odds Ratio	Upper 95% CI
Model 1					
Constant	-.32 [-.45, -.18]	.001			
Distance	1.34 [1.12, 1.55]	.001	3.07	3.81	4.73
Synt. Fun.	.44 [.18, .72]	.002	1.19	1.55	2.01
Sent. Pos.	-.32 [-.61, -.06]	.025	.56	.73	.95
Model 2					
Constant	-.34 [-.55, -.11]	.004			
Distance	1.55 [1.29, 1.81]	.001	3.66	4.70	6.03
Synt. Fun.	.68 [.41, .99]	.001	1.45	1.97	2.67
Sent. Pos.	-.55 [-.89, -.24]	.002	.42	.58	.79
Genre (1)	1.23 [.97, 1.51]	.001	2.66	3.41	4.37
Genre (2)	-1.65 [-1.90, -1.42]	.001	.15	.19	.25

Table 2: Logistic models of predictors of text-based anaphoric demonstrative variation (95% BCa bootstrap confidence intervals based on 1000 samples in brackets). Reference categories were previous sentence(s) (for Distance), subject (for Syntactic Function), initial (for Sentence Position), and reviews [for Genre (1) in the comparison to news reports and for Genre (2) in the comparison to wikipedia texts].

7.2.2 Demonstrative pronouns vs. demonstrative NPs

To test the hypothesis that referential distance should play a different role for demonstrative pronouns compared to demonstrative NPs, as proposed by Ariel (1988), we carried out separate binary logistic regression analyses for these two types of demonstrative use. These analyses were identical in setup to the overall analysis reported above, but carried out separately for demonstrative pronouns vs. demonstrative NPs.

The analysis on demonstrative pronouns ($n = 701$) showed that a model that included genre as an additional predictor [Model 2: $\chi^2(5) = 255.18$, $p < .001$, $R^2 = .31$ (Cox-Snell), .41 (Nagelkerke)] again explained significantly more variance in the data compared to a model that included distance, syntactic function, and sentence position [Model 1: $\chi^2(3) = 27.81$, $p < .001$, $R^2 = .04$ (Cox-Snell), .05 (Nagelkerke)]. Indeed, including genre as a predictor led to 27% (Cox-Snell) to 36% (Nagelkerke) of additional variance in writers' choice of demonstrative pronoun form being explained compared to the simpler model. A similar pattern of results was found when this analysis was carried out on only the demonstrative pronouns with a nominal antecedent ($n = 210$). Indeed, including genre as a predictor significantly improved the model and led to 35–47% of additional variance explained in the dependent variable by genre, also in this smaller dataset. In all models, distance was also a significant predictor.

The analysis on demonstrative NPs ($n = 1311$) yielded similar results, both for the important role of genre and the smaller but significant effect of distance. The model that included genre as an additional predictor [Model 2: $\chi^2(5) = 476.10$, $p < .001$, $R^2 = .31$ (Cox-Snell), .41 (Nagelkerke)] thus explained significantly more variance in the data compared to a model that included only distance, syntactic function, and sentence position [Model 1: $\chi^2(3) = 190.15$, $p < .001$, $R^2 = .14$ (Cox-Snell), .18 (Nagelkerke)]. Hence, including genre as a predictor led to 17% (Cox-Snell) to 23% (Nagelkerke) of additional variance in the writers' choice of demonstrative NP form being explained. A similar pattern of results was found when the same analysis was carried out on only the demonstrative NPs with a nominal antecedent ($n = 990$). Indeed, including genre as a predictor significantly improved the model and led to 15–20% of additional variance explained in the dependent variable by genre, also in this smaller dataset, and also here distance was a significant predictor of demonstrative variance.

Table 3 shows the contributions of the individual predictors to the model that showed the best fit (Model 2), presented separately for the analyses on demonstrative pronouns vs. demonstrative NPs, regardless of type of antecedent. In all models, distance was hence a significant predictor of demonstrative variance. Note that sentence position ($p = .076$) did not contribute significantly to the model that explained most variance in the writer's choice of demonstrative NP.

	<i>b</i>	Sig. (2-tailed)	Lower 95% CI	Odds Ratio	Upper 95% CI
Dem Pronouns					
Constant	.32 [.04, .66]	.032			
Distance	.98 [.55, 1.50]	.001	1.73	2.68	4.14
Synt. Fun.	.95 [.31, 1.59]	.002	1.44	2.60	4.67
Sent. Pos.	-.88 [-1.50, -.30]	.004	.23	.42	.75
Genre (1)	1.04 [.61, 1.49]	.001	1.84	2.83	4.35
Genre (2)	-3.00 [-3.76, -2.44]	.001	.03	.05	.10
Dem NPs					
Constant	-1.06 [-1.39, -.75]	.001			
Distance	1.78 [1.46, 2.13]	.001	4.35	5.93	8.10
Synt. Fun.	.73 [.39, 1.09]	.002	1.42	2.07	3.02
Sent. Pos.	-.33 [-.70, .02]	.076	.49	.72	1.05
Genre (1)	1.58 [1.26, 1.97]	.001	3.51	4.87	6.76
Genre (2)	-.99 [-1.33, -.69]	.001	.27	.37	.52

Table 3: Logistic models (final) of predictors of text-based anaphoric demonstrative variation (95% BCa bootstrap confidence intervals based on 1000 samples in brackets). Reference categories were previous sentence(s) (for Distance), subject (for Syntactic Function), initial (for Sentence Position), and reviews [for Genre (1) in the comparison to news reports and for Genre (2) in the comparison to wikipedia texts]. Separate analyses were carried out for demonstrative pronouns and demonstrative NPs.

7.3 Additional analyses

7.3.1 Text-based demonstratives and referent introduction

Given the lower activation status of familiar *thatN* in the givenness hierarchy (Gundel et al. 1993), one would hypothesize more distal than proximal demonstrative NPs when these are used to introduce a new referent in discourse. In the three corpora, we found 193 first mention demonstratives, distributed over three classes: recognitional demonstratives ($n = 23$), indefinite *this* ($n = 1$) and restrictive *those/that* demonstratives ($n = 169$), suggesting support for this hypothesis. Restrictive *those/that* demonstratives were all distal by definition. They were all elliptic, consisting of a (verbal) post-modification only, as in Example 1 below. Recognitional demonstratives, as in Example 2, were based on generic familiarity rather than private common ground (as in spoken discourse). They were all observed in the reviews corpus, found to be predominantly distal (66,6%), representing typical cases of recognitional or familiar *thatN* (Gundel et al 1993; Himmelmann 1996; Cornish 2001). The remaining cases were proximal, and

relied on similar familiarity inferences (mainly based on knowledge of the book reviewed), but could not be substituted by an indefinite article.

- (1) For **those** who don't know Gibran, get to know his work. (R3332)²
- (2) But, you can tell why the hippie-set loved this book: it is spirituality devoid of religion. Namely, this book can make you feel all “cosmic” without all **that** pesky Christian morality. (R3421)

Finally, as mentioned in Section 2, we classified all demonstratives for which a linguistic trigger could be found as anaphoric, cataphoric, or first mention, depending on the location of the trigger (respectively before or after the demonstrative or after but in the same NP position of the demonstrative). This caused these classes to cover not only a majority of regular cases, but also a minority of borderline cases with slight changes in the denotation or semantic interpretation of the referent. Although the coding variables did not allow for a detailed analysis of borderline cases, some classes could easily be detected, and showed a distal preference, thus supporting the idea of distal demonstratives being associated more with (slightly) new referents. For example, the bridging inference demonstratives found in the corpus preferred a distal demonstrative (12 out of 16), as observed elsewhere (Lücking 2018). We also found a productive class of deferred elliptic those/that anaphors ($n = 156$), as in Example 3, where the demonstrative picks out the head noun of the antecedent NP (Cubans) to create a new entity (wet foot Cubans).

- (3) Under the government's wet foot/dry foot policy, Cubans who set foot on U.S. soil are generally allowed to stay, while **those** intercepted at sea are usually returned to Cuba. (N623)

7.3.2 Text-based demonstratives and referent prominence

The idea of demonstratives expressing not so much the discourse structural position of the referent, but rather the writer's strategy to present a referent as *prominent* or *noteworthy*, can be supported by a preference for proximal demonstratives in reference to singular, human, and named entities (Kirsner 1979). Our data in **Table 1** confirm the overall preference for proximal singular demonstratives, but show a mixed picture for human referents. The number of observed human ($n = 209$) and named human ($n = 46$) demonstratives was relatively small. The majority of human referents ($n = 160$) consisted of restrictive *those* demonstratives, which were by

² In the examples presented in the text, we have underlined the antecedent, highlighted the critical demonstrative in boldface, and added the source (News, Wikipedia, Reviews) and the corpus ID number, referring to the coding used in the corpus presented by Maes et al. (2022).

definition all distal. Also the remaining cases were mostly distal ($n = 39$ out of 49). Named referents, on the other hand, were observed to predominantly elicit a proximal demonstrative ($n = 38$ out of 46). Nevertheless, many of them referred to protagonists in the book reviews from the review corpus, and had a displaced exophoric flavor. Furthermore, Kirsner (1979) suggested that demonstratives followed by a new noun, compared to a repetition of the antecedent noun, should be more prominent and thus proximal. The bottom row of **Table 1**, however, shows a more pronounced preference for distal NPs in these cases in our analysis.

Finally, our data do not allow us to test the effect of more fine-grained semantic features of lexical items, as has been done by Rocca and colleagues (2019). There is only one class of demonstratives in the text fragments we analyzed that resembles Rocca et al.'s experimental situation of a demonstrative that is linguistically and mentally connected to one singular lexical item only, namely demonstratives involved in a relationship between an unmodified anaphor noun and the same unmodified antecedent noun. All other demonstratives are related to a combination of lexical items: first mention demonstratives are by definition modified, thus including a combination of items (e.g., *that pesky + Christian + morality* R3421). The same holds for all anaphoric demonstratives: demonstrative NPs with a non-nominal antecedent connect an abstract anaphor noun (e.g. *this fact* W2755; *that case* N33; *this shift* N169, etc.) with a non-nominal antecedent; modified demonstratives combine the semantics of head noun and modifiers (e.g. *this/that poor + animal* R5480); anaphor nouns with different noun antecedents induce synonymous or (most often) hyperonymic relations between these nouns (e.g. *trail* <— *this route* W2752; *Motorola Inc.* <— *this company* N284; *October* <— *that month* N96). The only category applicable here is nearly empty: unmodified anaphor nouns with the same antecedent noun turn out to have modified antecedents, inducing categorical relations between anaphor and antecedent (e.g. *dry climate* <— *this climate* W2698; *a colossal eruption around 1750 BC* <— *this eruption* W3085; *a contemporary man in his 90's who lives in a nursing home* <— *this man* R4377). Only one exceptional demonstrative, shown in Example 4, showed a syntactic and mental connection between only one singular lexical element (a proper name) and a demonstrative:

(4) I'm proud to live in France, but **this** France disappoints me. (N618)

In other words, all (but one) anaphoric demonstratives in the corpus are for their interpretation dependent on different lexical items and combinations thereof, which raises the question: which of these should we consider the source of the referent-intrinsic features expected to have an effect on demonstrative choice? Based on our findings, it seems that any hypothesized relation between variation in type of demonstrative used and a referent's fine-grained semantic properties is neither applicable nor relevant to endophoric demonstratives.

8 Discussion

The corpus study reported in this paper confirms the importance of discourse genre in explaining variation in writers' choice of endophoric demonstratives. As predicted by our recent theoretical framework (Peeters et al. 2021), discourse genre explained significantly more demonstrative variation in the wide variety of text fragments we analyzed than local discourse variables related to the cognitive status or prominence of the referent. In different discourse genres, the position of the writer differs vis-à-vis the described entities and the addressee. The substantial variation we observed in the overall use of proximal vs. distal demonstratives across the different corpora shows that writers may implicitly make use of sociocultural genre knowledge when writing a text, as evident from the demonstratives they use in expressions of endophoric reference. As such, a clear preference for distal demonstratives is found when the addressee is considered more prominent in the given discourse setting (as in news reports), whereas an overall preference for proximal demonstratives is observed when writers feel more responsibility for the produced discourse themselves, as in an expository context (e.g. wikipedia texts). In such expository contexts, proximal demonstratives hence indicate that the referent is psychologically situated near the writer, whereas in interactional and narrative discourse the writer uses distal demonstratives to reach out to the addressee (Peeters et al. 2021). Thus, these demonstrative preferences are derived from genre specific assumptions on the interaction between writer, referent, and addressee. These assumptions in turn cover the typical characteristics of text goals and genres, which we experience in everyday life and gradually internalize as socio-cultural genre knowledge (Clinton et al. 2020; Mar et al. 2021). At the same time, the genre default is defeasible and thus leaves room for other variables and explanations to play their part. In the remainder of this section, we will discuss to what extent our quantitative and qualitative observations offer support for the idea that earlier theories of endophoric demonstrative reference can be seen as providing reasons for endophoric demonstratives to deviate from pre-existing genre defaults.

8.1 Demonstrative variance and activation-based proposals

The quantitative analyses presented above show that activation-based variables (referential distance, syntactic function, sentence position) explained a significant part of variance in the writers' choice of demonstrative type across the different corpora, though clearly smaller than genre. Some of the results can be seen as support for the idea that demonstratives are sensitive to the activation level of referents, in particular the observation that proximal demonstratives more often have a subject function and are found in sentence-initial position than distal demonstratives (subject: 55,5% vs. 40,9%; sentence-initial: 55% vs. 44,3%). When we take these variables as being associated with highly accessible or given referents, their distribution supports the two activation-

based theories. We also found more first-mention *familiar thatN* (i.e. distal demonstrative NPs) than first-mention proximal demonstrative NPs, as predicted by the givenness hierarchy.

Other activation-based predictions, however, were not supported by our results. Referential distance significantly explained some variance in the data, but we found more distal than proximal demonstratives with antecedents in the same sentence (vs. in a previous sentence; 45,2% vs. 19,2%), while accessibility theory would predict the opposite.

Accessibility theory furthermore predicted that pronominal demonstratives should occur more often in subject and initial position and near their antecedents compared to demonstrative NPs. Our quantitative results showed significant, but again opposite, effects of syntactic function (subject vs. non-subject) in separate analyses for demonstrative pronouns and demonstrative NPs. Text-based demonstrative pronouns were indeed present in subject position ($n = 500$) more often than in non-subject position ($n = 201$), whereas for demonstrative NPs the opposite pattern was observed ($n = 511$ vs. $n = 800$ respectively). A demonstrative's sentence position (initial vs. non-initial) only significantly explained variance in the observed demonstrative type (proximal vs. distal) for demonstrative pronouns, and not for demonstrative NPs. Demonstrative pronouns were indeed found more often in sentence-initial ($n = 495$) than in sentence non-initial position ($n = 206$).

8.2 Demonstrative variance and referent prominence

Earlier work on endophoric demonstrative variation has hypothesized that the *prominence* of a particular referent may influence a writer's choice of demonstrative type. Specifically, it was proposed that human, singular, and named entities would by definition possess a high degree of prominence, making such referents ideal candidates for proximal demonstratives (Kirsner 1979). In addition, proximal demonstratives have been argued to more often come with new (as opposed to repeated) nouns, when compared to the critical noun present in their antecedents (Kirsner 1979). Our data in **Table 1** do not reveal a stable connection across corpora between proximal demonstratives and the discourse-structural prominence of referents (i.e. their subject and sentence-initial position) or the intrinsic prominence features of referents (i.e. human and named human referents). In addition, we did not find more proximal demonstratives with new nouns compared to with repeated nouns. As most of these anaphor nouns are attenuated classifiers of the antecedent noun (e.g. *October* <— *that month* N96), we consider them 'different' rather than 'new', and thus unreliable as a measure of referent prominence.

Yet, some observations may support the idea of proximal demonstratives being associated with more prominent referents. First, as for activation-based proposals, the higher proportion of proximal demonstratives in subject function and sentence-initial position can also suggest a higher association of proximal demonstratives with prominent referents. Second, we also

observed a preference for proximal (vs. distal) modified NPs across the three corpora. Apparently, when writers decide to predicate information on a well-established referent using a modified demonstrative NP, they more often use a proximal than a distal demonstrative (17,3% vs. 8,3%). This preference can be explained as a strategic choice of the writer to present new information on an existing discourse referent preferably from their own perspective, thus expressing the same idea of interactional inferences as proposed in Peeters et al. (2021), which may be suggested by local discourse conditions and not necessarily by pre-existing genre assumptions. This observation is in line with the idea of genre preferences being defeasible and thus being able to be overruled when certain conditions apply. In this case, it is plausible to assume that a writer uses a proximal demonstrative to render a particular referent more prominent.

Finally, the text-based demonstratives in our corpus all showed semantic relations exceeding the simple relation between one demonstrative and the semantics of one lexical item. Therefore, we were not able to verify the effect of any other intrinsic semantic feature of lexical items or referents. We deem it unlikely that our data are special in this respect, and thus cast doubt on whether it is possible to find or even define effects of intrinsic, semantic features of referents in endophoric demonstratives. This hence questions the generalizability to natural multi-word texts of observed referent-intrinsic, semantic effects on a writer's demonstrative choice, as observed in experimental settings where nouns are presented in isolation (e.g. Rocca et al. 2019; Rocca & Wallentin 2020).

9 Conclusion

In conclusion, the data we have presented in this paper support the idea of socio-cultural genre knowledge being the main driver of demonstrative variance in endophoric demonstratives. Text genres can be seen as carrying a default assumed psychological distance between writer and referents, as well as an assumed interactional relationship between writer and addressee. The literature shows many examples with clear proximal or distal preference that can be and have been explained in this vein. The main contribution of this paper is that we have shown similar overall preferences in different text genres for the large majority of demonstratives that can easily be replaced by the other variant.

We observed partial support for the validity of theoretical proposals based on the local discourse context, motivated either by general discourse mechanisms (such as the degree of activation of referents) or by incidental strategies used by writers, e.g. to highlight new information on an activated referent or to introduce new referents. As we see it, we consider these as reasons to overrule the default preference set by the genre. Yet, the effects of the variables associated with the general mechanism of referent activation were small (in the case of subject function and sentence-initial position) and sometimes even showed the opposite pattern compared to earlier

theoretical expectations (in the case of referential distance). The distribution of some other variables differed more substantially, such as the preference for proximal modified anaphoric NPs, and for distal NPs referring to new or slightly new referents, thus supporting the idea of writers incidentally using a writer-near demonstrative to predicate new information about an established referent, and a reader-near demonstrative to introduce a new referent or signal a change to an existing referent. Importantly, the latter preferences can perfectly be explained in terms of the interaction between writer, referent, and addressee as proposed in Peeters et al. (2021), this time not as a pre-existing genre assumption, but as suggested by local context: writers prefer to present new information on established referents from their own perspective, while they prefer to appeal to the addressee when new referents have to be created or when the representation of activated referents has to be adapted.

On a more general note, studies on endophoric demonstratives have indeed shown ample evidence that writers can use demonstrative variation freely to suggest a variety of subtle pragmatic inferences, such as when suggesting prominence of referents, positive or negative attitudes towards referents, or invitations or appeals to the addressee. For all of these inferences, translated in a variety of concepts as listed in Section 3.2, we consider the presumed psychological distance and the presumed relation with referent and addressee in the mind of the writer as the common denominator best explaining demonstrative variance in text. By using the method of an exhaustive analysis of a large corpus representing clear discourse genres, we were able to capture the effect of higher order sociocultural knowledge of discourse goals and genres. Our analysis does not invalidate the use of individual examples as a basis for theoretical proposals of demonstrative variance, but presents a more generalizable way to distinguish between local discourse conditions and higher order considerations in the mind of the writer as an explanatory basis of demonstrative variance. Likewise, it does not intend to discourage the use of experimental methods to elicit demonstratives or acceptability ratings from naïve participants in the lab or in online studies. However, it does suggest the importance of manipulating, or at least taking into account, genre and goal information in such experimental tasks, to be able to capture not only the effect of local discourse variables, but also of the higher order default demonstrative preferences connected to different discourse genres.

Given the subtlety of these higher order pragmatic inferences, we assume that cognitive abilities and stylistic, rhetorical skills of individual writers must lead to substantial variation in their choice of demonstrative type (Peeters et al. 2021). As we selected corpora produced by many different writers, we expect however genre preference to be stronger than individual differences, and indeed hope to confirm such a result in future experiments in which respondents have to choose a proximal or distal demonstrative in fragments clearly coming from more speaker- or addressee-oriented text genres.

Finally, we note that the evidence presented in this paper is restricted to endophoric demonstratives in English. Specific typicalities, such as degree modifiers or cases of indefinite *this*, are likely to have limited scope across the languages of the world. Furthermore, it remains unclear whether and to what extent our dichotomous (*this* vs. *that*) view on endophoric variance is applicable to languages with a richer and more hybrid demonstrative system. For the major assumptions made in this paper, however, we assume a broad generalizability across languages. Indeed, as evident from the wide range of earlier work discussed in this paper, fundamental notions such as psychological distance, implicit genre knowledge, and the assumed interaction between the writer and their addressee should play a role in the minds of writers around the world.

Data availability statement

Data associated with this paper can be found on the Open Science Framework via https://osf.io/2y6ce/?view_only=3c14c34e91f046de9a82c26c94ad13f2

- Supplementary file 1: Corpus_Coding_SecondCoding. The corpus of text fragments used in the current study, the coding of all demonstratives including the code book used, the subset of fragments coded by the second coder (“recoding”), and the outcome of the comparison between the coding of the first and second coder.
- Supplementary file 2: Dataset_Quantitative_Analysis. The dataset used in the quantitative analyses reported above.

Acknowledgements

The authors thank three anonymous reviewers for valuable suggestions on an earlier draft of this work, and Thiago Castro Ferreira and Emiel van Miltenburg for help in the construction of the corpora used.

Funding information

D.P. was supported by a Veni grant (275-89-037) awarded by *De Nederlandse Organisatie voor Wetenschappelijk Onderzoek* (NWO, the Dutch Research Council).

Competing interests

The authors have no competing interests to declare.

References

- Acton, Eric K. & Potts, Christopher. 2014. That straight talk: Sarah Palin and the sociolinguistics of demonstratives. *Journal of Sociolinguistics* 18(1). 3–31. DOI: <https://doi.org/10.1111/josl.12062>
- Ariel, Mira. 1988. Referring and accessibility. *Journal of Linguistics* 24(1). 65–87. DOI: <https://doi.org/10.1017/S0022226700011567>
- Ariel, Mira. 1990. *Accessing antecedents*. Routledge.
- Auer, Peter. 1981. Zur indexikalitätsmarkierenden Funktion der demonstrativen Artikelform in deutschen Konversationen. In Hindelang, Götz & Zillig, Werner (eds.), *Sprache: Verstehen und Handeln*, 301–311. Max Niemeyer Verlag. DOI: <https://doi.org/10.1515/9783111611877.301>
- Bartkute, Darija. 2020. Spatial deictics and translational implicitation: evidence from a corpus-based analysis of english and lithuanian fictional discourse. *International Journal of Language and Linguistics* 8(5). 229–239. DOI: <https://doi.org/10.11648/j.ijll.20200805.15>
- Belz, Anja & Kow, Eric & Viethen, Jette & Gatt, Albert. 2010. Generating referring expressions in context: the GREC task evaluation challenges. In Kraemer, Emiel & Theune, Mariet (eds.),

- Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation*, 294–327. Springer. DOI: https://doi.org/10.1007/978-3-642-15573-4_15
- Best, Rachel M. & Floyd, Randy G. & McNamara, Danielle S. 2008. Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology* 29(2). 137–164. DOI: <https://doi.org/10.1080/02702710801963951>
- Blagoeva, Roumiana. 2004. Demonstrative reference as a cohesive device in advanced learner writing: a corpus-based study. *Advances in Corpus Linguistics* 49. 297–307. DOI: https://doi.org/10.1163/9789004333710_018
- Botley, Simon & McEnery, Tony. 2001a. Demonstratives in english: a corpus-based study. *Journal of English Linguistics* 29(1). 7–33. DOI: <https://doi.org/10.1177/00754240122005170>
- Botley, Simon & McEnery, Tony. 2001b. Proximal and distal demonstratives: a corpus-based study. *Journal of English Linguistics* 29(3). 214–233. DOI: <https://doi.org/10.1177/00754240122005341>
- Brown-Schmidt, Sarah & Byron, Donna K. & Tanenhaus, Michael K. 2005. Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language* 53(2). 292–313. DOI: <https://doi.org/10.1016/j.jml.2005.03.003>
- Bühler, Karl. 1934. *Sprachtheorie*. Fischer.
- Byron, Donna & Allen, James. 1998. Resolving demonstrative anaphora in the TRAINS93 corpus. In Botley, Simon & McEnery, Tony (eds.), *New Approaches to Discourse Anaphora: Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*, 68–81.
- Chen, Rong. 1990. English demonstratives: A case of semantic expansion. *Language Sciences* 12(2). 139–153. DOI: [https://doi.org/10.1016/0388-0001\(90\)90009-6](https://doi.org/10.1016/0388-0001(90)90009-6)
- Cheshire, Jenny. 1996. That jacksprat: An interactional perspective on English that. *Journal of Pragmatics* 25(3). 369–393. DOI: [https://doi.org/10.1016/0378-2166\(95\)00032-1](https://doi.org/10.1016/0378-2166(95)00032-1)
- Clinton, Virginia & Taylor, Terrill & Bajpayee, Surjya & Davison, Mark L. & Carlson, Sarah E. & Seipel, Ben. 2020. Inferential comprehension differences between narrative and expository texts: a systematic review and meta-analysis. *Reading and Writing* 33(9). 2223–2248. DOI: <https://doi.org/10.1007/s11145-020-10044-2>
- Çokal, Derya & Sturt, Patrick & Ferreira, Fernanda. 2014. Deixis: *this* and *that* in written narrative discourse. *Discourse Processes* 51(3). 201–229. DOI: <https://doi.org/10.1080/0163853X.2013.866484>
- Çokal, Derya & Sturt, Patrick & Ferreira, Fernanda. 2018. Processing of *it* and *this* in written narrative discourse. *Discourse Processes* 55(3). 272–289. DOI: <https://doi.org/10.1080/0163853X.2016.1236231>
- Cornish, Francis. 2001. 'Modal' *that* as determiner and pronoun: the primacy of the cognitive-interactive dimension. *English Language and Linguistics* 5(2). 297–315. DOI: <https://doi.org/10.1017/S1360674301000259>
- Danon-Boileau, Laurent. 1984. That is the question. In Grésillon, Almuth & Lebrave, Jean-Louis (eds.), *La langue au ras du texte*, 31–55. Presses Universitaires de Lille.

- Diessel, Holger. 1999. *Demonstratives: Form, function and grammaticalization*. John Benjamins Publishing. DOI: <https://doi.org/10.1075/tsl.42>
- Doran, Ryan B. & Ward, Gregory. 2019. A taxonomy of uses of demonstratives. In Gundel, Jeanette & Abbott, Barbara (eds.), *The Oxford Handbook of Reference*, 236–259. Oxford University Press. DOI: <https://doi.org/10.1093/oxfordhb/9780199687305.013.12>
- Elbourne, Paul. 2008. Demonstratives as individual concepts. *Linguistics and Philosophy* 31(4). 409–466. DOI: <https://doi.org/10.1007/s10988-008-9043-0>
- Fossard, Marion & Garnham, Alan & Cowles, H. Wind. 2012. Between anaphora and deixis: The resolution of the demonstrative noun phrase “that N”. *Language and Cognitive Processes* 27(9). 1385–1404. DOI: <https://doi.org/10.1080/01690965.2011.606668>
- Fraser, Thomas & Joly, André. 1979. Le système de la deixis. Esquisse d’une théorie d’expression en anglais. *Modèles Linguistiques* 1(2). 97–157.
- Fraser, Thomas & Joly, André. 1980. Le système de la deixis (2). Esquisse d’une théorie d’expression en anglais. *Modèles Linguistiques* 2. 22–49.
- Fukumura, Kumiko & van Gompel, Roger P.G. 2015. Effects of order of mention and grammatical role on anaphor resolution. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 41(2). 501–525. DOI: <https://doi.org/10.1037/xlm0000041>
- Glover, Kelly D. 2000. Proximal and distal deixis in negotiation talk. *Journal of Pragmatics* 32(7). 915–926. DOI: [https://doi.org/10.1016/S0378-2166\(99\)00078-8](https://doi.org/10.1016/S0378-2166(99)00078-8)
- Goethals, Patrick. 2007. Corpus-Driven hypothesis generation in translation studies, contrastive linguistics and text linguistics: a case study of demonstratives in spanish and dutch parallel texts. *Belgian Journal of Linguistics* 21(1). 87–103. DOI: <https://doi.org/10.1075/bjl.21.07goe>
- Goethals, Patrick. 2013. Demonstratives on the move: what translational shifts tell us about demonstrative determiners and definite articles in Spanish and Dutch. *Linguistics* 51(3). 517–553. DOI: <https://doi.org/10.1515/ling-2013-0019>
- Gray, Bethany. 2010. On the use of demonstrative pronouns and determiners as cohesive devices: A focus on sentence-initial this/these in academic prose. *Journal of English for Academic Purposes* 9(3). 167–183. DOI: <https://doi.org/10.1016/j.jeap.2009.11.003>
- Grosz, Barbara J. & Weinstein, Scott & Joshi, Aravind K. 1995. Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2). 203–225. DOI: <https://doi.org/10.21236/ADA324949>
- Gundel, Jeanette K. & Hedberg, Nancy & Zacharski, Ron. 1988. On the generation and interpretation of demonstrative expressions. In Vargha, Dénes & Hajičová, Eva (eds.), *Proceedings of the 12th Conference on Computational Linguistics*, 216–221. Association for Computational Linguistics. DOI: <https://doi.org/10.3115/991635.991679>
- Gundel, Jeanette K. & Hedberg, Nancy & Zacharski, Ron. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2). 274–307. DOI: <https://doi.org/10.2307/416535>

- Gundel, Jeanette K. & Hedberg, Nancy & Zacharski, Ron. 2004. Demonstrative pronouns in natural discourse. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, 81–86.
- Halliday, Michael A. K. & Hasan, Ruqaiya. 1976. *Cohesion in English*. Routledge.
- Hanks, William F. 2005. Explorations in the Deictic Field. *Current Anthropology* 46(2). 191–220. DOI: <https://doi.org/10.1086/427120>
- Hauenschild, Christa. 1982. Demonstrative pronouns in Russian and Czech - deixis and anaphora. In Weissenborn, Jürgen & Klein, Wolfgang (eds.), *Here and There: Cross-linguistic Studies on Deixis and Demonstration*, 167–186. Amsterdam: John Benjamins Publishing. DOI: <https://doi.org/10.1075/pb.iii.2-3.09hau>
- He, Ruining & McAuley, Julian. 2016. Ups and Downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, 507–517. DOI: <https://doi.org/10.1145/2872427.2883037>
- Himmelman, Nikolaus P. 1996. Demonstratives in narrative discourse: a taxonomy of universal uses. In Barbara A. Fox (ed.), *Studies in Anaphora*, 205–254. John Benjamins Publishing. DOI: <https://doi.org/10.1075/tsl.33.08him>
- Jackson, Clare. 2013. ‘Why do these people’s opinions matter?’ Positioning known referents as unnameable others. *Discourse Studies* 15(3). 299–317. DOI: <https://doi.org/10.1177/1461445613480587>
- Jin, Haeyun. 2019. On the anaphoric use of demonstratives this/these in L2 academic writing. *Journal of English for Academic Purposes* 38. 62–74. DOI: <https://doi.org/10.1016/j.jeap.2019.01.002>
- Kaiser, Elsi & Trueswell, John C. 2008. Interpreting pronouns and demonstratives in Finnish: Evidence for a form-specific approach to reference resolution. *Language and Cognitive Processes* 23(5). 709–748. DOI: <https://doi.org/10.1080/01690960701771220>
- Kirsner, Robert S. 1979. Deixis in Discourse: an exploratory quantitative study of the modern dutch demonstrative adjectives. *discourse and syntax* 355–375. doi: https://doi.org/10.1163/9789004368897_016
- Kolhatkar, Varada & Roussel, Adam & Dipper, Stefanie & Zinsmeister, Heike. 2018. Anaphora with non-nominal antecedents in computational linguistics: a survey. *Computational Linguistics* 44(3). 547–612. DOI: https://doi.org/10.1162/coli_a_00327
- Labrador, Belen. 2011. A corpus-based study of the use of Spanish demonstratives as translation equivalents of English demonstratives. *Perspectives* 19(1). 71–87. DOI: <https://doi.org/10.1080/0907676X.2010.481047>
- Laczkó, Krisztina. 2010. Demonstrative pronouns in spatial deixis, discourse deixis, and anaphora. *Acta Linguistica Hungarica* 57(1). 99–118. DOI: <https://doi.org/10.1556/ALing.57.2010.1.5>
- Lakoff, Robin. 1974. Remarks on this and that. In Fox, Robert A. & La Galy, Michael W. & Bruck, Anthony (eds.), *Papers from the tenth regional meeting*, 345–356. Chicago Linguistic Society.
- Lenko-Szymanska, Agnieszka. 2004. Demonstratives as anaphora markers in advanced learners’ English. In Aston, Guy & Bernardini, Silvia & Stewart, Dominic (eds.), *Corpora and Language Learners*, 89–108. John Benjamins Publishing. DOI: <https://doi.org/10.1075/scl.17.07len>

- Levinson, Stephen C. 2004. Deixis. In Horn, Laurence R. (ed.), *The handbook of pragmatics*, 97–121. Oxford: Blackwell. DOI: <https://doi.org/10.1002/9780470756959.ch5>
- Levinson, Stephen C. & Cutfield, Sarah & Dunn, Michael & Enfield, Nick & Meira, Sergio & Wilkins, David (eds.). 2018. *Demonstratives in Cross-Linguistic Perspective*. Cambridge University Press. DOI: <https://doi.org/10.1017/9781108333818>
- Linde, Charlotte. 1979. Focus of attention and the choice of pronouns in discourse. In Givón, Talmy, *Discourse and Syntax*, 337–354. Academic Press. DOI: https://doi.org/10.1163/9789004368897_015
- Lücking, Andy. 2018. Witness-loaded and witness-free demonstratives. In Coniglio, Marco & Murphy, Andrew & Schlachter, Eva & Veenstra, Tonjes (eds.), *Atypical Demonstratives: Syntax, Semantics and Pragmatics*, 255–284. De Gruyter. DOI: <https://doi.org/10.1515/9783110560299-009>
- Lyons, John. 1977. *Semantics*: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511620614>
- MacLaran, Rose. 1980. On two asymmetrical uses of the demonstrative determiners in English 18(9–10). 803–820. DOI: <https://doi.org/10.1515/ling.1980.18.9-10.803>
- Maes, Alfons. 1996. *Nominal Anaphors, Markedness and Coherence of Discourse*. Leuven: Peeters.
- Maes, Alfons & Krahmer, Emiel & Peeters, David. 2022. Understanding demonstrative reference in text: A new taxonomy based on a new corpus. *Language and Cognition*, 1–23. DOI: <https://doi.org/10.1017/langcog.2021.28>
- Maes, Alfons & Noordman, Leo G.M. 1995. Demonstrative nominal anaphors: a case of nonidentificational markedness. 33(2). 255–282. DOI: <https://doi.org/10.1515/ling.1995.33.2.255>
- Mar, Raymond A. & Li, Jingyuan & Nguyen, Anh T.P. & Ta, Cindy P. 2021. Memory and comprehension of narrative versus expository texts: A meta-analysis. *Psychonomic Bulletin & Review* 28(3). 732–749. DOI: <https://doi.org/10.3758/s13423-020-01853-1>
- McCarthy, Michael. 2002. It, this and that. In Coulthard, Malcolm (ed.), *Advances in Written Text Analysis*, 266–275. Routledge.
- McNamara, Danielle S. & Ozuru, Yasuhiro & Floyd, Randy G. 2011. Comprehension challenges in the fourth grade: the roles of text cohesion, text genre, and readers' prior knowledge. *International Electronic Journal of Elementary Education* 4(1). 229–257.
- Niimura, Tomomi & Hayashi, Brenda. 1994. English and Japanese demonstratives: a contrastive analysis of second language acquisition. *Issues in Applied Linguistics* 5(2). Retrieved from <https://escholarship.org/uc/item/66t5t4zf>. DOI: <https://doi.org/10.5070/L452005199>
- Niimura, Tomomi & Hayashi, Brenda. 1996. Contrastive analysis of English and Japanese demonstratives from the perspective of L1 and L2 acquisition. *Language Sciences* 18(3). 811–834. DOI: [https://doi.org/10.1016/S0388-0001\(96\)00049-6](https://doi.org/10.1016/S0388-0001(96)00049-6)
- Oh, Sun-Young. 2001. A focus-based study of English demonstrative reference: with special reference to the genre of written advertisements. *Journal of English Linguistics* 29(2). 124–148. DOI: <https://doi.org/10.1177/00754240122005260>

- Oh, Sun-Young. 2009. Korean college students' use of english demonstratives in argumentative essays. *English Teaching* 64(1). 51–78. DOI: <https://doi.org/10.15858/engtea.64.1.200903.51>
- Passonneau, Rebecca J. 1989. Getting at discourse referents. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, 51–59. Vancouver, British Columbia, Canada: Association for Computational Linguistics. DOI: <https://doi.org/10.3115/981623.981630>
- Pavesi, Maria. 2013. This and that in the language of film dubbing: a corpus-based analysis. *Meta: Journal Des Traducteurs / Meta: Translators' Journal* 58(1). 103–133. DOI: <https://doi.org/10.7202/1023812ar>
- Peeters, David & Krahmer, Emiel & Maes, Alfons. 2021. A conceptual framework for the study of demonstrative reference. *Psychonomic Bulletin & Review* 28(2). 409–433. DOI: <https://doi.org/10.3758/s13423-020-01822-8>
- Peeters, David & Özyürek, Asli. 2016. This and that revisited: a social and multimodal approach to spatial demonstratives. *Frontiers in Psychology* 7. DOI: <https://doi.org/10.3389/fpsyg.2016.00222>
- Petch-Tyson, Stephanie. 2000. Demonstrative expressions in argumentative discourse: A computer corpus-based comparison of non-native and native English. In Botley, Simon P. & McEnery, Tony (eds.), *Corpus-based and Computational Approaches to Discourse Anaphora*, 43–64. John Benjamins Publishing. DOI: <https://doi.org/10.1075/scl.3.02pet>
- Poesio, Massimo & Modjeska, Natalia N. 2005. Focus, activation, and this-noun phrases: an empirical study. In Branco, António & McEnery, Tony & Mitkov, Ruslan (eds.), *Anaphora Processing: Linguistic, cognitive and computational modelling*, 429–442. John Benjamins Publishing. DOI: <https://doi.org/10.1075/cilt.263.24poe>
- Potts, Christopher & Schwarz, Florian. 2010. Affective 'this'. *Linguistic Issues in Language Technology* 3. 1–30. DOI: <https://doi.org/10.33011/lilt.v3i.1219>
- Prince, Ellen F. 1981. On the interfacing of indefinite-this NPs. In Joshi, Aravind K. & Webber, Bonnie L. & Sag, Ivan A. (eds.), *Elements of Discourse Understanding*, 231–250. Cambridge University Press.
- Ribera, Josep & Cuenca, Maria. 2013. Use and translation of demonstratives in fiction: a contrastive approach (English-Catalan). *Catalan Review: International Journal of Catalan culture*. 27(1). 27–51. DOI: <https://doi.org/10.3828/CATR.27.1.27>
- Roberts, Craige. 2002. Demonstratives as definites. In van Deemter, Kees & Kibble, Rodger (eds.), *Information sharing: Reference and Presupposition in Language Generation and Interpretation*, 89–136. CSLI Press.
- Rocca, Roberta & Tylén, Kristian & Wallentin, Mikkel. 2019. This shoe, that tiger: Semantic properties reflecting manual affordances of the referent modulate demonstrative use. *PLoS ONE* 14(1). e0210333. DOI: <https://doi.org/10.1371/journal.pone.0210333>
- Rocca, Roberta & Wallentin, Mikkel. 2020. Demonstrative reference and semantic space: a large-scale demonstrative choice task study. *Frontiers in Psychology* 11. DOI: <https://doi.org/10.3389/fpsyg.2020.00629>

- Schlegloff, Emanuel A. 1996. Some practices for referring to persons in talk-in-interaction: a partial sketch of a systematics. In Fox, Barbara A. (ed.), *Studies in Anaphora*, 437–486. John Benjamins Publishing. DOI: <https://doi.org/10.1075/tsl.33.14sch>
- Schnedecker, Catherine. 2006. SN démonstratifs « prédicatifs » : qu'est-ce qui limite leur apport informatif ? *Langue française* 152(4). 39–55. DOI: <https://doi.org/10.3406/lfr.2006.6634>
- Scott, Kate. 2013. This and that: A procedural analysis. *Lingua* 131. 49–65. DOI: <https://doi.org/10.1016/j.lingua.2013.03.008>
- Sichel, Ivy & Wiltschko, Martina. 2021. The logic of Person markedness: Evidence from pronominal competition. *Language* 97(1). 42–71. DOI: <https://doi.org/10.1353/lan.2021.0001>
- Sidner, Candace. 1983. Focusing in the comprehension of definite anaphora. In Brady, Michael & Berwick, Robert C. (eds.), *Computational Models of Discourse*, 267–330. MIT Press. Retrieved from <https://ci.nii.ac.jp/naid/10000030796/>
- Smith, John C. 1995. L'évolution sémantique et pragmatique des adverbes déictiques ici, là et là-bas. *Langue française* 107(1). 43–57. DOI: <https://doi.org/10.3406/lfr.1995.5304>
- Stirling, Lesley & Huddleston, Rodney. 2002. Deixis and anaphora. In Huddleston, Rodney D. & Pullum, Geoffrey K. (eds.), *The Cambridge grammar of the English language*, 1449–1564. Cambridge, UK; New York: Cambridge University Press. DOI: <https://doi.org/10.1017/9781316423530.018>
- Strauss, Susan. 2002. This, that, and it in spoken American English: a demonstrative system of gradient focus. *Language Sciences* 24(2). 131–152. DOI: [https://doi.org/10.1016/S0388-0001\(01\)00012-2](https://doi.org/10.1016/S0388-0001(01)00012-2)
- Swierzbinska, Bonnie. 2010. Demonstratives' special place in the English reference system: why that's important for English language learners. *Language and Linguistics Compass* 4(10). 987–1000. DOI: <https://doi.org/10.1111/j.1749-818X.2010.00248.x>
- Tomasello, Michael. 2008. *Origins of human communication*. MIT Press. DOI: <https://doi.org/10.7551/mitpress/7551.001.0001>
- Toole, Janine. 1996. The effect of genre on referential choice. In Fretheim, Thorstein & Gundel, Jeanette K. (eds.), *Reference and Referent Accessibility*, 263–290. John Benjamins Publishing. DOI: <https://doi.org/10.1075/pbns.38.16too>
- Uryupina, Olga & Artstein, Ron & Bristot, Antonella & Cavicchio, Federica & Delogu, Francesca & Rodriguez, Kepa J. & Poesio, Massimo. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU Corpus. *Natural Language Engineering* 26(1). 95–128. DOI: <https://doi.org/10.1017/S1351324919000056>
- Vanderbauwhede, Gudrun & Desmet, Piet & Lauwers, Peter. 2011. The shifting of the demonstrative determiner in French and Dutch in parallel corpora: from translation mechanisms to structural differences. *Meta: Journal Des Traducteurs / Meta: Translators' Journal* 56(2). 443–464. DOI: <https://doi.org/10.7202/1006186ar>
- Voorhees, Ellen & Graff, David. 2008. *AQUAINT-2 information-retrieval text research collection*. Linguistic Data Consortium, Philadelphia.
- Weaver III, Charles A. & Kintsch, Walter. 1991. Expository text. In Barr, Rebecca & Kamil, Michael L. & Mosenthal, Peter M. & Pearson, P. David (eds.), *Handbook of Reading Research, Volume 2*, 230–245. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

- Webber, Bonnie Lynn. 1988. Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, 113–122. Retrieved from https://repository.upenn.edu/cis_reports/458. DOI: <https://doi.org/10.3115/982023.982037>
- Wolfe, Michael B.W. & Mienko, Joseph A. 2007. Learning and memory of factual content from narrative and expository text. *British Journal of Educational Psychology* 77(3). 541–564. DOI: <https://doi.org/10.1348/000709906X143902>
- Wu, Yíán. 2004. *Spatial Demonstratives in English and Chinese: Text and Cognition*. John Benjamins Publishing. DOI: <https://doi.org/10.1075/pbns.126>
- Zhang, Jing. 2015. An analysis of the use of demonstratives in argumentative discourse by chinese EFL learners. *Journal of Language Teaching and Research* 6(2). 460–465. DOI: <https://doi.org/10.17507/jltr.0602.29>
- Zwaan, Rolf A. 1994. Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(4). 920–933. DOI: <https://doi.org/10.1037/0278-7393.20.4.920>

