The landscape of affective meaning

by

Víctor Carranza Pinedo

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

University of Milan & École Normale Supérieure

Supervised by Elisa Paganini and Márta Abrusán

April 2022

Abstract

English version

Swear words are highly colloquial expressions that have the capacity to signal the speaker's affective states, i.e., to display the speaker's feelings with respect to a certain stimulus. For this reason, swear words are often called *expressives*. Which linguistic mechanisms allow swear words display affective states, and, more importantly, how can such 'affective content' be characterized in a theory of meaning? Even though research on expressive meaning has produced models that integrate the affective aspects of swear words in a compositional framework, there is extensive evidence that swear words cannot be assigned a single or stable affective interpretation across contexts. For example, even though expletive adjectives (e.g., *damn*), particularistic insults (e.g., *bastard*) and slurs (e.g., *wop*) typically express (and elicit) negatively valenced affective states, they can also be interpreted positively in some contexts. Thus, inspired in recent developments in formal sociolinguistics, I propose an 'indexical' approach to affective meaning. Under this approach, an affective expression is associated with a set of affective qualities, anyone of which may emerge at a given context depending on the interpreter's prior assumptions about the speaker's affective states and/or relation with the target of the swear word. To define this set, also called 'indexical field', I will use the dimensions pleasure, arousal and dominance, standardly employed in cognitive psychology to characterize and measure affective episodes. In this dissertation, thus, the affective meaning of an expression is given by the set of affective states it typically conveys within a linguistic community, but its interpretation at a given context is established by taking into account the interpreter's prior assumptions about the speaker's affective states and/or attitudes with respect to the target of the affective expression.

Abstract

Italian version

Le parolacce sono espressioni altamente colloquiali che hanno la capacità di segnalare gli stati affettivi del parlante, cioè di mostrare i suoi sentimenti rispetto a un certo stimolo. Per questo motivo, le parolacce sono spesso chiamate termini *esspresivi*. Quali meccanismi linguistici permettono alle parolacce di mostrare stati affettivi, e, cosa più importante, come può essere caratterizzato tale 'contenuto affettivo' in una teoria del significato? Anche se la ricerca sul significato espressivo ha prodotto modelli che integrano gli aspetti affettivi delle parolacce in un quadro compositivo, c'è un'ampia evidenza che alle parolacce non può essere assegnata un'interpretazione affettiva unica o stabile nei vari contesti. Per esempio, anche se gli aggettivi imprecativi (ad esempio, *maledetto*), gli insulti particolaristici (ad esempio, *stronzo*) e gli epiteti razziali (ad esempio, *crucco*) esprimono tipicamente (e suscitano) stati affettivi con valenza negativa, possono anche essere interpretati positivamente in alcuni contesti. Così, ispirato dai recenti sviluppi della sociolinguistica formale, propongo un approccio 'indicizzato' al significato affettivo. Secondo questo approccio, un'espressione affettiva è associata a un insieme di qualità affettive, ognuna delle quali può emergere in un dato contesto a seconda delle ipotesi precedenti dell'interprete sugli stati affettivi del parlante e/o sulla relazione con il bersaglio della parolaccia. Per definire questo insieme di qualità, chiamato anche 'campo indiciale', utilizzerò le dimensioni piacere, eccitazione e dominanza, comunemente impiegate in psicologia cognitiva per caratterizzare e misurare gli episodi affettivi. In questa dissertazione, quindi, il significato affettivo di un'espressione è dato dall'insieme degli stati affettivi che essa trasmette tipicamente all'interno di una comunità linguistica, ma la sua interpretazione in un dato contesto è stabilita tenendo conto delle ipotesi precedenti dell'interprete sugli stati affettivi e/o gli atteggiamenti del parlante rispetto al bersaglio dell'espressione affettiva.

Abstract

French version

Les jurons sont des expressions colloquiales qui ont la capacité de signaler les états affectifs du locuteur, c'est-à-dire d'afficher ses sentiments à l'égard d'un certain stimulus. Pour cette raison, les jurons sont souvent appelés *expressifs*. Quels sont les mécanismes linguistiques qui permettent aux jurons d'afficher des états affectifs et, plus important encore, comment peut-on caractériser un tel 'contenu affectif' dans une théorie de la signification ? Même si la recherche sur la signification expressive a produit des modèles qui intègrent les aspects affectifs des jurons dans un cadre compositionnel, il existe de nombreuses preuves que les jurons ne peuvent pas se voir attribuer une interprétation affective unique ou stable dans tous les contextes. Par exemple, même si les adjectifs explétifs tels que *sacré*, les insultes particularistes tels que *connard*) et les termes discriminatoires tels que *rital*) généralement expriment (et suscitent) des états affectifs à valence négative, ils peuvent également être interprétés positivement dans certains contextes. Ainsi, en m'inspirant des développements récents en sociolinguistique formelle, je propose une approche 'indexicale' de la signification affective. Selon cette approche, une expression affective est associée à un ensemble de qualités affectives, dont chacune peut émerger dans un contexte donné en fonction des hypothèses préalables de l'interprète sur les états affectifs du locuteur et/ou sa relation avec la cible du juron. Pour définir cet ensemble, également appelé 'champ indiciel', j'utiliserai les dimensions plaisir, excitation et dominance, classiquement employées en psychologie cognitive pour caractériser et mesurer les épisodes affectifs. Ainsi, dans cette thèse, la signification affective d'une expression est donnée par l'ensemble des états affectifs qu'elle véhicule typiquement au sein d'une communauté linguistique, mais son interprétation dans un contexte donné est établie en tenant compte des hypothèses préalables de l'interprète sur les états affectifs et/ou les attitudes du locuteur à l'égard de la cible de l'expression affective.

Contents

Chapter 1

Introduction

This dissertation is about curse words, that is, terms that allow speakers express intense emotions as a primary aspect of their meaning. The central proposal can be summarized as follows: people interpret curse words probabilistically. More specifically, a curse word is interpreted by reasoning about i) the likelihood that a stereotypical speaker will utter a curse word given that they experience a particular emotion and ii) the prior probability that the actual speaker is disposed to feel some emotion with respect to the curse word's target. Among the wide array of curse words that can be found, I will concentrate on expletive adjectives, particularistic insults and slurring terms. In this chapter, I introduce and discuss some theoretical background before moving to the analysis of curse words.

Curse words (i.e., insults, profanities, slurs, etc.) have received extensive attention by linguists and philosophers in recent years (Kaplan, 1998; Potts, 2004; Blakemore, 2011). The interest on curse words and cursing partly stems from the fact that their utterance display, rather than report, the affective states experienced by the speaker (e.g., emotions, mood, sentiments, etc.). Thus, even though the utterances in (1a-b) may be employed to convey the

same state of affairs (e.g., the speaker's surprise), what distinguishes them is that each does it 'through different *modes* of expression' (Kaplan, 1998, p. 17). In particular, while (1a) shows the speaker as being shocked, (1b) describes him as being in such state.

(1) a. Damn!
b. I am shocked!

It should be noted, though, that this same feature (i.e., curse words' capacity to immediately display affective states) partially explains why their (socio-)linguistic study remained marginal. In the early studies of brain lesions, researchers noticed that aphasic patients, who are unable to produce fluent speech, nonetheless retain their ability to use curse words (Lordat, 1843). And, in the non-aphasic, it was observed that cursing typically arises during strong emotional episodes, as uncontrolled verbal reactions (Jackson, 1958). Thus, cursing was primarily linked to the sub-cortical areas of the brain, which are also in charge of reflexive, automatic vocalizations such as laughing or shouting. For these reasons, cursing was not considered to be purposeful, and sometimes not even part of speaker's 'genuine' language (Pinker, 1995).

However, there is now considerable evidence that curse words can be 'purposeful and rule-governed' (Jay, 2000, p. 17). From a semantic perspective, extensive data shows that there is a grammar to the distribution of curse words. In this line of research, linguists have been concerned with the way in which semantic composition with emotive expressions takes place (Potts, 2004; McCready, 2010; Gutzmann, 2015). A second line of research is found in sociolinguistics, which looks at ways in which speakers employ curse words as part of their linguistic identity rather than as out of frustration, anger or passion. In this framework, curse words are seen as resources that assemble with non-linguistic elements with the aim of making salient a recognizable

persona (Burridge and Mulder, 1998). Finally, an emergent area of research situates the use of curse words, and impolite behavior in general, within a general theory of rational-linguistic behavior. The interest on this aspect of curse words stems from the fact that, in many contexts, their use requires that speaker and listener coordinate on a common interpretation of the affective information displayed for communication to be successful (McCready, 2012).

Important results have been obtained on each of these areas. However, with some exceptions (e.g., McCready, 2012), the domains of research mentioned above tend not to interact with each other. Research on the grammatical features of curse words tend not to consider observations made within the sociolinguistic literature about their use; socio-linguistic and game-theoretic accounts have started to investigate strategic uses of curse words but without proposing a formal theory of how affective information is transmitted. A formal theory that can connect the various aspects of affective communication seems necessary, specially taking into account the growing interest on the illocutionary and perlocutionary effects of harmful and oppressive speech. The aim of the present dissertation is thus to propose a general pragmatic model in which extant hypotheses about the nature of emotional states, and the interpretation of affective expressions can be tested and compared. Since emotional expressions, like curse words, are interpreted via cross-modal integration of different cues at the same time (e.g., facial expressions, intonation, other utterances), locating our model into a broader framework of how we communicate emotions in interactive settings (Ong et al., 2015) is left for the future.

This chapter introduces curse words and their subject matter, i.e., affective states. Section 1.2 distinguishes curse words from other emotionally charged expressions. Section 1.3 briefly presents the three types of curse words that will be studied in later chapters. Section 1.4 introduces a model to char-

acterize affective states, and briefly discusses the reasoning process behind understanding other's underlying emotions. Section 1.5 summarizes the main contributions of this dissertation.

1.1 Types of affective expressions

As in any scientific endeavor, we need to carefully delimit the set of data before embarking on the analysis. However, what exactly counts as an affective term? And which criteria should we use to distinguish among affective terms? There is a wide range of natural language expressions that might be thought of as having an affective function as part of their meaning. Therefore, our task in this section will be to make explicit the criteria we will use to distinguish affective from non-affective expressions and, among the former, those that will be the subject of this dissertation.

A standard strategy in the study of affective expressions has been to, first, look for those expressions that are intuitively considered to trigger inferences about the speaker's emotions and, second, organize them according to their distribution and behavior within different syntactic environments. However, a problem with such strategy is that is doesn't give us much insight into what makes an expressions 'affective' in the first place and, moreover, that types of affective expressions don't necessarily map to types of grammatical categories (e.g., the class of slurs include nouns, adjectives and verbs, as we will see in Chapter 6).

Instead, we will consider what properties characterize affective states in general and then classify expressions according to their relation to such properties. In cognitive psychology, affective phenomena (emotions, moods, sentiments, etc.) are characterized in terms of their valence (also called 'pleasure'), arousal and dominance properties (Mehrabian and Russell, 1974). Pleasure refers to how pleasant or not the subject evaluates an stimulus (e.g., whereas

happiness is felt as pleasant, fear is felt as unpleasant); *arousal* refers to how excited or energetic the subject feels with respect to the stimulus (e.g., while surprise is felt as arousing, boredom is felt as non-arousing); and *dominance* refers to how in control of its environment the subject feels with respect to the stimulus (e.g., whereas anxiety is non-dominant, anger is dominant).

First, consider evaluative expressions, exemplified by personal-taste predicates (e.g., *fun*, *boring*, *tasty*, etc.). Evaluative terms are considered to carry appraisal as part of their meaning, and thus are typically used in value judgments that express the speaker's positive or negative attitudes towards the object of the predication (Vayrynen, 2013). In that sense, evaluatives can be understood as affective terms but primarily linked to the first dimension, i.e., pleasure. For example, by uttering (2a), the speaker expresses how pleasant they find John:

(2) a. John is **fun**.
b. Doing homework is **boring**.

Second, consider descriptive expressions such as *womanizer*, *hookup* or *brothel*. In Warriner et al. (2013) study, it is observed that participants systematically associate these terms with high degrees of intensity. Thus, they can be understood as being affective but primarily linked to the second dimension, i.e, arousal. For example, by uttering (3a), the speaker not only describes Mary and Alex as engaged in some form of intimacy, but also expresses their strong reaction as provoked by it. Notice that, even though such judgment may be also incidentally associated with a positive or negative evaluation of Mary and Alex, it doesn't need to.

(3) a. Mary and Alex **hooked up** at the party.
b. John is a **womanizer**.

Finally, consider curse words such as *damn* or *bastard*. The observations in Jay (2000), then confirmed by the normative data in Janschewitz (2008), indicate that cursing expressions are not only associated with a particular (typically negative) evaluation, but also yield higher arousal ratings than non-curse words. Therefore, curse words can be understood as affective, but linked to the pleasure and arousal dimensions simultaneously (and, in cases we will see in Chapter 6, to the dominance dimension as well). For example, by uttering (4a), the speaker typically expresses both her negative attitudes towards the dog (e.g., her displeasure) with an additional high degree of affect (e.g., with excitement):

(4) a. John owns a **damn** dog.
b. My boss is a **bastard**.

Perhaps due to their multidimensional affective character, curse words are not only used to express the speaker's emotions, but also to elicit them in others. Indeed, curse words may strategically aim at threatening, harming or joking, rather than at merely expressing the speaker's emotions (Jay, 1992). For example, slurs' effect is not limited to the expression of the speaker's negative evaluation of the target, but also include the allocation of a subordinate social role (Popa-Wyatt and Wyatt, 2017). In stark contrast, expressions that are exclusively linked either to the valence or arousal dimensions have a more limited range of effects. To wit, even though evaluative (e.g., *fun*) or arousing (e.g., *hook up*) words may sometimes have an influence on the speaker's relation to others, they are not typical resources in speech-acts that aim at shaping such relation.

In sum, what makes an expression affective? That it systematically triggers inferences that make reference to one or more affective dimensions, that is, to a certain degree of pleasure, arousal or dominance. How can affective

expressions be distinguished from each other? Primarily, by the type(s) of affective dimension(s) they are systematically associated with and, as we will see in the next section, by how such association is grammatically manifested. For the purposes of this dissertation, I will only focus on the third type of affective expressions listed above. That is, on expressions that are systematically linked to more than one affective dimension simultaneously, primarily exemplified by curse words.

1.2 Curse words: definitions and examples

Let's turn to the data. Thanks to the work of Potts (2004, 2007a,b), who picks up the discussion initiated in (Kaplan, 1998), curse words have increasingly received attention in the semantics and pragmatics literature. Examples in English include the adjectives in (5); examples in Spanish can be found in (6):

(5) a. I hear your **damn** dog barking. (Gutzmann, 2015)
b. **Fucking** Mike Tyson got arrested for domestic violence. (McCready, 2012)

(6) a. Llevo toda la tarde con la **dichosa** ponencia.
'I have been working on the bleeding presentation the whole afternoon.' (Padilla Cruz, 2018)
b. No he visto al **puto** perro.
'I haven't seen the bloody dog'.

With some exceptions, which we will study later, these modifiers signal the speaker's affective states without describing any specific state of affairs. In other terms, they contribute affective but no truth-conditional meaning. As a result, from a truth-conditional perspective, adding or omitting them does

not alter the content asserted by the host utterance. Following Cruse (2006) and Gutzmann (2015), I will call these expressions *expletive adjectives*. These expressions are the subject of Chapter 4.

Now, expletive adjectives are often classified alongside pejorative nouns and adjectives that occur as non-restrictive modifiers. Similarly to expletive adjectives, pejorative non-restrictive modifiers don't contribute to the truth-conditions of their host utterance, so they can be omitted without altering the main at-issue content asserted by it Potts (2004). Following Saka (2007a), I will call these expressions *particularistic insults*:

(7) a. That **bastard** Kresge is famous. (Potts, 2007a)
b. That **idiot** Kresge dropped the bottle again. (Gutzmann, 2015)

(8) a. Ese **idiota** de ahí me está mirando.
'That idiot there is staring at me.'
b. El **rata** de mi jefe me regaló una PC antigua.
'My jerk boss gave me an old PC.'

However, a basic feature that distinguishes expletive adjectives from particularistic insults (more will be discussed later) is that the latter also have uses where they contribute truth-conditional content. That is, uses where they cannot be possibly omitted without altering the content and grammatically of the host utterance. Thus, in contrast to expletive adjectives, particularistic insults not only display the speaker's affective states towards a target, but also describe it in a certain way:

(9) a. John is a **jerk**. (Beller, 2013)
b. I knew John would be a **bastard** and run.

(10) a. Juan es solo un **pendejo** en un carro muy caro.

'Juan is just a dick in a fancy car.'

b. Gabriel es un **huevon**, siempre me hace enojar.
'Gabriel is a fuckhead, he always makes me angry.'

Now, depending on their degree of descriptiveness, particularistic insults can be distinguished in two sub-types. On the one hand, what I call *thin PIs* (also called 'all-purpose pejoratives' in the literature), express the speaker's affective states towards someone or something but are too lean to indicate the property that evokes such affective reaction. On the other, *thick PIs* express the speaker's affective states but, additionally, make explicit the descriptive basis of the reaction. Examples of thin PIs include the aforementioned *jerk* and *bastard*, and examples of thick PIs include *crook*, who refers to the criminal, and *wimp*, who refers to the coward.

(11) a. Bringing in a **crook** to run the company is a new low.
b. John is being a **wimp** but Alex is being **wimpier**.

It is worth noting that some PIs may have thin or thick uses depending on the context. In some situations, calling someone a *rata* (i.e., 'rat') in Spanish may merely express the speakers heightened negative evaluation of the target, but in others it can also describe it as being greedy. Particularistic insults will be studied in Chapter 5.

The final group of expressions studied in this dissertation are slurs. This class includes terms like *Boche*, which refers to German people, and the Spanish *Sudaca*, which refers to South-American people. By uttering (12b), the speaker asserts that he likes the music from South-America, but at the same time expresses a derogatory attitude towards South-American people:

(12) a. Lessing was a **Boche**.

b. Me gusta la musica **sudaca**.
'I like South-American-PEJ music.'

If one wants to get rid of the negative and potentially harmful effect of slurs without altering the truth-conditions of the host utterance, one should replace the slur by a neutral counterpart that makes reference to the same group (Gutzmann, 2015):

(13) a. Lessing was German.
b. Me gusta la musica sudamericana.
'I like South-American music.'

However, slurs not only indicate that the speaker evaluates negatively the group targeted. In addition, slurs express that the members of the target group are *lesser*, unworthy, or undeserving of respect as individuals (Jeshion, 2013). To capture the specific scornful denigration expressed by slurs, I will argue that they are linked to the pleasure and dominance dimensions. By uttering a sentence like (12b), the speaker not only expresses her negatively valenced attitudes towards South-Americans, but also that he sees himself as dominant with respect to them. Slurs will be the subject of Chapter 6.

A recurrent theme throughout this dissertation is that, even though expletive adjectives, particularistic insults and slurs are considered to signal negatively valenced affective states (e.g., contempt), they are largely underspecified with respect to the emotions they can express. Even though the expressions in (14-16) display a negativity bias, they can be also interpreted as expressing positive affective states in some contexts (Potts, 2004; McCready, 2012). For example, the felicitous continuations in (14) show that,

(14) The **fucking** dog was barking all night.

a. ...I am glad he chased the robbers away.
b. ...I couldn't sleep well because of that.

(15) Hey, **bastard**!

a. ...You have been such a good friend.
b. ...You forgot to pay the rent.

(16) There are **Queers** in the show.

a. ...It will be super fun.
b. ...We shouldn't have come.

Broadly speaking, our strategy to address this problem is that affective expressivity is determined contextually, by the interplay between listener's and speaker's expectations about each other's communicative intentions. In particular, I will argue that affective underspecification is resolved by interpreters by reasoning on the basis of what they assume about the speaker's feelings or affective dispositions. Moreover, since competent speakers are likely to be aware of curse words' underspecification, they will decide when and with whom to use these expressions without much worry about being misunderstood. Chapter 3 will present this proposal in detail.

The linguistic forms introduced in this section will comprise the empirical domain of this dissertation. To recap, I will only focus on those expressions which are linked to more than one affective dimension (i.e., pleasure, arousal, dominance) at the same time and, among these, on curse words. Now, since we aim at modeling we reason about other's affective states on the basis of some of the expressions they use, the next section will address two preliminary issues, namely, i) how is affective phenomena standardly measured/characterized in multidimensional psychological frameworks? and ii) how do people reason about other's underlying affective states on the basis of different contextual cues?

1.3 Emotions: measurement and appraisal

1.3.1 Pleasure, Arousal and Dominance

Affective dimensions will be used to inform our probabilistic model. Broadly speaking, we will use them to characterize the wide array of possible affective states expressed by curse words at a given utterance context. Yet, it should be noted that our proposal won't strictly depend on any specific theory of the psychometric properties of affective phenomena. Instead, we will propose a model in which different views about affective phenomena can be exploited in the linguistic analysis of affective communication.

What aspects characterize an affective episode? Mehrabian and Russell (1974) argue that affective episodes can be adequately described using three continuous, bi-polar and orthogonal dimensions: pleasure, arousal and dominance. This psychometric approach to affective phenomena, also known as the 'PAD' theory of emotions, originates in the work of Wundt (1896), and is widely applied in the analysis of affective episodes in a continuous rather than discrete framework. The three dimensions, briefly introduced above, are defined as follows:

- Pleasure: corresponds to a continuum ranging from negatively valenced affective states (e.g., sadness) to positively valenced ones (e.g., happiness). It is the evaluative component.
- Arousal: corresponds to the continuum ranging from low mental alertness (e.g., boredom) to high mental alertness (e.g., excitement). It is the physiological component.
- Dominance: corresponds to the continuum ranging from the sensation of being controlled (e.g., anxiety) to the sensation of being in control of one's surroundings (e.g., relaxation). It is a relational component.

Since Mehrabian and Russell (1974), different versions of this model have been proposed. For example, in Russell's later work (Russell, 1980, 1989), only the pleasure and arousal dimensions are taken into account. According to Russell, since dominance relates to the subject's relation to her environment during the affective episode, it doesn't capture an emotion's core quality. Yet, we will employ these dimensions if the informational impact of a natural language expression can be plausibly characterized by it, even though they may not necessarily reflect the core aspects of how emotions actually work.

How are affective states characterized using these dimensions? Mehrabian (1996), for example, employs these dimensions to define a 'temperament space', where high and low values in each dimension determine 8 basic affective states. In this framework, affective states characterize agent's emotional predispositions across representative life situations. As we can observe, 'hostility' corresponds to the [P-, A+, D+] state, 'anxiety' to the [P-, A+D-] state, etc. (see Tarasenko (2010), based on Mehrabian (1996)):

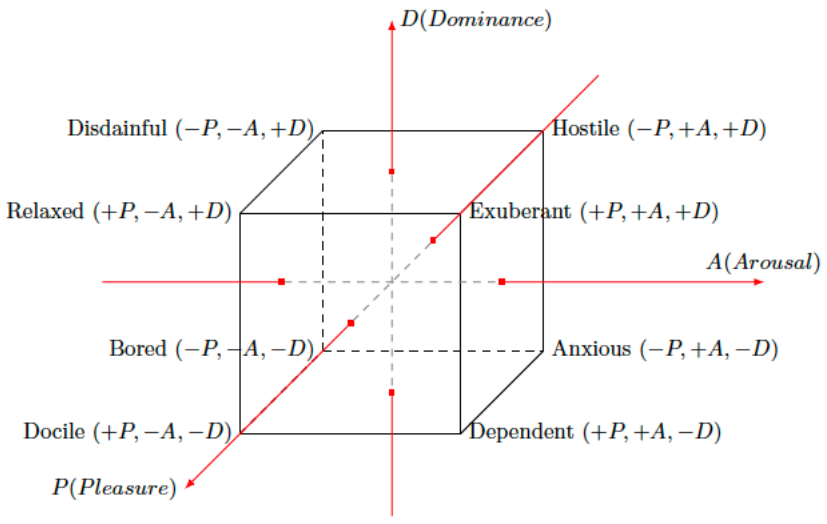


Figure 1.1: Mehrabian's 8 basic affective states (Tarasenko, 2010)

In contrast, in Reisenzein (1994) study, prototypical affective states are represented by a wedge in a two dimensional space whose axes represent degrees of pleasure and arousal. In this affective space, the location and orientation of each wedge represents the 'quality' of an affective episode, whereas its different points represent the different degrees of intensity in which it can be instantiated. The lower point of intensity is the one closer to the intersection of the two axes, which can be understood as a state of 'hedonic neutrality':

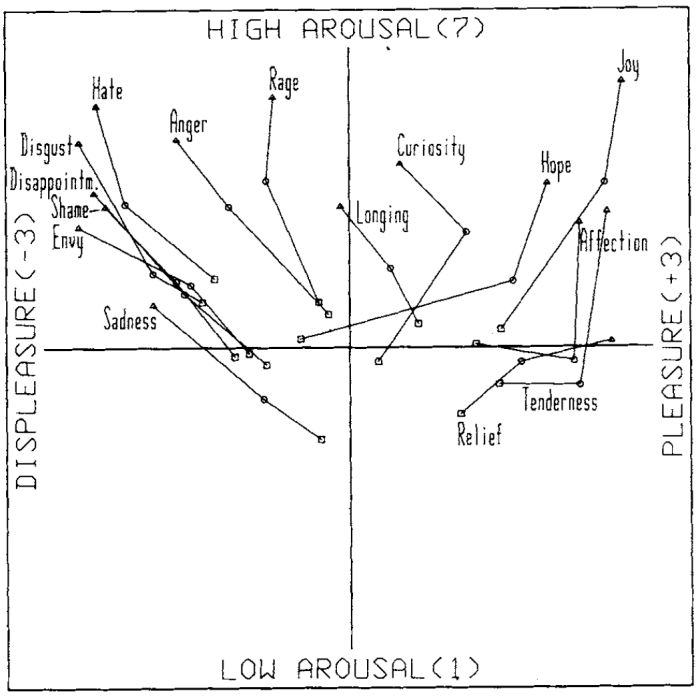


Figure 1.2: Mean pleasure of 30 affects (Reisenzein, 1994)

The comparison between this two models can help us understand some the basic features of affective episodes. First, Reisenzein's model incorporates a fourth element, namely the intensity in which an emotion-type can be experienced by a subject. In his framework, the intensity of an emotion depends on 'how extreme' an affective state scores on one of the dimensions, independently of whether such value is positive or negative. Indeed, high arousal

emotions (e.g., surprise) and low arousal emotions (e.g., frustration) can be nonetheless intense. Second, whereas Mehrabian's model focus is on 'standing' affective states, Reisenzein focus is on 'occurrent' affective states (using Lyons (1980)'s terminology). Imagine that Mary says that John is angry with a dog. Such utterance can be interpreted as indicating an occurrent state, i.e., as indicating that John is at the grip of anger at the utterance's context, or as indicating a standing state, i.e., as indicating that John is disposed to feel anger towards the dog in specific situations (even though, at the utterance's context, he is not). As we will see in later chapters, this ambiguity is also reflected in the interpretation of curse words, which can be interpreted as expressing standing or occurrent emotions depending on the situation.

1.3.2 Affective cognition

Our model aims at capturing the way in which interpreters infer speakers' underlying emotional states based on their use of a curse word. Thus, two preliminary questions arise: i) how do lay individuals reason about other's emotions? and ii) what distinguishes curse words from other types of affective cues?

Let's address (i). Intuitively, we assume that other's emotional states arise as reactions to specific events, e.g., we expect someone to be happy if she is hired in a company and sad if she is fired. However, we not only use the outcome of events to infer someone's emotions, but our assumptions about their beliefs or desires, e.g., if we know that a new job won't satisfy someone's career goals, then our original inference (e.g., that they are happy if they are hired) will be defeated. In this case, we reason 'forward' about how events and mental states cause an individual's emotions.

However, we also expect emotional states to, in turn, cause specific types of actions (e.g., fleeing, attacking, etc.) and expressions (e.g., facial expressions,

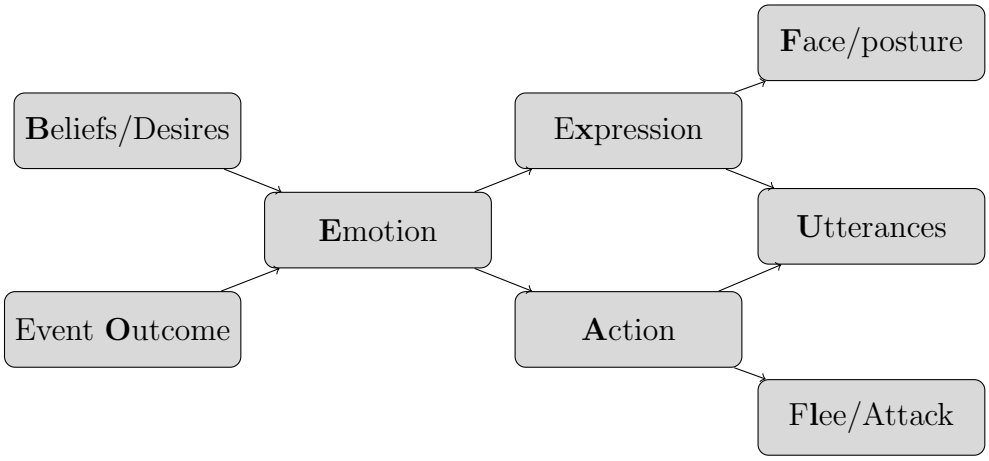


Figure 1.3: Lay theories of emotions (based on Ong et al. (2019))

body posture, vocalizations, etc.). For example, if we see an individual fleeing from a dog, we may infer that he fears it; if we observe an individual smiling while fleeing from a dog, we may infer that they are in fact playing, so the original inference will be overridden. In this case, we reason 'backwards' from the actions and expressions which are typically caused by an individual's emotions (Hess and Hareli, 2015, 2017).

Thus, even though the causal flow from events and mental states to emotions and, in turn, from emotions to expressions and actions is unidirectional (see figure 1.3), information can flow in different directions (Ong et al., 2015).

A clear way to represent this flow of reasoning, and the one that we will use in the model proposed in Chapter 3, employs Bayes' theorem (Ong et al., 2015, 2019; Saxe and Houlihan, 2017). In the basic case, e.g., inferring an agent's emotion **e** after he performs an action **a** (i.e., P(e|a), read as 'the probability that someone is feeling **e** based on action **a**'), we can combine the likelihood that the action **a** is performed given that the agent is experiencing an emotion **e** (i.e., P(a|e)) and the prior probability that emotion **e** occurs (i.e., P(e)), and then divide the result by the probability of action **a** occurring in the first

place (i.e., P(a)):

$$P(e|a) = \frac{P(a|e)P(e)}{P(a)} \tag{17}$$

Moreover, Bayes' theorem can also be used to represent the integration of multiple cues during the inference of an emotional state (Zaki, 2013; Ong et al., 2019). For example, we can use this formula to infer an agent's emotion **e** from her actions **a** and expressions **x** (i.e., P(e|a, x). Now, it is worth noting that these various emotional cues can be either in conflict or in harmony with each other. In our previous example, observing someone fleeing from a dog while smiling suggested conflicting degrees of pleasure (i.e., negative and positive, respectively). While it has been standardly argued that, in general, facial expressions dominate other types of affective cues, some studies suggest that cues' degree of reliability is highly sensible to the context (Hess and Hareli, 2015; Kayyal et al., 2015). For example, depending on the type of dog (e.g., a Bullmastiff) and the way in which the agent is running, his smile may not prevent us from inferring that he fears the dog.

Let's address (ii). In the figure 1.3, we assumed a strict distinction between affective expressions (e.g., facial expressions, posture), which are standardly viewed as involuntary 'read-outs' of affective states, and actions (e.g., fleeing a situation, gestures), which are instead viewed as dependent on the agent's intentionality (Townsend et al., 2017). Yet, even though actions may be more easily controllable than expressions, both types of affective reactions seem to occur with different degrees of intentionality. For example, gestures may be more easy to exaggerate or inhibit compared to facial expressions, but that doesn't prevent individuals from regulating their facial expressions depending on who they are interacting with and their social goals. Moreover, even preschooler children are proficient at regulating crying and talking about their affective states (Cole, 1986), thus showing that the use of affective

expressions can be strategical from the early stages of development.

The distinction between actions and expressions has also been applied to verbal utterances. It was standardly assumed that, whereas utterances reporting events, including the speaker's own emotions (e.g., 'I am sad') are strategically controlled by the speaker, utterances expressing emotions (e.g., 'Fuck!') constitute reflexive, automatic vocalizations analogous to shouting or laughing (Jackson, 1958). However, evidence shows that cursing can also be strategic: people regularly modify, exaggerate or inhibit from cursing depending on their audience and the social goals they want to achieve (Jay, 1992, 2000). This is the reason why, throughout this dissertation, I will assume that, consciously or not, the use of curse words is linked to the agent's social goals, and thus can be studied as a form of interactive rational language use.

Finally, note that many actions and expressions are not typically associated with any particular emotion (e.g., closing our eyes, saying 'It is raining', etc.), so they cannot be used to reason about the speaker's emotions unless the context provides such link. For example, if it is common ground that someone hates the rain, his uttering 'It is raining' can be interpreted as signalling that he feels negatively. In contrast, curse words like *damn*, similarly to other types of affective cues (e.g., smiling), provide reliable information about the agent's underlying emotional states across different contexts. For example, the use of *damn* typically expresses negatively valenced states, so a model of affective communication needs to incorporate such negative bias, while at the same time explaining why in some contexts *damn* can also be interpreted positively.

With this background in place, we now move to the linguistic analysis of affective expressions in Chapter 2. There, we will analyze curse words as carrying expressive meaning. To conclude this chapter, I will summarize the main proposal of the dissertation, and the main arguments in each chapter.

1.4 Overview of this dissertation

The main argument of this dissertation is the following. Curse words are used to display the speaker's feeling towards a certain stimulus. Yet, they cannot be assigned a stable interpretation across contexts. For example, even though expletive adjectives (e.g., *damn*), particularistic insults (e.g., *bastard*) and slurs (e.g., *Sudaca*) typically express (and elicit) negatively valenced states, they can also be interpreted positively in some contexts. Thus, inspired in recent developments in formal sociolinguistics, I propose an 'indexical' approach to affective meaning. Under this approach, an affective expression is associated with a set of affective qualities, anyone of which may emerge at a given context depending on the interpreter's prior assumptions about the speaker's. To define this set, also called 'indexical field', I will employ the dimensions pleasure, arousal and dominance, introduced in the previous section. In a nutshell, in our model, a curse word is interpreted by reasoning about i) the likelihood that a 'stereotypical' speaker will utter the curse word given that they are experiencing some emotion and ii) the prior probability that the actual speaker is predisposed to feel a particular emotion with respect to the curse word's target.

Curse words are standardly classified as *expressives* (Kaplan, 1998; Potts, 2004). Thus, in Chapter 2, I start by analysing curse words *qua* expressive items. Throughout this chapter, I investigate the following questions: i) what kinds of content can be conveyed expressively? ii) how can expressive content be diagnosed? and iii) How should expressive content be explicated in a theory of meaning?

Then, in Chapter 3, I develop a probabilistic pragmatic approach to curse word's interpretation. First, I model affective information by extending Burnett (2017, 2019)'s pioneering work on identity construction through sociolinguistic variation. Then, I represent the interpretation of curse words as a Bayesian/probabilistic update of the listener's prior beliefs about the

speaker's affective tendencies, using as example expletive adjectives like *fucking*. Finally, even though the model that we will use throughout this dissertation is pragmatic, at the end of this chapter I will sketch a semantic implementation based on Potts (2007a,b)'s account, according to which expressives are represented as functions that directly update the context of interpretation.

In Chapter 4, I continue the analysis of expletive adjectives like *fucking* by focusing on two of their main linguistic properties. First, on the fact that they can receive 'non-local interpretations', that is, that they don't necessarily express emotions towards the (object referred by) the argument to which they apply. For example, in the sentence 'Alex says Jones forgot to buy the fucking pizza', *fucking* can be interpreted as conveying a (negative) attitude towards Jones despite its nominal-internal position in the syntax (Potts, 2004; Gutzmann, 2019). Second, on the fact that expletive adjectives can be used as degree modifiers. For example, the sentence *John is fucking intelligent* seems to indicate that John is very intelligent, and not just intelligent (Geurts, 2007).

In Chapter 5, I analyze particularistic insults. Here, I argue that particularistic insults are descriptive of behavioral traits (even though at different degrees of precision). Then, I apply the proposal elaborated in Chapter 3 to understand the way in which the utterance of a particularistic insult (e.g., *crook*) depends on the speaker's 'mood' (i.e., the publicly available information about the speaker's affective states) and his psychological relation with the target of the insult. Thus, this proposal will allow us understand those cases in which particularistic insults are used in friendly settings (e.g., in humorous, playful or romantic contexts).

In Chapter 6, I analyze slurring terms. Here, I argue that slurs are indexically associated with states defined along the pleasure and dominance relations. That is, that slurs not only display a negative evaluation of the target group,

but also that its members are *lesser* (according to the speaker). Then, I extend the proposal elaborated in Chapter 3 to understand the way in which slurs (e.g. *Boche*) are interpreted depending on i) the speaker's mood, ii) his relation with the slur's target, and iii) his social identity. This will allow us understand the situations in which slurs are used to build solidarity rather than to dehumanize.

Chapter 2

Affective meaning as expressive

Since Kaplan (1998)'s work on the meaning of *oops* and *ouch*, linguists and philosophers have taken a great interest in understanding the wide array of lexical items that display information rather than describe or entail it, discussed in formal semantics and pragmatics under the label of 'expressivity'. For example, upon hearing the following utterance,

(1) The {fucking, bloody, damn} dog is barking.

the addressee is likely to interpret the main clause as conveying (i) that a dog is barking and view the bracketed adjectives as (ii) expressing the speaker's affective states about the dog. Intuitively, (i) and (ii) are pieces of content that can be evaluated independently from each other: if there isn't a dog barking, the utterance comes out false, regardless of whether the speaker is in a certain affective state or not; conversely, if the speaker is not in the affective state indicated by the expressive items, the utterance typically comes out as odd, regardless of whether the dog is actually baking or not. In that sense, the contribution of the affective adjectives in (1) is 'independent'

(in a sense to be precised below) from the contribution of the host utterance:

(2) a. 'The {fucking, bloody, damn} dog is barking' is true if the dog is barking.

b. 'The {fucking, bloody, damn} dog is barking' is felicitously used if the speaker has a negative attitude towards the dog. (Gutzmann, 2019, p. 84)

This initial characterization of expressive content as contributing use-conditions rather than truth-conditions raises three basic questions:

A What kinds of content can be conveyed expressively?

B How can expressive content be diagnosed?

C How should expressive content be explicated in a theory of meaning?

The present chapter addresses these three questions with a special focus on curse words. By way of addressing A, Section 2.1.1 compares and contrast the type of information associated with affective and social expressives, and argues that affective expressives may have secondary social effects. Section 2.1.2 distinguishes between what we may call *conventional* expressivity, defined as expressive content that has achieved a certain degree of stability through repetition and circulation within a linguistic community, and *indexical* expressivity, defined as expressive content that is not grounded in convention but in the perceived co-occurrence between a sign and a phenomenon (Peirce 1955, reference). In this section, I argue that affective expressives display properties that are typically indexical rather than conventional.

By way of addressing B, Section 2.2 analyses the linguistic features standardly attributed to curse words. Section 2.2.1 roughly distinguishes those features that apply to all curse words from those that only apply to a subset of

them. Section 2.2.2 gives a closer look at three of their main properties: *independence* (i.e., the fact that curse words contribute to an independent meaning dimension), *perspective-dependence* (i.e., the fact that curse words are always evaluated with respect to an individual point of view) and what I call *affective underspecification* (i.e., the fact that the interpretation of curse words may considerably change depending on the utterance context).

By way of addressing C, Section 2.3 critically examines some of the theories that have been proposed to account for the main properties of curse words. In Section 2.3.1, I discuss semantic approaches that consider affective information to be a type of conventional implicature or presuppositional inference. Section 2.3.2 focuses on those pragmatic approaches that consider affective expressive information to be calculated as conversational implicatures. Finally, Section 2.3.3 focuses on two approaches that consider affective expressivity to constitute a distinguished type of meaning.

2.1 What kind of content can be expressive?

2.1.1 Affective vs. social expressivity

From a synchronic point of view, expressive meaning can be distinguished in 'affective' and 'social'. Affective expressives, as their name indicate, serve to display the speaker's affective states, which we can break up into specific dimensions such as valence, arousal and dominance. This group includes English curse words like *fucking*, endearing uses of Spanish diminutive suffixes such as *-ito*, sentence-initial interjection such as *man*, which indicate the speaker's surprise, etc. For example, by uttering (3a), the speaker typically signals that she is in a negatively valenced, aroused affective state with respect to Jones:

(3) a. **Fucking** Jones is coming to the party.

b. El perr**-ito** está durmiendo.
'The dog-DIM is sleeping.'
c. **Man**, Berlusconi wants to become President of Italy.

In contrast, social expressives display either the speaker's standing with respect to other individuals or the context degree of formality, which we can also break up in different dimensions such as social distance, psychological distance and formality (McCready, 2019). Social expressives include English titles such as *Mr./Mrs.*, the informal/formal Spanish second person pronouns *Tu/Usted*, Japanese honorific nominal modifiers such as *-san*, etc. For example, by uttering (4b), the speaker displays that her relation to Smith is formal rather than casual:

(4) a. **Mrs.** Smith is not available.
b. Alguien preguntó por **usted**.
'Someone asked for You-FORMAL.'
c. Jones**-san** ga nonde imasu.
'Jones-**SAN** is drinking.'

An issue that immediately arises is whether the distinction between affective and social information is well grounded. To wit, the semantics of honorifics has been assumed to involve the display of the speaker's sentiments towards the person addressed or referred (Kaplan, 1998; Potts and Kawahara, 2004). By uttering (4a), the speaker can be interpreted as showing her respect or admiration towards Mrs. Smith. However, emotions are not required to use honorifics felicitously. As McCready (2019) points out, in a context where it is common ground that the speaker doesn't respect or admire the addressee (e.g., a situation where Mrs. Smith is an universally despised boss), her use of *Mrs.* doesn't come out as odd. Therefore, honorifics primarily mark the speaker's social standing with respect to others, and only optionally trigger

implicatures regarding the speaker's sentiments.

Now, does the semantics of affective expressives involves a social component? In the semantic literature on curse words such as *fucking*, it is standardly assumed that these only display the speaker's affective states. However, many of their uses have an impact that goes well beyond the expressions of emotions. Observe the following utterance:

(5) I don't find my **fucking** cellphone. Have you seen it?

First, the adjective *fucking* belongs to a colloquial register, so its utterance may signal that the speaker wants to establish a more casual relationship between himself and the addressee. Second, *fucking* can also be used to intimidate, threaten or harass the addressee. In a situation where the speaker thinks the addressee hided the cellphone, (5) can be be interpreted as displaying anger towards him, rather than towards the cellphone. And, third, curse words such as *fucking* can also make part of the speaker's linguistic style. As (Burridge and Mulder, 1998, p. 13) observe, the use of taboo language functions as a 'desirable macho markers of gender identity in Australia', and, as Eggins and Slade (2004) point out, using curse words is often a resource to perform a macho identity and establish a 'leader' persona.

Yet, from a semantic point of view, social uses of curse words such as *fucking* are only optional. To wit, (5) can be felicitously uttered in i) a context where it is common knowledge that the speaker doesn't see the context as informal, ii) contexts where there is no intention to intimidate or threaten the addressee, or iii) contexts where curse words don't make part of the speaker's idiolect, but are used out of frustration or commotion. Thus, we may consider that curse words primarily display the speaker's emotions and, supplementary, may be used to convey information about the speaker's personae or his social relation with the addressee. Thus, the social aspects of

curse words arise secondarily in the same way that the affective aspects of honorifics arise secondarily. In that sense, we can conclude that the distinction between affective and social expressivity is well grounded.

2.1.2 Conventional vs. indexical expressivity

From a diachronic point of view, expressive meaning can be distinguished in conventional and indexical. On the one hand, conventional expressivity appears to be instantiated by those expressions like *fucking* or *Mr.*, which are widely recognized as lexically encoding a certain type of affective or social meaning. Arguably, their semantic stability is obtained through repetition and circulation within a linguistic community. Depending on the lexical item studied, this process is referred to as 'crystallization' (Jeshion, 2016), 'pragmaticalization' (Davis and Gutzmann, 2015) or 'enregistrement' (Agha, 2003), etc.

On the other hand, indexical expressivity is typically instantiated by variables, that is, contrast sets which include alternative ways of 'saying the same thing' (Labov, 1972). Indexical expressivity is grounded on the perceived association between the occurrence of a variable and some property of the speaker (Silverstein, 1976). For example, even though the following utterances have the same truth-conditions, the different ways of pronouncing (ING) tend to be associated with different social properties (Campbell-Kibler, 2005). In particular, the use of *-ing* tends to be associated with being competent (i.e., educated, articulate, etc.) but aloof (i.e., formal, unfriendly, etc.), and the use of *-in* tends to be associated with the opposite properties, that is, with being incompetent but friendly:

(6) a. John is fishing.
b. John is fishin'.

In the sociolinguistic literature, the set of qualities associated with a variant is called its 'indexical field', defined as a 'constellation of ideologically related meanings, any one of which can be activated in the situated use of the form' (Eckert, 2008, p. 453). Importantly, the use of a variable is interpreted depending on what other properties are believed to hold of the speaker rather than on the speaker's communicative intentions. Using Eckert (2008)'s terminology, what properties end up being 'activated' (e.g., assigned to the speaker) heavily depend on what it is previously believed about the speaker. For example, some people that employ the *-in'* variant can be seen as easy-going or friendly, but others can be seen as insincere or condescending[1].

It may be asked whether the dividing line between conventional and indexical expressivity is as sharp as it may seem at first sight. To wit, it can be observed that conventional expressives, and in particular curse words such as *fucking*, appear to have some 'traces' of indexical expressivity. First, the meaning associated with curse words such as *fucking* can be considered multilayered, as it relates to a constellation of affective dimensions such as valence, arousal and dominance. Second, *fucking*, and curse words in general, are also contingent on what other properties are assumed to hold of the speaker. For example, if the speaker of (7) is known to be a supporter of Berlusconi, her use of *fucking* will be more likely interpreted as indicating surprise or joy, but if he is known to be a critic of Berlusconi, *fucking* will normally be interpreted as indicating anger or frustration:

(7) **Fucking** Berlusconi wants to become President of Italy.

This last feature, henceforth 'affective underspecification', is not exclusive of curse words. Even though diminutive suffixes such as the Spanish *-ito* are typically used to convey endearment in familiar contexts, they can also be

[1]For a review of the main properties of sociolinguistic variables see Beltrama (2020)

used to indicate negatively valenced states, e.g., that the target is inferior or childish (de Klerk and Bosch, 1996). For example, the following utterance can be interpreted as indicating that the speaker feels positively with respect to the professor, or that he sees him or her as childish. In sum, morphemes indexing friendliness can be perceived as either intimate or disingenuous depending on who uses them:

(8) El profesors**-ito** está ocupado.
'The professor**-DIM** is busy.'

Therefore, conventional and indexical contents should not be taken as categorically distinct, but as a 'gradient cline between two phases of the same process' (Beltrama, 2020). Highly activated indexical associations could be analysed as parallel to conventional content and, conversely, highly context-sensitive conventional associations could be analysed as parallel to indexical content.

2.2 Properties of expressive meaning

2.2.1 Overview

Potts (2004, 2007a) groundbreaking work on expressives numerate the following properties as those that an account of conventional expressivity should explain:

- Independence: expressives contribute to a separate dimension of meaning.
- Non-displaceability: expressives invariably predicate something about the utterance situation.
- Perspective-dependence: expressive content is evaluated from a

particular perspective (often the speaker's).

- IMMEDIACY: expressives cannot be challenged and thus achieve their intended effect by being uttered.
- REPEATABILITY: repeating an expressive multiple times in the same utterance strengthens its content; it is not redundant.
- DESCRIPTIVE INEFFABILITY: expressive content cannot be effectively paraphrased using non-expressive terms.

However, there are other properties which are also discussed in the semantic literature about affective expressives, and specially about curse words:

- NON-LOCAL READINGS: expletive adjectives (e.g., *damn*) may have a syntactic realization that differs from their scope of semantic interpretation.
- AFFECTIVE UNDERSPECIFICATION: in general, curse words are underspecified with respect to the affective interpretations they can receive.

This section will critically analyze both the original properties in Potts (2004, 2007a) and those that have attracted the attention of researchers in the subsequent literature. In order to make the discussion more concise, I will only focus on how these properties are manifested by affective expressives, and in particular by curse words. Section 2.2.1 proposes an overview of these properties and organizes them with respect to whether they are manifested by various types of curse words ('main properties') or whether they only apply to a subset of them ('secondary properties'). Then, Section 2.2.2 zooms into three of the most important and perhaps difficult to conceptually pin down: independence, perspective-dependence and affective underspecification.

Main properties

Independence is considered the mark of expressive content. In order to diagnose this property, linguists use tests that show that expressives don't interact with various types of truth-conditional operators (e.g., negations, conditionals, disjunctions, etc.). To illustrate this property, consider the following utterances, where different expressives fall under the syntactic scope of these operators, but where there is no semantic interaction whatsoever:

(9) a. **Fucking** Alex is didn't came to the party.
b. If Jones is a **Boche**, then he will be on time.
c. Either Maria is a **Spic** or else she learnt to speak Spanish at high school.

In (9a), the utterance comes as false if Alex came to the party, independently of whether the speaker feels in a certain way with respect to them or not. In (9b), the utterance comes out as false if Jones is on time but is not German, independently of whether the speaker dislikes German people or not.

Non-displaceability makes reference to the displaceability property, namely, the ability to talk about objects located at different times, space, or modality. Expressives lack this ability, as they invariably convey something about the utterance situation. If expletive adjectives were displaceable, (10a) would have a reading in which the speaker felt annoyed about Jones yesterday. Similarly, if slurs were displaceable, (10b) would have a reading in which the speaker feels annoyed towards South-Americans only if the music comes from South-America:

(10) a. Yesterday, that **fucking** Jones came.
b. Posiblemente, esa es música es **sudaca**.
'Possibly, that music is South-American-PEJ.'

As McCready (2019) points out, non-displaceability follows from independence without further assumptions. To wit, if expressives systematically fail to scopally interact with truth-conditional operators, then, because displacement depends on the interaction with truth-conditional operators (e.g. *yesterday*, *possibly*), expressive content cannot displaced.

Perspective-dependence is standardly defined as the idea that expressive content is evaluated from a particular perspective. In the case of emotive expressives, this property thus points to the idea that terms like *fucking* or *Sudaca* involve a 'judge's' viewpoint. which often, but not always, corresponds to the speaker's own viewpoint. As it will be observed in Section 2.2.2, perspective-dependence has been mainly investigated by trying to account for those cases where expressives don't receive speaker-oriented interpretations. For example, *bastard* in (11) seems to convey the father's point of view rather than the speaker's:

(11) My father told me not to marry that **bastard** Webster.

Immediacy points to the performative character of expressivity: expressives, like 'performative' speech acts (e.g. marrying, baptizing, promising, etc.) immediately update the common ground without the mediation of a proposal. However, the utterance of a expressive may be better characterized as a display rather than as a performance: whereas the effects of performative speech acts can be undone if their illocutionary pre-conditions are not met (e.g. if, after a religious wedding, it is discovered that the priest wasn't officially ordained), expressive's effects on the context cannot be undone. To wit, uttering *fucking Jones* or *Jon is a Spic* displays hostility even if it is later known that the speaker didn't feel such hostility.

Descriptive ineffability refers to the idea that expressives are difficult, if not impossible, to paraphrase in non-expressive terms. However, the status of

ineffability as a diagnosis of expressive content is unclear. As Geurts (2007) observes, even though emotive information such as *fucking*, insults such as *bastard* or slurs such as *Spic* are difficult to paraphrase in descriptive terms, the same difficulty can be found in average truth-conditional expressions such as *the* or *green*. Therefore, descriptive ineffability can't be based on individual reports about how to conceptualize or define a particular expression. Instead, we could pin down descriptive ineffability by observing data about the difficulties to acquire expressive words: for individuals learning a new language, being competent with the contexts in which it is correct to use terms such as *the* or *green* is considerably less hard compared to learning the contexts in which to use emotive expressions such as *fucking*, due to the open-ended contextual factors that are involved in the latter's utterance.

Finally, it has also been observed that curse words are underspecified with respect to the affective interpretations they can receive. As McCready (2012) observes, the emotions that expletive adjectives like *fucking* end up conveying depend on various contextual factors. In (12a), *fucking* can be interpreted as positively valenced (e.g., if the speaker is a known supporter of Berlusconi), or as negatively valenced (e.g., if the speaker is known to be a critic of Berlusconi). Second, particularistic insults like *bastard* can also have affectionate uses, particularly when the speaker and addressee are close acquaintances (or presume to be so, as the advertisement in 12b illustrates). Third, it has been observed that slurs are interpreted radically differently when the speaker and addressee belong to the group denoted by the slur or when they don't: whereas in the first case the slur aims at consolidating or forge a solidarity relation, in the second it is used to express contempt towards the group targeted (12c):

(12) a. **Fucking** Berlusconi wants to be President of Italy.
b. Here's To You, Ya **Bastard**! You've been such a good friend to me through the years. I'm so grateful.

c. Me encanta la musica **sudaca**.
'I love South-American-PEJ music.'

Affective underspecification, as we will see in Section 2.2.3, will constitute one of the main challenges for building a semantic theory for curse words. Broadly speaking, the interpretation of a curse words depends on what other properties are known or assumed about the speaker. As we observed above, in our discussion of indexical expressivity, sociolinguistic variables such as *-ing* or *-in'* are also interpreted with respect to what it is known about the speaker in the utterance context, so there seems to be a common core between these two phenomena that is worth exploring.

Secondary properties

Repeatability, originally listed in Potts (2004), refers to the idea that the repetition of an expressive item heightens its effect. In the (13b), rather than being redundant, the repetition of the expletive adjective *fucking* conveys a higher degree of emotional intensity:

(13) a. **Fucking** Jones is coming.
b. **Fuck! Fucking** Jones is **fucking** coming.

However, repeatability doesn't seem ubiquitous in the expressive domain. Indeed, it isn't clear whether saying *Spic* or *bastard* repeatedly in a single utterance displays a stronger level of derision compared to using it once. Therefore, we may consider repeatability as a secondary property, as it only applies to expletive adjectives.

Non-local interpretations, also noticed in Potts (2004), points to the fact that expletive adjectives such as *fucking* often have a syntactic realization that differs from their scope of semantic interpretation. For example, in (14a),

fucking seems to apply to the sentence's subject despite its nominal internal position in the syntax. However, notice that this property is also exclusive of expletive adjectives. Other types of curse words, such as particularistic insults, (e.g., *crappy*), don't seem to have the kind of flexibility (Gutzmann, 2019):

(14) a. The dog is sleeping in the **fucking** couch.
→The speaker feels negatively about the dog.
b. The dog is sleeping in the **crappy** couch.
↛The speaker feels negatively about the dog.

To account for non-local interpretations, Frazier et al. (2014) propose that, appearances notwithstanding, expletive adjectives are not syntactically integrated in their host utterance. Rather, according to the authors, expletive expressives constitute independent speech acts that indicate the speaker's heightened emotional states. Frazier et al. (2014) view and non-local readings of expletive adjectives in general will be analyzed in Chapter 4.

2.2.2 Zooming in

Independence and not-at-issueness

In Potts (2004) original view, independence is a property that not only applies to expressives, but in general to different parenthetical expressions that, intuitively, introduce 'not-at-issue' content, that is, side-lined comments on the main point asserted by the rest of the utterance. To illustrate this, consider the following examples, which include (underlined) a non-restrictive relative clause (NRRCs) and a non-restrictive adjective (NRAs):

(15) a. I will visit my father, <u>who is sick</u>, in the hospital.
b. I will visit my <u>sick</u> father in the hospital

→I will visit my father in the hospital. *at-issue content*
→My father is sick. *not-at-issue content*

Even though the underlined expressions contribute truth-conditional content (i.e., the proposition that the speaker's father is sick), they systematically fail to interact with truth-conditional operators (e.g., negations, conditionals, disjunctions, etc.). To wit, the proposition expressed by (16a) comes out as false if the speaker visits his father in the hospital, independently of whether the father is sick or not. Similarly, the proposition expressed by (16b) comes out as false if the speaker visits his father but doesn't give him chocolates, independently of whether the father is sick or not:

(16) a. I will not visit my sick father in the hospital.
b. If I visit my father, who is sick, in the hospital, I will give him chocolates.

Therefore, given that parenthetical and expressive information are independent, it has been assumed that expressive information can be conceptualized as not-at-issue content. However, such assumption raises the following questions:

A Does the at-issue/not-at-issue divide map the propositional/expressive divide?

B Does not-at-issueness captures the informational status of expressives?

Let us address question A. To begin with, note that the content associated with the parenthetical expressions in (15) is not 'inherently' but only 'circumstantially' not-at-issue. As (17) illustrates, if the proposition associated with the parentheticals in (15) is instead expressed by a stand-alone utterance, it no longer has a not-at-issue status. Therefore, propositional information

obtains at-issue or not-at-issue status depending on how it is expressed:

(17) My father is sick. I have to visit him in the hospital.

Now, is expressive content also at-issue or not-at-issue depending on how it is expressed? At first sight, it seems that the emotive information associated with expletive adjectives like *damn* is only circumstantially not-at-issue. As (18b) illustrates, if we use the expletive adjective *damn* as a stand-alone utterance (e.g., *Damn!*), then it no longer appear to be a side-lined comment on some other proposition:

(18) a. I have to visit my **damn** father in the hospital.
b. **Damn!** I have to visit my father in the hospital.

However, it is unclear whether the stand-alone utterance of *Damn!* in (18b) expresses at-issue content or not. To wit, *Damn!* seems to denote a function that takes the follow-up sentence as argument and thus expresses the speaker's negative attitude towards the proposition expressed by it, i.e., the fact that he has to visit her father in the hospital. If such case, *Damn!* still fails to acquire an at-issue status despite appearing as a stand-alone expression. In that sense, we can be tempted to consider that expressive content is not-at-issue by 'by default' rather than circumstantially.

Yet, does not-at-issueness by default captures the informational impact of expressives? Let us now address question B. It is worth noting that there are many possible ways to define what is (not-)-at-issueness, so there are many ways to answer question B.

On one view, the at-issue/not-at-issue distinction captures how different pieces of information are packed together. As Abrusán (2011) observes, whereas at-issue information is such that participants pay attention to it

by default, not-at-issue content remains unnoticed unless contextual factors divert our attention to them. If one favours this way of conceiving (not-)at-issueness, then expressives will be more likely conceived as at-issue rather than not-at-issue. For example, by uttering the following sentence,

(19) Jones is **fucking** cooking.

the audience is unlikely to overlook or pay less attention to the inference triggered by *fucking* (i.e., that the speaker feels hostility towards Jones) than to the proposition that John is cooking, thus undermining the basic assumption that expressives provide not-at-issue content by default.

On another, more technical view, the at-issue/not-at-issue distinction captures how information behaves with respect to different linguistic environments, such as i) Questions Under Discussion (QUD), ii) direct denials and iii) rhetoric relations (Koev, 2018). In this framework, at-issue information is assumed to be able i) to provide a complete answer to the current QUD, or ii) to be available for direct denials, or iii) to be able to establish rhetoric relations, or else is not-at-issue. Now, can expressive information provide an answer to a QUD? It doesn't seems so. In (20), *Damn!* doesn't constitute neither a partial nor complete answer to the question made by B. Therefore, the content associated with *Damn!* is diagnosed as not-at-issue according to the first test.

(20) A: How are you feeling with respect to your father?
B: (??)Damn! I have to visit my father in the hospital.

With respect to the second test, i.e., direct denials, the outcome is similar. As illustrated by (21), the content associated with expressives such as *fucking* are not amenable for direct responses, thus indicating that they are not-at-

issue according to the second test:

(21) A: **Fucking** Jones is coming.
B(1): That is not true, he is not coming.
B(2): #That is not true, you don't feel hostility towards him.

According to the third test, a proposition is at-issue if a fresh uttered segment can attach to it by some appropriate coherence relation. Roughly speaking, this means that some clause is considered to provide at-issue content if the utterances coming after develops it (e.g., 'Jones is coming. He is staying for a bit.') or explains it (e.g., 'Jones is coming. He left his keys.'). Now, as (22a) shows, expressives don't seem able to establish rhetoric relations of the former type, i.e., don't seem able to further develop what has been previously said. However, as (22b) shows, they seem able to establish rhetoric relations of the latter type. In (22b), the follow-up sentence explains the utterance of *Fuck!*, thus establishing a rhetoric relation:

(22) a. I hit my thumb. **Fuck**!
b. **Fuck**! I hit my thumb.

The conclusion of this discussion is that the relation between expressive and not-at-issue information is not as straightforward as it is often assumed. If one considers that expressive information constitutes not-at-issue content, one commits to expressive information being not-at-issue 'by default' rather than pragmatically. Then, if one considers expressive information as not-at-issue by default, one cannot spell the at-issue/not-at-issue distinction in terms of how information is hierarchically organized in an utterance, but only in terms of how expressive content behaves with respect to various linguistic tests. And, finally, if one considers expressive information as not-at-issue in virtue of its behavior with respect to linguistic tests, one should acknowledge

that expressive is nonetheless diagnosed as at-issue according to the third test mentioned above, i.e., that expressives are able to establish rhetoric relations with new discourse segments.

Perspective-dependence and the semantic judge

Expressives are dependent on the speaker's perspective or viewpoint. By uttering

(23) Peter said that **fucking** Jones is coming.

bystanders come to know that the speaker is probably upset with Jones, even though the expressive occurs in the syntactic scope of a speech-report predicate.

However, there are exceptions to this phenomenon. That is, cases where expressives don't receive speaker-oriented interpretations. In the following examples, which are inspired on Kratzer (1999), *fucking* seems to convey the father's point of view rather than the speaker's, both in situations where the expressive occurs embedded (24a) and unembedded (24b):

(24) a. My father screamed that he would never allow me to marry that **fucking** Webster.
b. My father was always upset with Webster. **Fucking** Webster would marry her daughter soon.

To account for not speaker-oriented interpretations, Potts (2007a) proposes to link expressives to a contextual 'judge'. The idea of a judge comes from Lasersohn (2005) relativistic treatment of predicates of personal taste (PPTs), that is, predicates like *fun*, *horrible* or *delicious*. In this theory, PPTs are interpreted with respect to a judge parameter, which is typically

the speaker unless contextual factors make another perspective more salient. The application to expressives is straightforward: in (24), since it is unlikely that the speaker intends to express a negative attitude towards Webster, the value of the judge parameter is set to the speaker's father.

An issue for this strategy is that expressives and PPTs don't pattern with respect to perspective shifts (Hess, 2018). In speech reports, the perspective of a PPT (e.g., *horrible*) is that of the subject of the matrix clause. In contrast, expressives (e.g., *fucking*) have a strong bias to be interpreted as conveying the speaker's perspective. Even though we can imagine situations in which a speaker-oriented reading of (25a) is acceptable, or a non-speaker oriented of (25b) is acceptable, when uttered out of the blue these sentences receive drastically different interpretations:

(25) a. John said that he saw a **horrible** dog outside.
b. John said that he saw a **fucking** dog outside.

Another concern is that similar changes of perspective are also observed in items that are not usually classified as 'perspective-dependent'. For example, imagine a situation in which it is known that the speaker thinks that their uncle's home-made pizza is in fact a focaccia with extra ingredients. In that situation, the credence that the uncle's dish is a genuine pizza is attributed to the uncle rather than to the speaker:

(26) a. My uncle said that he prepares the best **pizza** in Milan.
b. My uncle thinks he is the best chef. The **pizza** he prepares is indeed unique.

Then, if perspective-shift is not exclusive of perspective-dependent expressions, it may be important to re-consider whether the semantic relativization

of expressives to a contextual judge is necessary, or if a more general pragmatic explanation can be proposed instead.

Yet a deeper problem with this relativization to a judge parameter is that the function of a 'perspective' in the case of PPTs is drastically different from its function in the case of emotive expressives. For example, in qualifying rollercoasters as fun, the semantic judge determines what features (or 'dimensions') of rollercoasters make them fun or not (e.g., speed, danger, etc.), how much weight each of those features have, and which is the threshold of *fun* in a scale derived from such dimensions. In contrast, even though expressives such as *fucking* in 'fucking rollercoaster' may trigger the inference that the speaker evaluates rollercoasters in a certain way, their use doesn't seem to require a scale in which rollercoasters can be ordered according to whether they count as a 'fucking' rollercoasters or not.

Instead, we can understand perspective-dependence as pointing to the fact that expressive content is always *about* someone, typically the speaker. Bracketing those contexts in which expressives receive not speaker-oriented interpretations, expressives' subject matter is the emotional state of the speaker in the utterance context. That is, emotive expressives would be not much 'dependent' on the speaker's viewpoint, but about the speaker's viewpoint. Comparing expressives with physical gestures may make this point clearer: even though uttering 'Jones' with a contemptuous [elated] facial expression may trigger the inference that the speaker assesses Jones negatively [positively], this doesn't make facial gestures relative or dependent upon a semantic judge like that involved in the assessment of roller-coasters as fun. The speaker's facial expression displays how the speaker's feel about Jones, rather than how he evaluates him with respect to others.

In sum, we have observed that, despite the rich literature on expressive's perspective-shift, it is unclear how to understand expressive's perspective-dependence itself. The suggestion sketched in this section is that expressive's

perspective-dependence is not relative to a semantic judge (in the same way that evaluative expressions such as *fun* are), but relative to the subject matter of expressive content (i.e., typically, but not always, the individual who talks).

Affective underspecification

In the literature on expressive meaning, it is argued that curse words are conventionally associated with affective states (Potts, 2004). Thus, to represent their meaning, curse words are associated with 'use-conditions' rather than truth-conditions. That is, the meaning of a curse word (e.g., *fucking*) is represented by the set of contexts in which it is appropriate or felicitous to use them, rather than by the set of worlds in which they are true (Kaplan, 1998; Gutzmann, 2015). For example, the meaning of *fucking* in (1a), here repeated, is represented as the set of contexts in which the speaker feels a negatively valenced affective state towards the dog, and infelicitous otherwise:

(27) a. 'The fucking dog is barking' is true if the dog is barking.
b. 'The fucking dog is barking' is felicitously used if the speaker has a negative attitude towards the dog.

However, it has also been observed that curse words are largely underspecified with respect to the emotions they can express (Potts, 2004; McCready, 2012). Even though the expressions in (28) display a bias towards negatively valenced interpretations, they can also express other types of affective states:

(28) a. The **fucking** dog started to bark. Fortunately, he chased away the robbers.
b. Hey, **bastard**! You have been such a good friend.
c. Hay musica **Sudaca**. Suena bien.
'There is South-American-PEJ music. It sounds nice.'

The expletive adjective in (28a) can be interpreted as indicating the speaker's alleviation that the dog was there to chase the robbers. The particularistic insult in (28b), despite its negative connotations, can be interpreted as friendly, playful and non-face threatening when used in certain contexts. Finally, the slur in (28c), which typically express contempt towards South-Americans, can also be used by the members of the group derogated without expressing that the speaker considers South-Americans to be lesser than others.

We may try to solve this problem in various ways. One would be to assume that these expressions are systematically ambiguous. In such view, a term such as *fucking* would be ambiguous between positive and negative readings:

(29) a. $[\![\text{fucking(x)}]\!]_1$ = the speaker feels negatively towards x.
b. $[\![\text{fucking(x)}]\!]_2$ = the speaker feels positively towards x.

However, this proposal is problematic for various reasons. One of them is that *fucking* can also indicate neutrally valenced affective states such as surprise. In the following example, *fucking* expresses the speaker's perplexity about the weather, without necessarily evaluating it positively or negatively:

(30) It is **fucking** sunny and rainy outside!

Another problem is that this solution cannot be easily extended to insults like *bastard* and slurs like *Sudaca*. To wit, positive interpretations of these expressions are only possible when the speaker and target are close friends (in the case of insults) or belong to the group derogated (in the case of slurs). Other ambiguous expressions (e.g., *bank*) do not place restrictions on who can use them to express one or the other of their meanings (Ritchie, 2017).

A different solution would be to assume that expressives are not conventionally associated with positive or negative evaluations, but with a certain

degree of arousal (e.g., excitation, energy) (Potts, 2004; Potts and Schwarz, 2008). In this framework, an adjective like *fucking* would invariably indicate that the speaker is in a high degree of arousal, independently of whether such arousal is also positively or negatively valenced:

(31) $[\![\text{fucking(x)}]\!]^c$ = the speaker feels in a heightened emotional state towards x at c.

Yet, replacing valence by arousal is also problematic. To wit, even though expletive adjectives like *fucking* are typically interpreted as indicating a high degree of arousal, it is possible to use them in contexts where the speaker doesn't feel very excited. In the following examples, assuming the intonation is flat, *fucking* doesn't need to be interpreted as indicating that the speaker feels excited at the utterance context, but only that he feels negatively:

(32) a. I've done everything to reach this point and now that I'm here, I'm fucking bored.
(from the film 'Mr. Nobody', 2009).
b. I'm fucking bored, man. There ain't shit to do on this bus.
(from 'Jay And Silent Bob Strike Back').

Moreover, other expressive items, such as slurs, don't seem to require that the speaker is feeling a heightened emotion to be felicitously uttered. To wit, slurs are part of the ideolect of the racist or homophobe, not only when he is excited or in the grip of anxiety, but in general.

In Chapter 3, I propose to solve this problem by postulating that emotive expressives are indexically (rather than conventionally) associated with various affective dimensions simultaneously, namely, valence, arousal and dominance. In this proposal, underspecification is resolved by us, hearers, by reasoning

about what we know about the speaker and what the 'normal' interpretation of the expressive item is. In other terms, I will defend the idea that affective expressivity is determined contextually, by the interplay between listener's and speaker's expectations about each other.

Notice that affective underspecification is not exclusive of expressive items, but can also be found among predicates of personal taste. As Stojanovic and Kaiser (2022) point out, some perspective-dependent predicates are underspecified with respect to the polarity of the evaluations they may give rise to. As the follow-ups in (33a-c) and (34a-c) illustrate, some predicates of personal taste (e.g., *intense*, *surprising*, *interesting*, etc.) can be interpreted in different ways depending on the context:

(33) The film was intense

a. ... and I like it.
b. ... and I don't like it.
c. ... and I don't like it nor dislike it.

(34) His CV is interesting

a. ... In the good way.
b. .. In the bad way.
c. ... but I cannot tell whether it is good or bad.

In these cases, the affective message of the predicate is recovered by reasoning about the speaker's preferences, the object of the predication, the utterance context, etc. Even though I will only focus on affective expressions as primarily exemplified by curse words, we should consider this problem as pervasive in perspective-dependent and, in particular, affective communication.

2.3 Previous theories of expressive meaning

Intuitively, sentences can carry different types of content simultaneously, including inferences about the speaker's feelings at the utterance context. Observe (35):

(35) Most of the damn students, those video-game addicts, failed the exam.

a. ENTAILMENT: Most of the students failed the exam.
b. PRESUPPOSITION: There is a unique group of students.
c. IMPLICATURE: Not all students failed the exam.
d. PARENTHETICAL: The students are video-game addicts.
e. ?: The speaker is upset with the students.

What type of content is (35e)? First, let me briefly describe each type of content in (35a-d), and then present the theories that associate expressive meaning with one of these categories.

Entailments amount to what is asserted, that is, what is added to the common ground if the conversational participants don't object to it. If no one objects to (35), then it becomes common ground that most of the students failed the exam. In contrast, presuppositions amount to what should be already contained in the common ground in order to interpret the utterance. If (35b) is not assumed to hold, i.e., if it is not considered true that there is a unique group of students, we cannot evaluate (35) as true or false. Importantly, the hallmark of presuppositions is that they are projective. That is, that they scope out from entailment-cancelling operators, including negations, modal operators, if-clauses or question-marks. In (36), all of the utterances trigger (35b) the presupposition that there is a unique group of students, but not (35a) the entailment that most of them failed the exam:

(36) a. Most of the damn students, those video-game addicts, didn't failed the exam.
b. It is possible that most of the damn students, those video-game addicts, failed the exam.
c. If most of the damn students, those video-game addicts, failed the exam, then the teacher will talk to their parents.
d. Did most of the damn students, those video-game addicts, failed the exam?

In contrast to entailments or presuppositions, implicatures are not linked to the conventional meaning of the expressions in the sentence. Instead, they are inferences calculated on the basis of rational principles of cooperation. Assuming that the speaker of (35) is informative, we can infer that he is no position to utter the most informative utterance, namely, '*All* of the damn students failed the exam'. As this sentence must be false, the reasoning follows, we obtain the implicature in (35c). Importantly, since implicatures are not lexically triggered, they cannot be considered to project (or fail to project) in the standard sense, that is, to survive (or be blocked) by the effect of entailment-cancelling operators.

Now, parenthetical content (also called 'conventional' implicatures, following Potts 2004), are triggered in virtue of the conventional meaning of an expression. The inference in (35d) would not exist if the parenthetical expression (i.e., 'those video-game addicts') were removed from the sentence. Moreover, like entailments, parenthetical content introduces new information into the common ground. If the context already entails that the students are video-game addicts, then uttering the parenthetical in (35) would sound repetitive.

(37) The students are video-game addicts. ??Most of the damn students, those video-game addicts, failed the exam.

However, like presuppositions, parenthetical expressions are typically projective. As the sentences in (36) illustrates, the parenthetical 'those video-game addicts' triggers the projective inference (35d) that the students are video-game addicts, despite the presence of entailment-cancelling operators.

Where does affective content fit in this taxonomy? It has been proposed that affective content is a particular type of parenthetical expression, presupposition or implicature. We will turn to these theories in Section 2.3.1, 2.3.2, and 2.3.3. Then, I will discuss a theory that considers expressivity to constitute a distinguished semantic category in Section 2.3.4.

2.3.1 The parenthetical approach (Potts, 2004)

Potts (2004) develops a multidimensional semantics designed to capture how conventional implicatures (CIs) contribute to the overall meaning of the utterances triggering them. In Potts (2004), CIs are defined as inferences triggered by linguistic expressions but which don't contribute to the utterance's at-issue content. In particular, CI-triggers include supplemental expressions (e.g., non-restrictive relative adjectives such as *sick* in 'my sick father') and expressives (e.g., expletive adjectives such as *fucking* in 'my fucking father'). As (38) illustrates, these expressions introduce content that is independent of the content conveyed by the host utterance:

(38) a. I will visit my **sick** father.
→I will visit my father. *At-issue inference*
→My father is sick. *CI inference*

b. I will visit my **fucking** father.
→I will visit my father. *At-issue inference*
→The speaker feels upset with their dog. *CI inference*

To give a compositional semantics for CI-triggers, Potts (2004) develops the

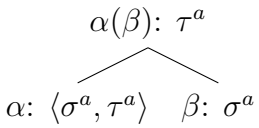
system $\mathcal{L}_{CI}$. The first innovation of $\mathcal{L}_{CI}$ consist in the distinction of at-issue and CI content at the level of types (see 39a and 39b). The second innovation is that it adds a clause that restricts how CI types may combine with at-issue types to form complex expressions (see 39d). As it can be observed, the system specifies that at-issue content can serve as argument to either at-issue (see 39c) and CI expressions (see 39d), but that CI content cannot be argument for an at-issue nor other CI expressions:

(39) Types for $\mathcal{L}_{CI}$ (simplified)

a. e^a and t^a are basic at-issue types.
b. e^c and t^c are basic CI types.
c. If σ and τ are at-issue types, then $\langle \sigma, \tau \rangle$ is an at-issue type.
d. If σ is a CI type, and τ an at-issue type, then $\langle \sigma, \tau \rangle$ is a CI type.
e. The set of types is the union of the at-issue and CI types.

The third innovation of $\mathcal{L}_{CI}$ consist in its 'tree-admissibility conditions'. These conditions regulate how expression of the various types should combine with each other during the semantic derivation in order to be well-formed. On the one hand, we have at-issue application, whose major difference with standard functional application is that it indicates that we are dealing with at-issue expressions.

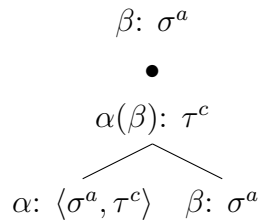
(40) At-issue application



On the other hand, we have CI application, which indicates that the output of the application of a CI-typed expression to an at-issue-typed expression

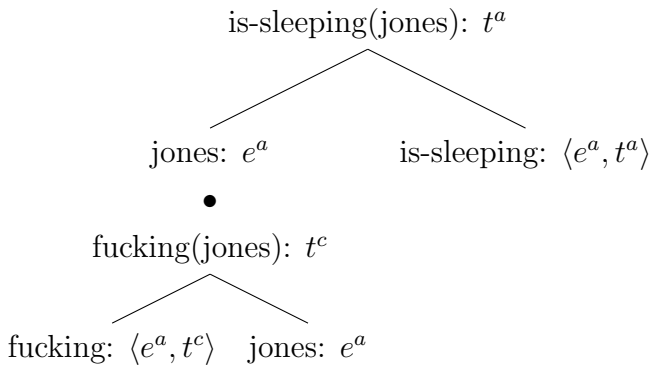
is both the functional application and the at-issue argument, which remains available for further derivations. Both types of content are isolated in the parse tree by means of the metalogical bullet '•', which distinguishes independent contents at the same node.[2].

(41) CI application



Consider the following example of CI application. In the derivation, the CI content (i.e., 'fucking(jones): t^c') is left behind during the semantic composition and the descriptive content (i.e., 'jones: e^a') is returned unmodified, thus remaining available to participate in further derivations:

(42) Fucking Jones is sleeping



[2]A third innovation of $\mathcal{L}_{CI}$, which we wont discuss here, is the 'parse-tree interpretation', according to which the denotation of a sentence is given by the interpretation of an entire semantic tree instead of just a single formula. However, we won't discuss this feature of the model in what follows.

$\mathcal{L}_{CI}$ achieves its goal of representing CI content (and thus expressives) as independent, that is, as content that cannot scopally interact with truth-conditional operators. Indeed, the fact that CI content is independent falls out from the the restrictions on well-typed expressions and well-formed combinations specified in (39): a semantic operator would need to have a type that takes CI content as input, which is not allowed by the definitions in (39).

The criticisms of $\mathcal{L}_{CI}$ mainly focus on the fact that is that it is too restrictive, as it doesn't include, e.g., 'hybrid expressives', that is, items that contribute to both dimensions of meaning at the same time (such as slurs). To solve this limitations, extensions of $\mathcal{L}_{CI}$ have been proposed, such as those in McCready (2010) and Gutzmann (2015). However, in what follows, I will focus on problems which have less to do with the restrictions imposed by the syntactic rules of $\mathcal{L}_{CI}$ than with its predictions about the interpretation of curse words.

The first observation is that $\mathcal{L}_{CI}$ assumes that expressive content is not-at-issue, i.e., information packed in a sidelined way. However, as mentioned in Section 2.2.2, the qualification of expressive content as not-at-issue doesn't help much in understanding its informational impact within a context. To wit, if we assume that the at-issue/not-at-issue distinction captures how information is hierarchically organized in an utterance, then it is unclear whether the affective information expressed by terms like *damn* or *Spic* can be considered less salient than others expressed by utterance.

The second observation is that $\mathcal{L}_{CI}$ expressives and parenthetical expressions don't contribute the same kind of content. To wit, the adnominal modifier *black* introduces the proposition that the speaker's dog is black, which can be evaluated as true or false according to the context. In contrast, the

expletive adjective *fucking* cannot be represented as propositional, as its affective content cannot be evaluated as true or false. Therefore, $\mathcal{L}_{CI}$ doesn't distinguish between side-lined truth-conditional expressions from expressive items.

(43) a. My black dog is coming.
b. My fucking dog is coming.

The difference between side-lined truth-conditional expressions and expressives can be observed in the ways they update the common ground. As the following exchanges illustrate, speech-act participants have the possibility to object to the content introduced by parenthetical expressions, but not to the content introduced by expressives:

(44) a. A: My black dog is coming.
B: Hey, wait a minute, your dog is not black.
b. A: My fucking dog is coming.
B: (??)Hey, way a minute, you are not upset with the dog.

Finally, $\mathcal{L}_{CI}$ remains silent about many features of expressive affective meaning, such as the fact that expressives can be interpreted as conveying different a wide array of emotions depending on the context of interpretation (see Section 2.2.2).

2.3.2 The presuppositional approach (Schlenker, 2007)

Schlenker (2007) argues that expressives are presupposition triggers. Roughly speaking, presuppositions are parts of the meaning of a linguistic expression that are marked as taken for granted by the speech-act participants. As observed above, presuppositions must be entailed by the utterance context

in order for the utterance to be felicitous. According to Schlenker (2007), expressives trigger a particular kind of presupposition, namely, a presupposition that is indexical (i.e., evaluated with respect to the utterance context) and attitudinal (i.e., that predicates something about the speaker's mental states). As (45) illustrates, presupposition triggers such as *knows* and emotive expressives such as *fucking* can be analyzed as triggering inferences that must be true in the utterance context in order for the host utterance to be felicitous:

(45) a. Mary **knows** Jones will visit my father.
→Mary beliefs Jones will visit my father. *asserted content*
→Jones will visit my father. *presupposition*

b. **Fucking** Jones will visit my father.
→Jones will visit my father. *asserted content*
→The speaker feels upset with Jones. *presupposition*

In order to capture expressive's presuppositional character, Schlenker (2007) gives expressive rules of context update such as the following:

(46) $[\![\text{fucking(x)}]\!]^{c,w} \neq \#$ if the agent of c feels negatively with respect to x in the world of c. If $\neq \#$, $[\![\text{fucking(x)}]\!]^{c,w} = x^{c,w}$.

Yet, since expressives normally contribute new information, they don't seem to pattern with presuppositions, which is information usually taken for granted. To solve this, Schlenker (2007) claims that expressive's presuppositions are systematically accommodated by the speech-act participants in virtue of the speaker's authority about their own mental states. In other terms, it would be senseless to challenge the content triggered by 'fucking Jones' (i.e., that the speaker feels upset with Jones) because it is information to which only the speaker has a privileged access.

As we can observe, Schlenker (2007) maintains Potts (2004)'s assumption that expressives contribute not-at-issue information, but formalizes it in a unidimensional framework. That is, expressives are analyzed as triggering truth-conditional content that is automatically accommodated in the common ground in virtue of the speaker's authority about their own mental states. In that sense, this proposal avoids the postulation of a separate meaning dimension, with the corresponding addition of new semantic types and rules of composition.

The first observation about the presuppositional account is that, even though both presuppositions and expressive inferences are projective, they don't project in the same way. In some situations, presupposition triggers can scopally interact with truth-conditional operators. For example, the possessive noun phrase 'Jone's son' doesn't project that John has a son when it occurs in a conditional whose antecedent already introduces the presupposition:

(47) a. Jone's son is probably sleeping.
$\rightarrow$Jones has a son.
b. If Jones has a son, Jone's son is probably sleeping.
$\not\rightarrow$Jones has a son.

In contrast, the expressive *fucking* is immune to such kind of binding. As (48b) illustrates, *fucking* still triggers the inference that the speaker is upset when it occurs in the consequent of a conditional (McCready, 2019):[3]

(48) a. Fucking Jones is probably sleeping.
$\rightarrow$The speaker is upset with Jones.

[3]However, as McCready (2019) points out, these examples are problematic because expressive content is descriptively ineffable, so any paraphrase of the expressive *fucking* that we may use in the antecedent of a conditional won't be enough to bind the expressive that occurs in the consequent.

b. If I am upset with Jones, fucking Jones is probably sleeping.
→The speaker is upset with Jones.

Another observation about the presuppositional account is that it also assumes that expressive content can be evaluated as true or false. Under this approach, an expressive (e.g., *damn*) is felicitously uttered if its presupposed content (e.g., that the speaker is upset) is true at the utterance's context. Thus, the presuppositional account doesn't distinguish between truth-conditional presupposed contents from expressive contents.

A supplementary issue with the presuppositional account is that it seems difficult to extend to the case of other kinds of expressive meaning, e.g., social meaning. In the presuppositional account, the meaning of the French second person pronoun *tu* is represented as follows:

(49) $[\![\text{tu}]\!]^{c,w} \neq \#$ if the agent of c believes in the world of c that they stand in a familiar relation to the addressee of c. If $\neq \#$, $[\![\text{tu}]\!]^{c,w} =$ the addressee of c.

As mentioned, Schlenker (2007)'s account is based on the idea that expressives trigger inferences that are systematically accommodated in the common ground because they are linked to the speaker's mental states. However, as observed in Section 2.1, honorifics convey information about the speaker's social relation to the addressee, rather than about the her mental states. Thus, if we specify the honorific update rules of *tu* as being about the speaker's social situation, then it becomes again unclear in which sense honorifics can be automatically accommodated in the common ground.

2.3.3 The implicature approach (Hom, 2012)

Hom (2012) notices that Potts (2004)'s account of terms like *fucking* cannot

be extended to curse verbs such as *to fuck up*. In atomic sentences like (50a), *to fuck up* triggers the inference that the speaker is upset with John. However, as the felicity of the follow-up in (50b) indicates, such inference disappears in conditional sentences:

(50) a. John fucked up another case.
→The speaker is upset with John.
b. If John fucks up another case, he will fired for it. (But I don't think he will because he is working much harder now.)

To propose a unified account of both expressive adjectives like *fucking* and verbs like *to fuck up*, Hom (2012) argues that, at the semantic level, they are descriptive terms: *fucking* refers to impermissible sexual intercourse (and, by the same token, *damned* to condemnation, *bloody* to being full of blood, etc.). Thus, in this account, the sentence in (51a) would be semantically equivalent to the one in (51b):

(51) a. John fucked up another case.
b. John had impermissible sex with another case.

However, the argument follows, since the reading in (51b) is likely to be considered odd in the utterance context, the speaker will be considered to have violated a conversational maxim in order to get the audience to understand the he has a 'extreme (affective) reaction' with respect to the situation described in the utterance (Hom, 2012, p. 399). This reaction would have the same degree of intensity as the negative attitudes culturally associated with pre-marital sex in the Western World. As a result, *fucked up* isn't interpreted literally and instead triggers an inference about the speaker's emotions.

As other conversational implicatures, Hom (2012) claims that expressive im-

plicatures can be cancelled. In (52), the speaker makes it clear that *fucking* is being used descriptively, thus blocking the affective inference:

(52) John is in the fucking couch. [Literally. I am not upset or surprised; John is actually in the couch where impermissible sexual intercourse takes place].

A first observation about this theory is that examples like (52) are problematic because expressive content is descriptively ineffable, so propositional paraphrases such as 'I am not upset or surprised' won't provide clear evidence about whether expressive's content can be cancelled or not.

A second observation of this theory is that, if it were the case that expressives like *damned* or *fucking* trigger expressive inferences because they are endowed with ideologies imposed by social institutions, then they would be able to trigger expressive inferences independently of the syntactic positions in which they occur. However, this is not what we observe. As (53b) illustrates, when *damned* occurs as a defining relative clause, it can only be interpreted literally, i.e., as communicating that the dog is condemned. This indicates that there is more than a pragmatic reasoning behind the triggering mechanisms of expressive content:

(53) a. The damned dog is on the couch.
b. ??The dog that is damned is on the couch.

Finally, this account cannot be easily extended to other expressive terms. For example, under this framework, slurs that derogate the same social group (e.g., the n-word and *spook*) would be predicted to express the same types of derogatory attitudes towards the group derogated. However, this is not what we observe: the n-word expresses a more negative evaluation of its target than

spook, even though both are associated with the same racial ideologies.

2.3.4 The context-update approach (Potts, 2007a)

In Potts (2007a), a new version of $\mathcal{L}_{CI}$ is developed in order to capture the semantic properties of expressives. In contrast to its predecessor, this new system uses a new expressive type ϵ, which denotes attitudinal relations between individuals. These attitudes are represented by real numbered intervals I $\sqsubseteq$ [-1, 1], according to their valence: while intervals (¡ 0) indicate negative emotions, intervals (0 ¡) indicate positive emotions. Moreover, these intervals relate two individuals and thus have the form 'aIb', which indicates the orientation of the expressive (i.e., that a the attitude I towards b):

(54) a. a[-1, -.9]b (a feels very negatively with b)
b. a[.6, 1]b (a feels very positively with b)

Accordingly, emotive expressives are represented as denoting intervals such as the following, where s represents the speaker and x the individual the attitude is about. Notice the interval associated with *fucking* is narrower than that associated with *damn*. Under this framework, this represents the fact that *fucking* conveys a stronger emotion than *damn* (Potts, 2007a, p. 20):

(55) a. $[\![\text{damn(x)}]\!] = \lambda\text{x.s[-.7, -.3]x}$
b. $[\![\text{fucking(x)}]\!] = \lambda\text{x.s[-1, -.8]x}$

In order to capture the effect of these denotations, Potts (2007a) proposes that the context includes a parameter c_ϵ, which is a set of all indices of the form aIb. This parameter keeps track of all the attitudinal relations between the individuals in the discourse domain. After an expressive such as *fucking* is uttered, the context is directly updated. This update can happen in two

ways: if c_ϵ does not contain any index of the form aI'b, then $c'_\epsilon = c_\epsilon \cup \{\text{aIb}\}$, and (ii) if it does contain such an index of the form aI'b, the aIb replaces aI'b, where it is also required that I $\sqsubseteq$ I':

(56) $c_\epsilon \approx c'_\epsilon$ iff c_ϵ and c'_ϵ differ at most in that

a. aIb $\in c'_\epsilon$; and

b. if c_ϵ contains an expressive index aIb, where I $\neq$ I', then aIb $\notin c'_\epsilon$ and I $\sqsubseteq$ I' (Potts, 2007a, p. 11).

In that sense, this framework formalizes the idea that an expressive's utterance directly displays how the speaker feels with respect to other individuals within the discourse domain. Thus, when 'fucking Jones' is uttered, it has two effects. On the descriptive dimension, it takes a descriptive argument (i.e., 'jones: e') as input and pass it unchanged for further derivations (as in $\mathcal{L}_{CI}$). However, on the expressive dimension, it alters the expressive index c_ϵ, outputting a context c'_ϵ which is just like c_ϵ except that a (the speaker) is experiencing the relevant negative feeling [-1, -.8] towards b (the target) in c'_ϵ. As in $\mathcal{L}_{CI}$, both effects are isolated using the metalogical bullet '•':

(57) $[\![\text{fucking}]\!]^c \bullet [\![\text{Jones}]\!]^c = [\![\text{Jones}]\!]^{c'}$, where c' is just like c except that speaker[-1, -.8]Jones

This treatment of expressives inherits the main advantages of $\mathcal{L}_{CI}$, namely, the compositional isolation of expressive effects in a different dimension. Moreover, the use of a dedicated expressive type ϵ denoting attitudinal relations rather than sidelined propositions allow us clearly distinguish between, on the one hand, the not-at-issue inferences triggered by parenthetical expressions and the contextual updates triggered by expressives. Thus, the account provides a solid foundation to understand affective expressivity.

However, the account also faces some limitations. First, as we observed in Section 2.2.2., expressives' can receive different interpretations depending on the utterance context. Thus, contrary to what (55a) indicates, *damn* can also be interpreted as positively valenced (e.g., 'the damn mathematician was brilliant!') and even as neutrally valenced (e.g., 'the damn weather is sunny and rainy!'). Moreover, even though *fucking* usually expresses stronger emotions than *damn*, this is not always the case. Depending on who utters the expressive (e.g., an unknown person), and in which circumstance (e.g., a confrontation), *damn* can convey more stronger emotions than *fucking*.

Second, the system seems too strict with respect to how an index 'aIb' can make its way into c_ε. In Potts (2007a), an index is included in c_ε only when the speaker a has uttered an expressive whose argument refers to b. As Potts (2007b) himself points out, this requirement needs to be relaxed, as non-linguistically expressed information about the speaker's emotions (e.g., gestures, tone of voice) can also have an effect on the context parameter c_ε. In the following chapter (Section 3.3), I will propose a model where an expressive index can be included in c_ε when there is evidential support that the speaker feels (or tends to feel) α with respect to b. Thus, the proposal will accommodate cases where the speaker provides non-verbal signals (e.g., smiles) about her affective states.

2.3.5 The game-theoretic approach (McCready 2012)

As we saw in Section 2.2.2, McCready (2012) observes that expressives can be interpreted in 'diametrically opposed ways', that is, as displaying positive or negative attitudes. Whereas, in (58a), *fucking* is interpreted positively, in (58b) it is interpreted negatively, as shown by the infelicity of the follow-up sentence:

(58) a. Fucking Mike Tyson won another fight. He is wonderful.

b. Fucking Mike Tyson got arrested again for domestic violence. #He is wonderful.

To analyse how expressives' underspecification is resolved, McCready proposes to use non-monotonic inference over a knowledge base, from which a 'normal' interpretation of the expressive is derived. This interpretation is based on what one would normally expect about the speaker's emotions. However, since the normal interpretation cannot be always identified with the interpretation selected by the hearer, McCready (2012) supplements the previous formalization with a game-theoretic model of how speaker and hearer's attempt to coordinate on a interpretation. In this game, the normal interpretation is used as an input that guides hearer's selection process, and thus that constraints speaker's decision about when to use the expressive without worrying about being misunderstood by the hearer.

First, how is the normal interpretation derived? McCready postulates different axiom schemas, i.e., statements representing our world knowledge, to derive the interpretation of expressives from the speaker's emotional state and, in turn, to derive the speaker's emotional state from different cues such as i) facts about the world, ii) the speaker's use of descriptive terms and iii) the speaker's use of other emotionally charged expressions (e.g., evaluatives). Then, the specific rules that instantiate these schema interact non-monotonically. For example, as domestic violence is widely seen as negative, we can infer that the speaker feels negatively about it and thus that *fucking* in (58b) expresses a negative attitude towards Mike Tyson. However, if we come to know that the speaker evaluates perpetrators of domestic violence positively, we will now infer that he feels positively and thus that *fucking* expresses a positive attitude towards Mike Tyson.

Second, how do speaker and listener coordinate on an interpretation? McCready (2012) models such interaction using signalling games (Lewis, 1979).

In this game, Nature chooses a state t from a set of states, where t represent the actual world. This set includes i) a state t_1 in which only the propositional content of the speaker's utterance is true, ii) a state t_2 in which the propositional content is true and the speaker feels positively and iii) a state t_3 in which the propositional content is true and the speaker feels negatively. Then, after observing this state, the speaker (or 'sender') chooses a message from a set of messages via a sender strategy. This set of messages includes i) a message m_1 with only propositional content (i.e., which is true in t_1) and ii) a message m_2 including propositional and underspecified affective meaning (i.e., a message which is true in either t_2 or t_3). After receiving either m_1 or m_2, the hearer selects a state t taking into account the prior probabilities that the speaker feel in a certain way or not (using the axiom schemas mentioned above), and the penalties for ignoring or misinterpreting the emotive aspects of the message received.

In contrast to the proposal we have analysed so far, McCready (2012) carefully investigates to what extent expressive's interpretation is not fixed by the context, but negotiated as part of a larger process of reasoning about the speaker's emotions and communicative intentions. Even though there is evidence that we strategically adjust our emotional signalling behavior, and that such emotional control enable us adapt to different social situations (Pollastri et al., 2018; de Melo and Terada, 2020), strategic aspects of affective communication have been usually neglected in the linguistics literature. Thus, McCready's pioneering work on expressive's interpretation will inform various aspects of the Bayesian model we will propose in the next section. Indeed, as she herself points out, her proposal can also be modelled in terms of Bayesian reasoning, where instead of deriving defeasible conclusions we obtain various conclusions that are held with different probabilities (p. 259). Now, before presenting our model, let's discuss some details in which our analyses will differ.

An general observation about the use of non-monotonic reasoning is that it assumes that some affective cues are more prevalent than others (p. 265). For example, that John's attitudes about specific individuals (e.g., John's love for his dog) would be considered more 'specific' and thus dominate over the speaker's 'global state' (e.g., John's grumpiness) in the inference of John's emotional state. Now, as we saw in the Introduction (Section 1.3), even though it has been argued that some affective cues (e.g., facial expressions) dominate over other types of affective cues (e.g., events, cultural norms, etc.), evidence shows that affective cues are rather weighed and combined optimally using statistical inference depending on each cue's reliability. This sensibility to the context can also be observed in the interaction of global states and attitudes: if John looks very upset, we will probably interpret his utterance of *fucking dog* as expressing anger rather than joy despite being common ground that he loves his dog. In this situation, John's global state is more reliable than his attitude towards the dog despite being less specific.

A first observation about the signalling game proposed by McCready is that it is based on the assumption that expressives such as *fucking* always express that the speaker is in a heightened emotional state, and thus that players only need to coordinate on the valence that this state has. However, as we saw in Section 2.2.2, even though expletives like *fucking* are typically interpreted as expressing arousing states, they may also display emotions that qualify as low in arousal (e.g., boredom, as in 'This fucking film is boring'). Now, a second observation is that the idea of a third player, Nature, who chooses a state t without any strategic concern, seems more appropriate for cases of reportative communication. That is, cases where the speaker observes a way the world is and then tries to communicate it to the hearer. Yet, in the case of affective communication, individuals have some control over how they want to be perceived by others depending on their social goals (Burnett, 2019). In other terms, curse words are not uttered to report our underlying affective states, but to 'perform' them across different social situations.

Chapter 3

A probabilistic pragmatics for affective meaning

3.1 Introduction

In Section 2.1.2 we briefly observed, first, that expressive meaning has been conceived as either conventional or indexical and, second, that even though curse words have been standardly classified as conventional expressives, they also share many features with indexical expressives. For example, affective expressives such as *damn* and sociolinguistic variables such as *-ing* and *-in* (e.g., *cooking* vs. *cookin'*) pattern along various dimensions:

1. Both affective expressives and sociolinguistic variables are associated with sets of properties of the speaker. For example, while using *-ing* typically signals social qualities like being competent but unfriendly, uttering *damn* typically signals affective states which score high on the arousal dimension but low in pleasure dimension (e.g., anger).

2. When the speaker utters an affective expressive or a sociolinguistic

variable, the properties that end up being attributed to him will depend on what other properties are believed about him. For example, in Podesva et al. (2015)'s study, it was found that the use of the /t/ release (e.g., uttering *wa*[t^h]*er* instead of *wa*[ɾ]*er*) is more associated with competence in the case of politician Condolezza Rice, but with unfriendliness in the case of Nancy Pelosi. Similarly, uttering *fucking Berlusconi* is typically interpreted negatively when it is known that the speaker doesn't like Berlusconi, but positively when it is known that the speaker likes him.

3. Speakers are aware of (1) and (2). That is, speakers are aware about these indexical associations and how their signals are constrained by what their audience assumes about them. In the study of sociolinguistic variation, there is evidence that speakers exploit this information strategically, to signal the social qualities that are the most useful for them depending on the audience (Labov, 2012). Although to a lesser extent, the psycho-linguistic literature on curse words has also pointed to how speakers inhibit or exaggerate their use of curse words depending on their audience and their social goals (Jay, 2000).

This parallelism points to a different way of understanding cursing expressions. On the one hand, cursing is considered a reaction, that is, an automatic emotional response of the speaker to a certain stimulus. On the other, in contrast, cursing can also be rational and strategic. As McCready (2012) points out, how curse words come to be (felicitously) uttered by the speaker and (accurately) interpreted by the audience heavily depends on the interplay between what the speaker intends to achieve by uttering a curse word and what the audience assumes about the speaker's emotions. Thus, a linguistic theory of cursing expressions should make explicit the link between, on the one hand, the set of qualities that a curse word stereotypically signals (i.e., its indexical associations) and what it is known or assumed in the context

about the speaker.

In this chapter, I propose a way to formally capture the interpretation of cursing expressions which highlights its context-dependent character. On the one hand, I propose that the meaning of cursing expressions is probabilistic. That is, that the meaning of a term such as *damn* is given by a probability distribution over the affective states that typically prompt their use. In particular, I associate a curse word m with the likelihood of uttering m given an affective state α (i.e., $\Pr(m|\alpha)$). On the other hand, I propose that curse words update the utterance context by conditioning what it is assumed about the speaker's feelings (or his affective dispositions) in such context. That is, by conditioning a probability distribution over the speaker's feelings relative to a specific stimulus (i.e., '$\Pr(\alpha)$', to be read as 'the probability that the speaker feels α with respect to a stimulus x'). Thus, under this approach, the interpretation of a curse word at a context is not fixed but dependent on what is assumed about the speaker.

This chapter is organized as follows. Section 3.2 presents Burnett (2017, 2019)'s pioneering work on identity construction through sociolinguistic variation, which will provide the basis for our indexical approach to affective expressivity. Section 3.3 presents our proposal, where curse words are indexically associated with different affective qualities (derived from the dimensions pleasure, arousal and dominance), anyone of which can be activated depending on what is previously assumed about the speaker. Even though the proposal we will develop is pragmatic, Section 3.4 will sketch a compositional implementation based on Potts (2007a,b)'s account, in which expressives are represented as functions that probabilistically update the context of interpretation. Section 3.5 discusses some of the features of the proposal (and the prospects of a semantic implementation).

3.2 Background: modelling social meaning

The proposal developed in Section 3.3 is inspired by Burnett (2017, 2019)'s pragmatic model of sociolinguistic variation. In her proposal, she applies a probabilistic game-theoretic framework to formalize how speakers and listeners reason about each other in the transmission of social information. In particular, her model makes it clear that the expression of social information is constrained by what the audience assumes about the speaker at the utterance context. This mechanism, I argue, is analogous to what happens during the transmission of affective information, where items such as *damn* can be interpreted in many different ways according to what it is assumed about the speaker. Before presenting my own proposal, I will briefly review game-theoretic pragmatics analyses of speaker-listeners interactions (Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013; Goodman and Frank, 2016), and how Burnett (2017, 2019) applies them to the case of social meaning in particular.

3.2.1 Game-theoretic pragmatics

Game theory is a formalism that describes strategic interactions. In game-theoretic pragmatics, it is standard to focus on 'signaling' games (Lewis, 1979). In these games, there are two players, the speaker (S) and the listener (L). Roughly speaking, it is assumed that S wants to transmit certain information to L and that L wants to learn such information. Based on the information that S desires to transmit, she picks a message to send to L. And, based on S's message, L interprets it by assigning it a meaning. Importantly, signaling games are cooperative, because both S and L win if L interprets the message in the way intended by S, and both lose otherwise.

In signaling games, the rules (also called 'solution concept') establish that S and L's actions are constrained by i) their prior beliefs about their interlocutor and ii) their reasoning about how their interlocutor will probably act.

Importantly, to characterize how S and L recursively reason about each other, we assume that they use Bayesian reasoning. In other terms, we assume that players draw a conclusion B after having observed event A (written 'P(B|A)', to be read as 'the probability of B given A') by combining two things:

- How likely they think A is to indicate B (written as 'P(A|B)', read as 'the likelihood of A given B').
- How likely they though B was in the first place (written 'P(B)', read as 'the prior belief that B is the case').

To illustrate how this Bayesian game-theoretic framework works, we will focus on its application to the calculation of scalar implicatures (Frank and Goodman, 2012; Goodman and Frank, 2016). Imagine that Bob lives with John and Mary. One day, Bob leaves three cookies on the dinner table. When Bob calls home, John tells him (1):

(1) Mary ate some of the cookies.

Bob will likely infer from (1) that Mary ate one or two cookies but probably not all three. An intuitive explanation is that speakers try to be as informative as possible so, if Mary had eaten all the cookies then John would have said (2), which is more informative than (1).

(2) Mary ate all of the cookies.

How do we formalize this reasoning in a probabilistic framework? First, we assume i) that w_i is the possible world in which Mary ate exactly i cookies, where $i = \{0, 1, 2, 3\}$. Furthermore, we suppose ii) that John saw Mary eating 2 cookies, so w_2 is the actual world. Finally, we also assume iii) that John can only choose from the following two messages. As the table below

shows, each message is paired with a meaning, which is the set of worlds in which it is true:

(3) Messages in the cookie example

Name	Message	Meaning
Some	Mary ate some of the cookies	w_1, w_2, w_3
All	Mary ate all of the cookies	w_3

Importantly, John's decision of picking one of these two messages to convey that we are in w_2 depends on what he assumes about his interlocutor's (i.e., Bob) beliefs concerning how many cookies have been eaten. In this framework, listener's belief states are treated as a prior distribution over possible worlds ($\Pr(w)$). For the purposes of the example, we assume that John thinks that Bob has no prior expectation about how many cookies Mary ate, so $\Pr(w)$ is represented as an uniform distribution over possible worlds:

(4) Bob's prior beliefs in w ($\Pr(w)$)

Possible world	w_0	w_1	w_2	w_3
$\Pr(w)$	0.25	0.25	0.25	0.25

Then, with Bob's prior beliefs in mind, John picks a message m to say. Following Franke (2009) and Frank and Goodman (2012), we assume that, after John picked a message m, Bob conditions his beliefs on m's meaning. Importantly, the conditioning proceeds in two steps: first, the listener restricts his attention to those worlds in which m is true and eliminates those in which m is false (that is, intersects w with m) and then readjusts their beliefs (that is, 'normalizes' the resulting measure):

(5) $$\Pr(w|m) = \frac{\Pr(\{w\} \cap [m])}{\Pr([m])}$$

In the cookie example, the result of the conditioning for all messages m (cf. the table in 3) based on Bob's priors (cf. the table in 4) are shown in the following table:

(6) Bob's beliefs in w after hearing m ($\Pr(w|m)$):

Message	w_0	w_1	w_2	w_3
All	0	0	0	1
Some	0	0.333	0.333	0.333

As we can observe, after uttering 'all', Bob is completely certain that Mary ate all of the cookies, i.e, that the world in which we are is w_3. In contrast, after uttering 'some', Bob is certain that Mary didn't eat zero cookies, but is equally uncertain about how many she ate.

At this point, we can already explain the implicature triggered by 'some', i.e., that if Mary had eaten all of the cookies, John would have said 'all' instead of 'some' because 'all' is more informative. To that effect, we assume that the probability of John using message m to convey world w is determined by optimizing the probability that the listener assigns to w after hearing m ($\Pr(w|m)$). For our present purposes, we only require that the optimization satisfies the rule in (7), i.e., that the speaker (S) picks a message m instead of m' in w iff the listener (L) would assign a higher probability to w after hearing m than m' (Qing and Cohn-Gordon, 2019):

(7) $P_S(m|w) > P_S(m'|w)$ iff $P_L(w|m) > P_L(w|m')$

Now, since $P_L(w_3|all) = 1 > P_L(w_3|some) = 0.333$, from this formula we have that $P_S(all|w_3) > P_S(some|w_3)$, i.e., that John would prefer to use 'all' rather than 'some' if Mary had eaten all three cookies.

In this brief exposition of game-theoretic pragmatics we saw some of its basic

features: i) that a message's meaning is represented by the set of possible worlds in which it is true, ii) that the listener's belief update proceeds by conditioning her prior beliefs on m's meaning, which results in a new probability distribution, iii) that conditionalization proceeds by elimination, i.e., by focusing on the worlds in which m is true and discarding the rest, and then by normalizing the resulting measure.

Note that this framework also allow us make quantitative predictions about the speaker's and listener's behaviors using Bayes' rule, based on the assumption that the speaker attempts to be informative but is not fully rational (e.g., that speakers are resource-bounded agents, with information processing limitations). That is, that even though the speaker tries to maximize his utility, he does not always pick the most optimal (i.e., informative) option. In Section 3.3.6, we will explore some of the constraints that affect speaker's choice of a message, but the prediction of the production probabilities of actual speakers will be left for a future project.

3.2.2 Burnett (2017, 2019) on social meaning

In this section, we will see how the game-theoretic pragmatic framework presented above has been used to formalize how speakers and listeners reason about each other during conversations to transmit socially relevant information. Compare the following utterances:

(8) a. I am walking
b. I am walkin'.

It has been observed that, even though the variants of (ING) don't have any impact on the utterance's truth-conditions, they convey various social properties about the speaker (Campbell-Kibler, 2005, 2007, 2008). In particular, the use of *-ing* tends to convey that the speaker is competent (i.e., educated,

articulate, etc.) but aloof (i.e., formal, unfriendly, etc.). In contrast, the use of *-in* tends to be associated with the opposite properties, i.e., being incompetent and friendly.

Moreover, there is evidence that speakers exploit these associations to 'build' an identity in conversational settings. In Labov (2012)'s study of Obama's speech in different contexts, it is observed that the rate of the *-in* variant is the highest during a barbecue but the lowest in a scripted acceptance speech, thus showing that Obama's choices regarding these variants considerably change depending on his audience and communicative goals. To understand this phenomenon in a probabilistic framework, we need to answer the following questions: i) what do possible worlds represent? and ii) how can we represent social meaning?

To answer (i), Burnett (2017, 2019) assumes that there is a set $\mathbb{P}$ of social properties a person can have. In Obama's example, we assume that the set of properties $\mathbb{P}$ includes 'competent', 'aloof' and their opposites. Given that some subsets of $\mathbb{P}$ are incoherent (e.g., an individual can't be competent and incompetent at the same time), the model includes the symbol '$>$', which introduce relations of incompatibility between the properties in $\mathbb{P}$:

(9) Properties in Obama's example:

a. $\mathbb{P} = \{\text{competence, incompetent, friendly, aloof}\}$

b. competent $>$ incompetent

c. friendly $>$ aloof

Then, we assume that the speaker (S) wants to convey to the listener (L) a particular persona, defined as a 'maximally compatible set of properties' (Burnett 2019). In other terms, a persona π is any subset of $\mathbb{P}$ that contains properties not banned by $>$. In Obama's example, we obtain four possible persona: the competent aloof persona, which is called the STERN LEADER,

the competent friendly persona, which is called the COOL GUY, etc.:

(10) Possible personae in Obama's example:

STERN LEADER	COOL GUY	ASSHOLE	DOOFUS
{comp., aloof}	{comp., friendly}	{incomp., aloof}	{incomp., friendly}

Now, to answer (ii), Burnett (2017, 2019) assumes that messages index a subset of the social properties in $\mathbb{P}$, which is called its 'indexical field' (Silverstein, 1979; Eckert, 2008). In its basic form (which Burnett 2017 calls 'Eckert fields'), messages are directly associated with the personae they signal. In Obama's example, the Eckert field associated with *-ing* includes 'competent' and 'aloof', which correspond to the STERN LEADER persona.

However, in order to accommodate indexical fields in a game-theoretic framework, Burnett adopts a different characterization. In her model, indexical fields (henceforth 'Eckert-Montague' fields) are defined as the sets of personae that a message has the 'potential' to build. Roughly speaking, a message m index a persona π iff π contains at least one property that is compatible with the variant. In Obama's example, the Eckert-Montague field associated with the (ING) variants are the following:

(11) Eckert-Montague fields associated with (ING):

π	STERN LEADER	COOL GUY	ASSHOLE	DOOFUS
⟦-ing⟧	1	1	1	0
⟦-in⟧	0	1	1	1

As we can observe, given that *-ing* is compatible with being **competent** and **aloof**, the only persona incompatible with it is DOOFUS. And, since *-in'* is compatible with both **incompetent** and **friendly**, the only persona incompatible with it is STERN LEADER.

The details of Burnett's framework that will follow parallel those observed for the calculation of scalar implicatures in the last section. First, the speaker S assumes that the listener L has prior beliefs concerning which persona S instantiates (i.e., $\Pr(\pi)$). In Obama's example, he may assume that, because he is the president, listener's priors slightly favor personae that are aloof:

(12) Listener's prior beliefs in Obama's example:

π	STERN LEADER	COOL GUY	ASSHOLE	DOOFUS
$\Pr(\pi)$	0.3	0.2	0.3	0.2

Now, assuming that Obama wants to convey a persona, he will choose a message. Once he chooses a message, L conditions his prior beliefs on the messages' social meaning, i.e., its Eckert-Montague field. This conditionalization proceeds in two steps: i) L intersects each possible personae π with the messages' field and ii) adjust his prior beliefs accordingly:

(13) $$\Pr(\pi|m) = \frac{\Pr(\{\pi\} \cap [m])}{\Pr([m])}$$

For example, after hearing *-ing*, L discards the personae that are incompetent and casual (i.e., L assigns 0% to the DOOFUS persona) and, after hearing *-in'*, L discards the personae that are competent and aloof (i.e., L assigns 0% to the STERN LEADER persona):

(14) L's beliefs after hearing m at the barbecue:

π	STERN LEADER	COOL GUY	ASSHOLE	DOOFUS
$\Pr(\pi\|ing)$	0.375	0.250	0.375	0
$\Pr(\pi\|in')$	0	0.286	0.428	0.286

At this point, we can explain why Obama would prefer to, for example, use the *-in'* variant rather than the *-ing* variant to build the COOL GUY (CG) persona by using the rule in (7), adapted below:

(15) $P_S(m|\pi) > P_S(m'|\pi)$ iff $P_L(\pi|m) > P_L(\pi|m')$

Given that $P_L(CG|in') = 0.286 > P_L(CG|ing) = 0.250$, we can conclude that $P_S(in'|CG) > P_S(ing|CG)$. That is, that assuming that Obama wants to convey the COOL GUY persona, he would prefer to use the *-in'* variant rather than the *-ing* variant.

In this brief exposition of Burnett (2017, 2019)'s model, we observed how social information is represented: a message's social meaning is the set of possible personae that the message has the potential to signal, given a set of social qualities and relations of compatibility among them. Moreover, in this model, the informational impact of social meaning is analogous to that of quantifiers such as 'all' vs. 'some': when they are used, the listener's beliefs are updated by eliminating those possible worlds (or possible personae) which are incompatible with the message's meaning and then normalizing the resulting measure.

3.3 The proposal: affective indexical fields

Curse words can express a wide range of affective states. In (16), *damn* can be interpreted as displaying frustration because the pizza arrived late, joy because it finally arrived, surprise because it unexpectedly arrived, or maybe a combination of all these states altogether:

(16) The damn pizza arrived.

As McCready (2012) points out, the interpretation of *damn* in (16) heavily depends on what it is assumed about the speaker's emotions. That is, on how the speaker probably feels in the context where (16) is uttered, which in turn is inferred on the basis of different affective cues: his desires, facial expressions, etc. (cf. figure 3.1). Thus, even though *damn* is more likely to be interpreted negatively, different interpretations will become salient depending on how the speaker's emotional states are perceived by the audience.

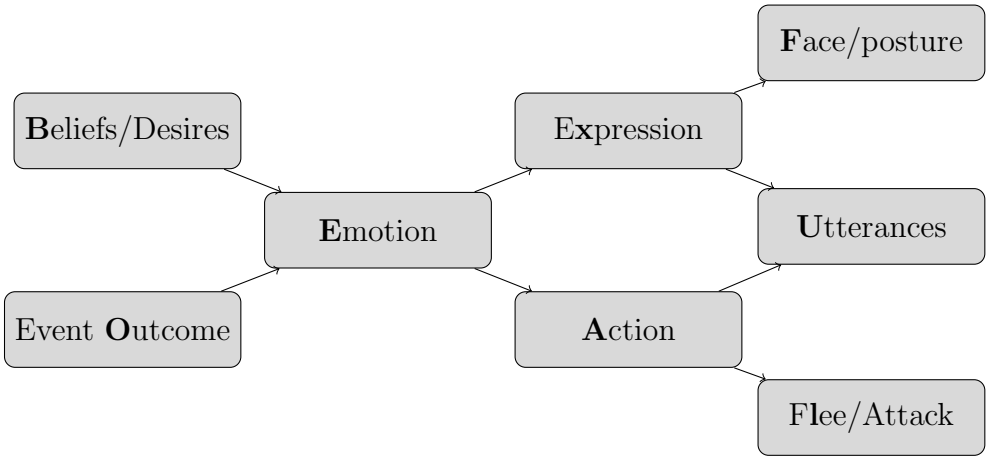


Figure 3.1: Lay theories of emotions (based on Ong et al. 2020)

In what follows, I propose a probabilistic pragmatic model for affective expressives. In this model, I associate curse words such as *damn* with indexical fields, which I represent as probability distributions that specify 'what it takes' to utter a given expression in typical situations. In order to introduce affective states, I use the dimensions pleasure, arousal and dominance, whose different configurations will determine the wide array of affective states that can be expressed.

Before modelling affective states in a probabilistic framework, we need to define the set of alternative messages that speakers can use in order to display their affective states to the listener. Following McCready (2012), we assume that this set includes a message that conveys propositional and underspecified

affective meaning ($m_{fucking}$) and a message that only conveys propositional meaning ($m_{-fucking}$):

(17) a. $m_{fucking}$: e.g., 'It is raining in fucking Lancaster.'
b. $m_{-fucking}$: e.g., 'It is raining in Lancaster.'

The main reason to choose these utterances as alternatives is that, as we saw in Chapter 2.2, expletives like *fucking* are always optional, i.e., the speaker can decide whether to use the expletive or not without altering the utterance's truth-conditions. Thus, this set is *prima facie* plausible as a representation of the listener's conceptualization of the decision situation after hearing an utterance like (17a). Moreover, this assumption is analogous to RSA models of vagueness which include a 'null' utterance as alternative (Lassiter and Goodman 2013), except that in our model one $m_{-fucking}$ is null only from the expressive point of view (cf. the discussion in Scontras et al. 2021).

Now, as before, we start by answering the following questions: i) what do possible worlds represent? and ii) how can affective meaning be represented?

3.3.1 Domain

To answer (i), we postulate a set $\mathbb{Q}$ of affective qualities. These characterize the possible affective states that an individual may experience. Following the discussion in Chapter 1.3, I assume that affective states are characterized by at least two orthogonal dimensions, pleasure and arousal (henceforth P and A)[1]. As we saw in Chapter 1.3, emotions can be represented as points in the space determined by the pleasure and arousal dimensions. Alternatively, they can be represented as wedges that also specify the different degrees of intensity in which an emotion can be instantiated:

[1]The third dimension, dominance, will be used in Chapter 6 to analyse slur's derogatory impact.

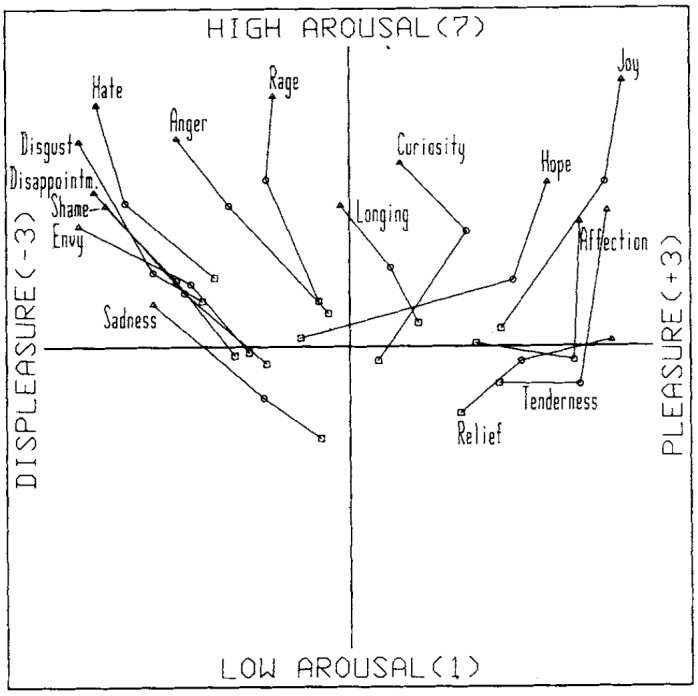


Figure 3.2: Mean pleasure of 30 affects (Reisenzein, 1994)

Now, pleasure and arousal are associated with different types of affective qualities. Pleasure determines a scale including negative ([P-]), neutral ([P±]) and positive ([P+]) affective states. In contrast, arousal determines a scale ranging from calm ([A-]) to aroused ([A+]) affective states, where there is no qualitatively 'neutral' arousal ([A±]), which instead can be represented as the absence of an emotion (i.e., apathy). Moreover, some combinations of these qualities are incoherent (e.g., an individual cannot be calm and aroused at the same time), so we use the symbol '>' to impose relations of compatibility. Notice that no affective state corresponds to [P±, A-] (unless one considers 'sleepiness' as an emotion), and that [P±, A+] corresponds to a 'pre-affective' state (i.e., excitement), so we add the constraints in (18e-f):

(18) $\mathbb{Q} = \{[P+], [P-], [P\pm], [A-], [A+]\}$

a. [P+] > [P-]
b. [P+] > [P±]
c. [P-] > [P±]
d. [A-] > [A+]
e. [A-] > [P±]
f. [A+] > [P±]

Now that we have defined $\mathbb{Q}$, we assume that the speaker (S) wants to convey to the listener (L) an affective state $\alpha \subset \mathbb{Q}$. Following Burnett (2017, 2019), we define an affective state α as a maximally compatible set of qualities in $\mathbb{Q}$. Thus, from $\mathbb{Q}$, we obtain four possible affective states AFF: the [P+, A-] affective state, which we label EASE, the [P+, A+] state, which we label JOY, etc.:

(19) Possible affective states AFF:

EASE	JOY	DISDAIN	ANGER
[P+, A-]	[P+, A+]	[P-, A-]	[P-, A+]

Notice that these labels should be understood as assembling different affective states. For example, JOY will be the name that represents [P+, A+] states in general (e.g., hope, affection, friendliness, etc.), and not only joy.

3.3.2 Indexical fields

To answer (ii), we first assume that the affective meaning of an expression m is a set of qualities in $\mathbb{Q}$, which we write $[\![m_{fucking}]\!] \subseteq \mathbb{Q}$. Thus, in the case of $m_{fucking}$, which is typically interpreted as signalling negative but arousing emotional states, we assume that it is associated with [P-, A+], i.e., ANGER.

However, as Burnett (2017, 2019) points out, we need to be able to talk about the set of affective states that a given message m has the potential to signal

(also called its 'Eckert-Montague' field). Thus, if we assume that $m_{fucking}$ = [P-,A+], then the only state that it cannot signal would be [P+,A-], i.e., EASE:

(20) Eckert-Montague field associated with $m_{fucking}$:

AFF	EASE	JOY	DISDAIN	ANGER
$m_{fucking}$	0	1	1	1

However, the conceptualization in (20) is problematic for our present purposes. On the one hand, it doesn't capture the fact that $m_{fucking}$ is associated with some emotions more than others. Indeed, $m_{fucking}$ is typically more connected to ANGER than to JOY or DISDAIN (Jay 2000). On the other, it doesn't capture the fact that, even though the use of curse words such as *fucking* predominantly signal [P-,A+] affective states, their use doesn't completely override the possibility that the speaker may experience a [P+, A-] state (e.g., contentment) later in the conversation. In sum, curse words such as *fucking* present a higher degree of conventionalization, so their indexical fields should incorporate explicit information about how they are typically interpreted.

For these reasons, I characterize indexical fields as a probability distribution $\Pr(m|\alpha)$, read as 'the likelihood of a message m given an affective state α'). In other terms, this distribution captures 'how strongly' a given expression m and an affective state α are associated, i.e., which emotions people usually experience when they use a particular cursing expression.[2]

[2]This characterization of indexical fields is inspired by Henderson and McCready (2019) analysis of dogwhistles, an extension of Burnett (2017, 2019) which incorporates the listener's beliefs about how personae and expressions are typically connected. As the authors note, this is just a trivial extension of Burnett (2017, 2019). Her analysis can be recovered by just assuming that the likelihood $\Pr(m|\alpha)$ is 1 whenever α is in the indexical field of m.

Thus, to capture the aforementioned properties of *fucking*, we assign a high value to $m_{fucking}$ displaying ANGER and a low value to $m_{fucking}$ displaying EASE. Assuming that the probabilities of the other two states are in between, we represent the indexical field of $m_{fucking}$ as follows:

(21) Indexical field associated with $m_{fucking}$:

AFF	EASE	JOY	DISDAIN	ANGER
$\Pr(m_{fucking}\|\alpha)$	0.3	0.4	0.6	0.7

What about the indexical field of $m_{-fucking}$? As we observe in (22), $\Pr(m_{-fucking}|\alpha)$ = 1 - $\Pr(m_{fucking}|\alpha)$. That is, we assign a low value to $m_{-fucking}$ displaying ANGER, and a high value to $m_{-fucking}$ displaying EASE:

(22) Indexical field associated with $m_{-fucking}$:

AFF	EASE	JOY	DISDAIN	ANGER
$\Pr(m_{-fucking}\|\alpha)$	0.7	0.6	0.4	0.3

Now, it may appear that (22) is not plausible as a representation of the emotions that are typically expressed when someone chooses an utterance without expressive items. Indeed, it is even unclear that 'neutral' utterances (e.g., 'It is raining in Lancaster') express any emotion at all. However, the same could be said about the *-ing* variant (e.g., 'cooking'), which doesn't seem to be linked to any particular persona when it is considered in isolation. Indeed, it is only when we contrast it to *-in* that *-ing* no longer appears as the 'neutral' variant, but as the variant that signals that the speaker is competent and aloof. Similarly, even though we need to empirically estimate the affective contrast between $m_{fucking}$ and $m_{-fucking}$ in future work, we can use the distributions in (21-22) to simulate the use of *fucking* in conversational exchanges.

3.3.3 Relativized prior beliefs

As in standard game-theoretic frameworks, we assume that the listener L has prior beliefs about how the speaker S feels before he talks, represented as a distribution over affective states ($\Pr(\alpha)$). However, it is worth noting that prior beliefs about the speaker's emotions are different from prior beliefs about the speaker's identity, as described in Burnett (2017, 2019)'s model. To wit, while sociolinguistic variables such as *-ing* are used to signal a persona, i.e., a general way in which the speaker desires to present himself to others, cursing expressions such as *fucking* signal the speaker's affective states with respect to particular stimulus. By uttering

(23) The fucking dog is sleeping

the speaker doesn't just signal that he is upset, but upset with the dog. Thus, we should be aware of the fact that cursing expressions often interact with other parts of speech, signalling affective states associated with what these parts refer to. In other terms, $\Pr(\alpha)$ should be relativized to specific stimuli, and thus cane read as 'the probability distribution that the speaker feels α with respect to stimulus x'.

Now, in a situation where L has no prior expectations about the speaker's feelings before he talks, we can represent L's beliefs as a uniform distribution over affective states:

(24) Listener's relativized prior beliefs:

AFF	EASE	JOY	DISDAIN	ANGER
$\Pr(\alpha)$	0.25	0.25	0.25	0.25

As we saw in Chapter 1.3, beliefs about an agents' emotions are determined by different types of affective cues. For example, by publicly available events

(e.g., winning a lottery, being fired, etc.) or actions (e.g., the agent's intonation, facial expressions, etc.). However, in subsequent chapters, we will analyse how less public factors can also determine how we reason about someone's emotional states. For example, as we will see in Chapter 5 and 6, the interpretation of some pejorative terms require us to reason on the basis of the speaker's psychological closeness to the target and her own social identity:

A Psychological closeness: this factor is determined by the past experiences and interactions between speaker and target (e.g., the frequency with which speaker and target engage in affiliating behavior or mock agressiveness). This factor will be useful to explain, for example, why individuals that don't belong to the group derogated by a slur, but which have a certain 'insider' status, can use the slur non-offensively (Ritchie, 2017).

B Social identity: 'identities' are labels that people use to group each other. When the speaker is identified with a label, such identification is interpreted as giving reasons to the speaker to feel in certain ways (Appiah, 2010). If the speaker is Catholic, it will be assumed that he tends to feel positively about the Catholic church and its teachings; if the speaker is South-American, it will be typically assumed that he doesn't feel that South-Americans are lesser or deserve to be discriminated. Even though these assumptions may prove to be incorrect, speaker's social identities guide how listeners think about their feelings.

In a context of utterance, the combination of public and private affective cues will influence the listener's expectations about the speaker's affective stance towards the target of the curse word. However, I remain neutral about the epistemological problem of deciding whether an agent's prior beliefs are formally constrained by rational principles (as objective Bayesians claim) or by non-rational processes such as socialization, evolution or free choice (as subjective Bayesian claim) (Talbott, 2001). Instead, the aim of our model will

be explaining how, even though curse words are typically interpreted as expressing negatively valenced states, they can have numerous interpretations depending on the listener's priors.

3.3.4 Context update (v. 1)

Given the above, context update proceeds as follows. Once S utters a cursing expression m whose argument is x, L's prior beliefs are updated by conditioning $\Pr(\alpha)$ on m's indexical field, $(\Pr(m|\alpha))$. In other terms, L i) combines the likelihood of m signalling an affective state with his prior beliefs about S's affective state or tendencies with respect to x, and then ii) readjust the resulting measure with a normalizing constant, that is, the sum of these terms computed for all affective states α:

$$\Pr(\alpha|m) = \frac{\Pr(\alpha) \times \Pr(m|\alpha)}{\sum_{\alpha} \Pr(\alpha) \times \Pr(m|\alpha)} \tag{25}$$

Here are two examples of how this work. If S utters *fucking dog*, but L has no prior expectations about S's feelings towards the dog, then we plug the uniform distribution in (24) and the probabilistic field associated with $m_{fucking}$ in the formula in (25). As a result, the probability that L assigns to each affective state after hearing *fucking dog* are the following:

(26) L's posterior beliefs after hearing $m_{fucking}$:

AFF	EASE	JOY	DISDAIN	ANGER
$\Pr(\alpha)$	0.25	0.25	0.25	0.25
$\Pr(m_{fucking}\|\alpha)$	0.3	0.4	0.6	0.7
$\Pr(\alpha) \cdot \Pr(m_{fucking}\|\alpha)$	0.075	0.1	0.15	0.175
$\Pr(\alpha\|m_{fucking})$	0.15	0.2	0.3	0.35

Similarly, the probability that L assigns to each affective state after hearing the same utterance without the expressive element, i.e., $m_{-fucking}$, are computed as follows:

(27) L's posterior beliefs after hearing $m_{-fucking}$:

AFF	EASE	JOY	DISDAIN	ANGER
$\Pr(\alpha)$	0.25	0.25	0.25	0.25
$\Pr(m_{-fucking}\|\alpha)$	0.7	0.6	0.4	0.3
$\Pr(\alpha) \cdot \Pr(m_{-fucking}\|\alpha)$	0.175	0.15	0.1	0.75
$\Pr(\alpha\|m_{-fucking})$	0.35	0.3	0.2	0.15

As we can observe, $\Pr(\text{ANGER}|m_{fucking}) > \Pr(\text{ANGER}|m_{-fucking})$. Following the rule in (15), adapted to the case of emotion signalling in (28), we can thus assume that the S would prefer to use $m_{fucking}$ in order to express ANGER.

(28) $\Pr_S(m|\alpha) > \Pr_S(m'|\alpha)$ iff $\Pr_L(\alpha|m) > \Pr_L(\alpha|m')$

Now, in a different type of situation, e.g., a situation where L assumes that S loves his dog, L's prior beliefs can be represented as favoring [P+] over [P-] affective states. As a result, we obtain that the L's posterior beliefs after hearing *fucking dog* now favour [P+] states:

(29) L's posterior beliefs after hearing $m_{fucking}$:

AFF	EASE	JOY	DISDAIN	ANGER
$\Pr(\alpha)$	0.40	0.40	0.10	0.10
$\Pr(m_{fucking}\|\alpha)$	0.3	0.4	0.6	0.7
$\Pr(\alpha) \cdot \Pr(m_{fucking}\|\alpha)$	0.12	0.16	0.06	0.07
$\Pr(\alpha\|m_{fucking})$	0.365	0.390	0.146	0.170

In this situation, the probability that L assigns to each affective state after hearing the same utterance without the expressive element, i.e., $m_{-fucking}$, are computed as follows:

(30) L's posterior beliefs after hearing $m_{-fucking}$:

AFF	EASE	JOY	DISDAIN	ANGER
$\Pr(\alpha)$	0.40	0.40	0.10	0.10
$\Pr(m_{-fucking}\|\alpha)$	0.7	0.6	0.4	0.3
$\Pr(\alpha) \cdot \Pr(m_{-fucking}\|\alpha)$	0.28	0.24	0.04	0.03
$\Pr(\alpha\|m_{-fucking})$	0.474	0.406	0.067	0.050

Interestingly, in this type of situation $\Pr(\text{JOY}|m_{fucking}) \approx \Pr(\text{JOY}|m_{-fucking})$, with a slight advantage in the case of $m_{-fucking}$. However, if we assume that the speaker wants to display EASE, $m_{-fucking}$ seems much more likely to be chosen than $m_{fucking}$.

3.3.5 Speaker's utilities

How do speaker's decide whether tho choose between $m_{fucking}$ and $m_{-fucking}$? By using the rule in (28), we have assumed that the speaker's utilities (U_S) for a variant m to signal an affective state α are only guided by informativity. That is, that speaker and listener are aware that the former is trying to give to the latter the most information possible about their emotional states. Following Frank and Goodman (2012), the informativity of a message with respect to an affective state α is given by the natural logarithm (ln) of the prior probability of α conditioned on the indexical field of m:[3]

(31) $U_S(\alpha, m) = ln(\alpha|m)$

[3]For simplicity, we also assumed that there are not cost differences between each alternative message. Due to the fact that $m_{fucking}$ is longer than $m_{-fucking}$, we could consider the former more costly to produce than the latter. However, I leave this investigation for future work.

However, it is unlikely that speakers are only guided by informativity when deciding whether to use an impolite (and potentially aggressive) term like *fucking* or not. Thus, following Henderson and McCready (2019), we assume that speakers are sensitive to whether the listener is likely to 'approve' the emotion signalled by the speaker or not. For example, in contexts where there is an imbalance in power derived from the social roles of speaker and listener (i.e., employee vs. company boss), the listener is likely to penalize the expression of intense negatively valenced affective states. Thus, in these situations, the speaker may prefer to inhibit from using $m_{fucking}$ even though it is the most informative alternative.

Thus, we assume that the utility calculation takes into account the message's social value, which assigns a function v_L to the listener. This function assigns a positive or negative real number to each persona representing the listener's (dis)approval of an affective state. In this case, as Henderson and McCready (2019) point out, utilities can be calculated over message-listener pairs, in order to represent which message should be sent by the speaker given the way it might be received by the listener (positively or negatively) and the message's indexical field ($\Pr(\alpha|m)$:

(32) $U_S(\alpha, m) = \sum_\alpha ln(Pr(\alpha|m)) + v_L(\alpha)Pr(\alpha|m)$

We could also add a function v_S, which assigns a positive real number to each affective state representing the speaker's preferences (Burnett, 2019). For example, in situations where the speaker aims at threatening or intimidating the listener, v_S would assign a high value to ANGER over other affective states, independently of whether ANGER is penalized in the listener's function v_L. However, for simplicity, we won't take such factor into account.

Now, in the first case described above, i.e., the interaction represented in (26-27), we can assume that the context is such that the listener, which we

may call 'prudish listener', is adverse to [P-] states and favours [P+] states:

(33) $v_L(\alpha)$ for the prudish listener:

AFF	EASE	JOY	DISDAIN	ANGER
Values	100	100	-100	-100

Given these values for $v_L(\alpha)$, and the formula in (32), we can derive the utilities of the alternatives $m_{fucking}$ and $m_{-fucking}$:

(34) U_S:

Message	$m_{fucking}$	$m_{-fucking}$
Values	-40.758	24.242

As we can observe, even though $\Pr(\text{ANGER}|m_{fucking}) > \Pr(\text{ANGER}|m_{-fucking})$, and thus $m_{fucking}$ is more informative in case S desires to signal ANGER, S would overall prefer to inhibit from using $m_{fucking}$ when interacting with a prudish listener.

3.4 A compositional implementation

As we saw in the last section, expressives such as *fucking* or *damn* not only display the speaker's global affective states but also compositionally interact with other parts of speech. That is, *damn dog* don't merely indicates that the speaker is globally angry or surprised, but angry or surprised with respect to the dog. Thus, even though that I will continue using the pragmatic probabilistic framework in the chapters to follow, in this section I sketch a compositional implementation based on Potts (2007a,b) theory of expressives, analysed in Section 2.3.4. Given that Potts (2007a,b) proposes a logic based on his previous work on conventional implicatures ($\mathscr{L}_{CI}$), I will

refer to this probabilistic extension as $\mathscr{L}_{Pr}$. Refining many features of this compositional implementation will be left for future work.

3.4.1 Type system

Based on Potts (2007a), we distinguish descriptive and expressive denotations syntactically, that is, by postulating descriptive and expressive types. Importantly, the rule in (35d) specifies that complex expressive types have expressive types within its domain but not in its range. This rule thus ensures that expressives are independent: there are not expressions that can take expressive content as an argument, so expressives can't scopally interact with truth-conditional operators:

(35)
a. e and t are basic descriptive types.
b. ε is an expressive type.
c. If σ and τ are descriptive types, then $\langle \sigma, \tau \rangle$ is a descriptive type.
d. If σ is a descriptive type, then $\langle \sigma, \varepsilon \rangle$ is an expressive type.
e. The set of types is the union of the descriptive and expressive types.

Given that we will only focus on expressive adjectives such as *damn* in this section, we won't add more rules to this list. However, it is important to notice that a more complex type system will be required to deal with other types of curse words, such as pejoratives or slurs (Gutzmann, 2015; McCready, 2010).

3.4.2 Probabilistic expressive indices

As in Potts (2007a,b), expressive indices will be the main object manipulated by expressive denotations. In $\mathscr{L}_{Pr}$, indices are defined as probability distributions about the speaker's affective states α.

In order to define α, we follow Burnett (2017, 2019): we assume that α is an affective state derived from the structure $\langle \mathbb{Q}, > \rangle$ where:

(36) a. $\mathbb{Q} = \{p_1, ..., p_n\}$ is a finite set of affective qualities.
b. $>$ is a relation on $\mathbb{Q}$ that is irreflexive and asymmetric.

In this structure, $\mathbb{Q}$ is a set of relevant affective qualities (e.g., [P-], [A+], etc.), and $>$ encodes relationships of compatibility between them. From this relational structure, we derive affective states α, which are particular collections of affective qualities that 'go together'. Thus, the set of possible affective states AFF is the maximally consistent set of properties α, defined as follows:

(37) α is an affective state ($\alpha \in$ AFF) iff
a. $\alpha \subseteq \mathbb{Q}$ and there are no $p_1, p_2 \in \alpha$ such that $p_1 > p_2$
b. There is no $\alpha' \in$ AFF such that $\alpha \subset \alpha'$

Now that we have defined α, we define expressive indices as probability distributions over the possible affective states of a with respect to b:

(38) An expressive index is a triple $\langle a \ \mathbf{Pr}_\alpha \ b \rangle$ where
a. a and $b \in D_e$
b. $\mathbf{Pr}_\alpha$ is the probability distribution over affective states α that a may feel with respect to stimulus b.
c. α is an affective state derived from the structure $\langle \mathbb{Q}, > \rangle$.

In our proposal, we assume a relational structure $\langle \mathbb{Q}, > \rangle$ that is similar to the one we proposed in Section 3.3.1. That is, we assume that the affective qualities in $\mathbb{Q}$ are derived from the dimensions pleasure and arousal. There-

fore, we assume that individuals can only convey four types of affective states (e.g., the [P+, A+] state, which we label JOY, the [P-, A+] state, which we label ANGER, etc.).

A conceptual advantage of this definition of expressive indices is that it avoids an issue observed in Geurts (2007). To wit, indices of the form $\langle a\ \mathbf{I}\ b\rangle$ in Potts (2007a,b)'s system can be systematically mapped to propositions of the form 'a is at the expressive level **I** towards b'. That is, to descriptions of the world as being a certain way, which is at odds with the basic assumption that affective meaning is expressive rather than descriptive. In $\mathscr{L}_{Pr}$, in contrast, indices are associated with a probability distribution, which can be paired with evidential judgments of the form 'It is likely that a feels α with respect to b' or 'There is 75% possibility that a feels α with respect to b'. Given that these judgments don't describe the world as being a certain way, i.e., that there is no way the world could be that would make them true or false, we can avoid Geurts (2007)'s objection.

3.4.3 Context

Following Potts (2007a,b), we assume that the context is a Kaplanian tuple extended with an expressive parameter c_ε which keeps track of the conversational participant's feelings:

(39) A context is a tuple $c = (c_s, c_t, c_w, c_\varepsilon)$ where c_s is the speaker of c, c_t is the time of c, c_w is the world of c and c_ε is the expressive parameter.

In $\mathscr{L}_{Pr}$, c_ε is defined as a set of expressive indices of the form $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$. However, such modification requires a different interpretation of the probability function (**Pr**) in the expressive indices. In Section 3.3, we have understood $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ in subjectivist (i.e., Bayesian) terms. That is, as representing the

listener's degree of confidence that a has a tendency to feel α with respect to stimulus b. However, in a subjectivist interpretation of expressive indices, c_ε would contain multiple indices of the form $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ for every salient pair of entities a and b, depending on each listener's credences. In that case, there would be not one but multiple contexts of interpretation, and therefore possibly multiple interpretations of the same expressive term.

Thus, to determine a unique context of interpretation, we will understand the probability function in $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ in 'evidential' terms. In the evidential conception of probabilities, these are conceived as the (uniquely determined) 'intrinsic plausibility' of an hypothesis given the available evidence, independently of the agents actual credences (Williamson, 2002). Accordingly, we interpret $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ as measuring the intrinsic plausibility that a feels (or tends to feel) α with respect to b given the available evidence, independently of what each speech-act participants believes about the speaker's affective states. From this it follows that c_ε can contain at most one index of the form $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$, for every salient pair of entities a and b, and thus that there is a unique context of interpretation.

It may be argued that the objectivist view of probabilities, and thus of expressive indices, is not suitable to understand expressive meaning as an interactive phenomenon. However, the boundaries between subjective and evidential probabilities aren't sharp: agents are rationally constrained by objective evidential support relations (Hajek, 2002). Therefore, what speech-act participants come to believe about the speaker's feelings is constrained by what an evidential index $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ specifies. Moreover, we may conceive objective probabilities as the subjective probabilities that a perfectly rational agent possess. In such case, we would benefit from the applications that subjective probabilities have, such as their interactive character (Eder, 2019).

3.4.4 Context update (v. 2)

In $\mathscr{L}_{Pr}$, c_ε can be updated by expressive's indexical fields (i.e., $\Pr(m|\alpha)$) in two ways:

- Case 1: if there is no evidence about whether a feels α with respect to b in c, i.e., if c_ε does not include any index of the form $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$, then $c'_\varepsilon = c_\varepsilon \cup \langle a\ \mathbf{Pr}_\alpha\ b\rangle$, where $\mathbf{Pr}_\alpha = \Pr(m|\alpha)$. This ensures that, whenever there is no evidence about whether a feels α with respect to b, an expressive m is interpreted with its 'normal' interpretation.

- Case 2: if there is evidence about whether a feels α with respect to b in c, i.e., if c_ε does include an index of the form $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$, then $\langle a\ \mathbf{Pr'}_\alpha\ b\rangle$ replaces $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$, where $\langle a\ \mathbf{Pr'}_\alpha\ b\rangle$ is the result of conditioning $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ with $\Pr(m|\alpha)$.

Conditioning works in the way specified by the formula in (25). Given the above, we have the following relations between expressive indices:

(40) $c_\varepsilon \approx_{a,b}^{Pr'(\alpha)} c'_\varepsilon$ iff c_ε and c'_ε differ at most in that

a. $\langle a\ \mathbf{Pr'}_\alpha\ b\rangle \in c'_\varepsilon$; and

b. if c_ε contains an expressive index $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$, where $\mathbf{Pr}_\alpha \neq \mathbf{Pr'}_\alpha$, then $\langle a\ \mathbf{Pr}_\alpha\ b\rangle \notin c'_\varepsilon$; and $\langle a\ \mathbf{Pr'}_\alpha\ b\rangle$ is the result of conditioning $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ by the indexical field of an expressive m (i.e., $\Pr(m|\alpha)$).

Therefore, we can talk about contexts in which there is not evidence about whether a feels α with respect to b. For example, in a context where there is no evidence about whether Tom hates Jerry, his utterance of *damn Jerry* will be interpreted by using the indexical field of *damn*, which assigns 0.7% to *damn* displaying ANGER, and only 0.3% to it displaying EASE:

(41) There is no evidence about whether Tom likes Jerry in c

AFF	EASE	JOY	DISDAIN	ANGER
$\varnothing$	?	?	?	?
$\Pr(damn\|\alpha)$	0.3	0.4	0.6	0.7
$[\![tom]\!]$ $\mathbf{Pr}_\alpha$ $[\![jerry]\!]$	0.3	0.4	0.6	0.7

At this point, it may be asked why, in cases where there is no evidence about whether a tend to feel α with respect to b in c, c_ε don't just includes an index $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ which represents a uniform distribution over affective states? The main reason is that the evidence about the affective tendencies of a are always limited (e.g., it is unlikely that there is evidence about a's affective tendencies with respect to Saturn). Therefore, including potentially unlimited indices $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ in the context of interpretation is implausible as a representation of its configuration.

Now, in a context where there is evidence that Tom loves Jerry, e.g., where there is an index $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ which assigns 0.4% to Tom loving Jerry, his utterance of *damn Jerry* will be interpreted by combining $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ with *damn*'s indexical field, thus giving as result the new index $\langle a\ \mathbf{Pr'}_\alpha\ b\rangle$ which specifies that, this time, the probability of Tom feeling JOY towards Jerry is 0.39%:

(42) There is evidence that Tom loves Jerry in c

AFF	EASE	JOY	DISDAIN	ANGER
$[\![tom]\!]$ $\mathbf{Pr}_\alpha$ $[\![jerry]\!]$	0.40	0.40	0.10	0.10
$\Pr(damn\|\alpha)$	0.3	0.4	0.6	0.7
$[\![tom]\!]$ $\mathbf{Pr'}_\alpha$ $[\![jerry]\!]$	0.365	0.390	0.146	0.170

Finally, following Potts (2007a), we add the following rule, which specifies how the relations on sets of expressive indices in c_ε affect the relation between contexts c during the update:

(43) $c \approx^{\mathbf{Pr}_\alpha}_{a,b} c'$ iff $c_\varepsilon \approx^{\mathbf{Pr}_\alpha}_{a,b} c'_\varepsilon$

It is important to notice that, in $\mathscr{L}_{Pr}$, an index $\langle a\ \mathbf{Pr}_\alpha\ b\rangle$ belongs to c_ε iff there is evidential support that a feels (or tends to feel) α with respect to b. Such evidential support can be determined by verbal acts (e.g., by uttering expressives) or non-verbal acts (e.g., by the speaker's facial expressions, tone of voice, etc.). Therefore, in contrast to Potts (2007a,b), our model doesn't require that an expressive should be uttered in order for the context to include an index about the speaker's affective states (see Section 2.3.4).

In sum, expressives either introduce new expressive indices on the context or manipulate existing ones. When Tom utters *damn Jerry* and there is not evidence about whether he likes Jerry or not, the utterance introduces a new index specifying a probability distribution that favors ANGER over other affective states. But, one there is evidence that Tom likes Jerry, the utterance manipulates the existing index by conditioning it, thus outputting a new index that favors JOY.

3.4.5 Denotations

The denotation of expressives such as *damn* is here represented as a mapping from prior to posterior probability distributions. In formal terms,

(44) a. $[\![\text{damn}]\!] : \langle \text{e}, \varepsilon\rangle$

b. $[\![\text{damn}]\!]^c$ is the function f such that $f([\![a]\!]^c)(\text{c}) = \text{c'}$, where

(i) $c \approx^{\mathbf{Pr'}_\alpha}_{c_s,[\![a]\!]^c} c'$;

(ii) $\mathbf{Pr'}_\alpha = \mathbf{Pr}_\alpha$ conditioned by $\Pr(damn|\alpha)$.

(iii) $\Pr(damn|\alpha)$ assigns 0.7% to ANGER, 0.3% to JOY, etc.

Now, distinguishing mild expressives such as *damn* from strong expressives such as *fucking* requires specifying different indexical fields for each. In the

case of *damn*, we can assume that its indexical field assigns 0.7% to *damn* displaying ANGER and 0.3% to it displaying EASE. To distinguish it from *fucking*, the latter's indexical field may, for example, assign 0.8% to *fucking* displaying ANGER:

(45) a. $[\![\text{fucking}]\!] : \langle \text{e}, \varepsilon \rangle$

b. $[\![\text{fucking}]\!]^c$ is the function f such that $f([\![a]\!]^c)(\text{c}) = \text{c'}$, where

(i) $c \approx^{\mathbf{Pr'}_\alpha}_{c_s, [\![a]\!]^c} c'$;

(ii) $\mathbf{Pr'}_\alpha = \mathbf{Pr}_\alpha$ conditioned by $\Pr(fucking|\alpha)$.

(iii) $\Pr(fucking|\alpha)$ assigns 0.8% to ANGER, 0.10% to JOY, etc.

Importantly, Potts (2007b) notices that, even though expressive denotations such as (44) or (45) (that is, functions from contexts to context) capture the immediacy of expressives, they don't fully account for their compositionality. To wit, when [[damn]] combines with [[jerry]], the result is an altered context, but [[jerry]] is returned unmodified, thus remaining available to participate in further derivations. To solve that, Potts (2007) introduces the compositional operator '•' which is defined as follows[4]:

(46) Where α is of type $\langle \sigma, \varepsilon \rangle$, and β is of type σ:

$$[\![\alpha]\!]^{c'} \bullet [\![\beta]\!]^c = [\![\beta]\!]^{[\![\alpha]\!]^{c'}([\![\beta]\!]^c)(c)}$$

The composition of $[\![damn]\!]$ and $[\![jerry]\!]$ using '•' is the following. In the context c in which there is evidence that Tom likes Jerry, c' specifies that it is more likely that Tom feels positively about Jerry:

[4]This operator should not be confounded with the metalogical bullet '•' in (Potts, 2004), which only indicates that expressive and descriptive types are independent in the derivation.

(47) $[\![damn]\!]^{c'} \bullet [\![jerry]\!]^{c} = [\![jerry]\!]^{c'}$, where c' is just like c except that it includes an index $[\![tom]\!]$ $\mathbf{Pr}_{\alpha}$ $[\![jerry]\!]$ which favors JOY over ANGER.

And, in a context c where it is not known whether Tom likes Jerry, c' specifies that it is more likely that Tom feels negatively about Jerry:

(48) $[\![damn]\!]^{c'} \bullet [\![jerry]\!]^{c} = [\![jerry]\!]^{c'}$, where c' is just like c except that it includes an index $[\![tom]\!]$ $\mathbf{Pr}_{\alpha}$ $[\![jerry]\!]$ which favors ANGER over JOY.

For our present purposes, we will adopt this compositional operator as such, as it is required to account for the fact that expressives such as *damn* not only update the context but leave their arguments available for further derivations.

3.4.6 Discussion

Before concluding this chapter, let's discuss some aspects of $\mathscr{L}_{Pr}$ that need to be revised in later research, and provide some ideas about how such revision may be done.

First, we have replaced the subjectivist (Bayesian) conception of the probability function **Pr** by the objectivist (evidential) one. However, can the subjectivist approach be maintained? Doing this would require developing a different conceptualization of the context of interpretation and of context update. For example, we could assume a Stalnakerian conception of the context of interpretation instead of a Kaplanian one. In such framework, we would conceive expressive's informational impact in the context as updating not the common ground (i.e., what the speech-act participants take for granted) but only the commitments of the listener (i.e., what the listener assumes about the speaker's feelings at a given point in the conversation). However, this would require speaker's use of expressives to have a direct or privileged access to the listener's commitments: whereas propositional infor-

mation updates the listener's beliefs only if the listener doesn't object to it, expressive information is 'imposed' without the requirement that the listener previously approves its content or not.

Second, there are situations in which the use of *damn* doesn't signal the speaker's affective states but that of other salient individuals. As observed in Section 2.2.2, this phenomenon can occur when the expressive is embedded in a speech-report predicate (e.g., 'My father told me not to adopt that damn dog') or not (e.g., 'My father dislikes dog. I shouldn't have adopted the damn dog.'). To account for these cases, we may consider that the speaker is using the expressive *damn* in an echoic manner. That is, that the speaker is implicitly attributing the use of the expressive to someone else (an individual, group of individuals, or even a generic agent) who takes responsibility for the use of the expressive. In such case, the expressive cannot be interpreted with respect to the actual context, but with respect to the context of the person whose speech the speaker is reporting (Recanati, 2019).

Finally, it is worth noting that this chapter hasn't touched into questions about the pragmatic impact of affective expressions on context. However, the pragmatic impact of curse word arguably varies depending on the type of expression. For example, whereas expletive adjectives like *damn* can be used as weapons to threaten or intimidate, they don't have the same perlocutionary effects of slurs on their targets, which dehumanize and assign a subordinate role to their targets (Jeshion, 2013; Popa-Wyatt and Wyatt, 2017).

3.5 Summary

This chapter presented a pragmatic probabilistic framework to represent the transmission of affective information through expressive items (Section 3.3) and sketched a compositional implementation (Section 3.4). Even though I

will only employ the former pragmatic framework in subsequent chapters, we should bear in mind that, in the case of expressive meanings, the conventional (i.e., semantic) and indexical (i.e., pragmatic) distinction should not be taken as representing two categorically distinct meanings, but a 'gradient cline between two phases of the same process' (Beltrama, 2020) (see Section 2.1.2). The main takeaways from the discussion in this chapter are the following:

- Affective and social information pattern along various dimensions. For example, what affective of social properties end up being assigned to the speaker depend on what it is previously believed or known about them.

- The interpretation of curse words such as *damn* is probabilistic. That is, is given by a probability distribution over the affective states that typically prompt their use ($\Pr(\text{damn}|\alpha)$).

- Curse words update the utterance context by conditioning what is is known about the speaker's feelings in such context, represented by the distribution $\Pr_\alpha$, to be read as 'the probability that the speaker tends to feel α with respect to the expressive's target'.

- A compositional account of curse words such as *damn* is possible, but requires that we interpret the probability function $\mathbf{Pr}$ in objective (evidential) terms rather than subjective (Bayesian) terms. However, objective probabilities can 'permeate' the credences of the speech-act participants in a conversation.

Chapter 4

Expletive adjectives

4.1 Introduction

At this point of the dissertation, we have analysed canonical and less known features of expletive adjectives like *damn*, giving a special emphasis to their underspecified character (Section 2.2). As a result of such investigation, we have proposed a theory in which expressive adjectives update the interpreter's assumptions about the speaker's affective states (defined by means of the two affective dimensions pleasure and arousal). Moreover, we also offered a compositional implementation of such idea, in which terms like *damn* denote functions from contexts to contexts that update an expressive parameter in the context:

(1) a. $[\![\text{damn}]\!] : \langle \text{e}, \varepsilon \rangle$

b. $[\![\text{damn}]\!]^c$ is the function f such that $f([\![a]\!]^c)(\text{c}) = \text{c'}$, where

(i) $c \approx^{\mathbf{Pr'}_\alpha}_{c_s, [\![a]\!]^c} c'$;

(ii) $\mathbf{Pr'}_\alpha = \mathbf{Pr}_\alpha$ conditioned by $\Pr(damn|\alpha)$.

(iii) $\Pr(damn|\alpha)$ assigns 0.7% to ANGER, 0.3% to EASE, etc.

The main goal of our proposal has been to explain the fact that expressives like *damn* can be used to convey a wide array of emotions depending on what other (affective) properties are assumed to hold of the speaker in the utterance context.

However, there are two other issues that have attracted the attention in the semantic and pragmatic literature on expletive adjectives. The first is that, across different languages, many expletive adjectives give rise to degree interpretations when they modify scalar items. For example, in (2a), *fucking tall* may be interpreted as equivalent to *very very tall* (Geurts, 2007):

(2) a. The conference is **fucking** long.
b. El boleto de avión está **pinche** caro.
'The flight is fucking expensive.'

The second issue is that expletive adjectives seem to receive 'non-local' interpretations in some contexts. That is, expletive adjectives don't necessarily express emotions towards the (object referred by) the argument to which they apply. For example, in (3), *fucking* can be interpreted as conveying, e.g., a (negative) attitude towards Jones, despite its nominal-internal position in the syntax (Potts, 2004; Frazier et al., 2014; Gutzmann, 2019):

(3) Alex says Jones forgot to buy the **fucking** pizza.

This chapter addresses these two features of expletive adjectives. Section 4.2.1 analyzes the syntactic environments in which degree uses of expletive adjectives arise, and argue that degree uses don't necessarily lose their affective connotations. Section 4.2.2 argues that expletive adjectives cannot be considered ambiguous between affective and degree uses, and Section 4.2.3 proposes that these readings arise instead as pragmatic enrichments. In few

terms, in example (2a), the conference's length is interpreted as mapping the length of the affective intensity associated with the indexical meaning of *damn*.

In Section 4.3.1, I review the empirical data about non-local readings, and argue that affective expressive morphemes also seem to display the same kind of syntactic flexibility, so an explanation of non-local readings should be structurally similar in both cases. In Section 4.3.2, I argue that current proposals to account for non-local interpretations are too extreme, and that a moderate solution is preferable. Thus, Section 4.3.3 proposes that, whereas interpreting the meaning of *damn* on the basis of the speaker's emotions requires listener's to reason 'forward' from the speaker's emotions to the verbal expressions of their emotions ($\mathrm{P}(m|\alpha)$), non-local interpretations require listener's to reason 'backwards' from the speaker's emotions to the potential events e that caused the emotions ($\mathrm{P}(e|\alpha)$) (see also Chapter 1.3.2).

4.2 Degree readings

4.2.1 The data

As Geurts (2007) and Morzycki (2011) observe, expletive adjectives can be interpreted as degree words in certain environments. In general, for degree interpretations to arise, the head nouns or adjectives modified by the expletive adjective are required to be scalar. In (4), the bracketed expletive adjectives are interpreted as moving something up an scale of intelligence. That is, as indicating that Frege was very intelligent:

(4) Frege was {fucking, goddamn, bloody} intelligent.

Importantly, notice that, when expletive adjectives receive degree interpretations, they make a truth-conditional contribution to their host utterance.

For example, if we omit the expletive adjective in (5), the resulting utterance will be deviant, thus indicating that the truth-conditions expressed have been altered:

(5) a. Einstein was intelligent, but Frege was **fucking** intelligent.
b. #Einstein was intelligent, but Frege was intelligent.

Another way to see the truth-conditional contribution of expletive adjectives *qua* degree modifiers can be observed in entailment patterns. Arguably, (6) doesn't entail (6b), i.e., the speaker will leave if the conference is very very long, not just long:

(6) a. If the conference is **fucking** long, I will leave.
b. $\not\rightarrow$ If the conference is long, I will leave.

It is worth noting that this phenomenon can be observed across languages. In Spanish, expletive adjectives (e.g., *puto*, *pinche*) and even adverbs (e.g., *jodidamente*, *estupidamente*) can also receive degree interpretations (Padilla Cruz, 2018). In (7b), for example, the speaker qualifies the sandwich as better than the standard:

(7) a. Curro es un **puto** crack.
'Curro is a fucking crack.'
b. Este sandwich esta **jodidamente** bueno.
'This sandwich is fucking good.'

Now, observe that expletive adjectives are not the only non-truth-conditional (or expressive) phenomenon that has been observed to give rise to the intensification of a scalar noun or adjective. For example, the prosodic stress

in (8a) raises the standard of *intelligent* to a higher degree, thus affecting the utterance's truth-conditions (Kennedy, 2007). And, moreover, the length of the vowels in the sign 'long' (i.e., the number of its replications) in (9a) seem to map directly onto the duration of the object of which *long* is predicated (i.e., *the-conference*) thus affecting the host utterance's truth-conditions (Schlenker, 2018):

(8) a. Einstein was intelligent, but Frege was INTELLIGENT.
b. #Einstein was intelligent, but Frege was intelligent.

(9) a. If the conference is looooooong, I will leave.
b. $\not\rightarrow$ If the conference is long, I will leave.

A crucial question that arises at this point is whether degree interpretations of expletive adjectives maintain their affective connotation or not. Even though we don't have precise data about whether expletive adjectives that are used as degree modifiers still trigger inferences about the speaker's emotions, it seems that they don't necessarily lose this function. That is, by uttering

(10) Frege was **fucking** intelligent.

the speaker may be communicating that Frege was very very intelligent and, at the same time, that she is in a heightened emotional state at the utterance's context. If that is the case, the degree interpretation of *fucking* would come in addition to a still active affective interpretation.

4.2.2 Previous proposals

Morzycki (2011) analyzes why expletive adjectives (interpreted as degree modifiers) are incompatible with AP-modifying measure phrases:

(11) #Rufus is seven feet {goddman, fucking, bloody} tall.

For the purposes of his investigation, he assumes that expletive adjectives and expletive adjectives interpreted as degree modifiers constitute two different, but homophonous, classes of modifiers. In that sense, the same item (e.g., *fucking*) would be systematically ambiguous between affective (i.e., truth-conditionally empty) and degree (i.e., truth-conditionally relevant) interpretations. The (simplified) versions of their respective denotations would be the following, where (12a) represents *fucking* as a function that takes an argument x and outputs that the speaker c_s feels in a certain way with respect to it at the utterance context c, and (12b) represents it as a function that takes an argument and denotes that the speaker 'stands in the fucking relation to it':

(12) a. $[\![\text{fucking(x)}]\!]^c = \lambda\text{x}.[\![\text{x}]\!]^c \bullet \textbf{feel}(c_s, \text{x})$
b. $[\![\text{fucking(x)}]\!]^c = \lambda\text{x}.\textbf{fucking}(c_s, \text{x})$

There are reasons however to be unsatisfied with the assumption that *fucking* is ambiguous between degree and non-degree interpretations. First, if degree uses of expletive adjectives were conventionalized as part of the lexicon, one would expect there to be translation differences among expletive adjectives of different languages. However, as observed in the last section, this is generally not the case: expletive adjectives in different languages systematically receive degree interpretations when they modify scalar adjectives or nouns.

Second, as we observed in the last section, there are reasons to believe that expletive adjectives don't lose their affective impact when they have degree interpretations. Assuming that items such as *fucking* are ambiguous between exclusively affective and degree interpretations would make the latter interpretation incompatible with the inference that the speaker is in a heightened

emotional state. This is not what we intuit, though: in (5a), here repeated, *fucking* can simultaneously indicate that Frege was more intelligent than Einstein, and at the same time that the speaker feels in a certain way about that:

(13) Einstein was intelligent, but Frege was **fucking** intelligent.

These reasons suggest that degree interpretations follow somehow from the affective content of expletive adjectives. That is what I will argue in the next section.

4.2.3 Towards an explanation

As observed in Section 4.2.1, prosodic stress and iconic enrichments can also trigger degree interpretations. However, notice that explaining why this is the case doesn't require postulating an enriched lexicon that would explain why they contribute to an utterance's logical form.

For example, in Schlenker (2017), iconic representations are considered to enrich an utterance's logical form by specifying that the object that an expression (e.g., *long*) is true of (e.g., a conference) should preserve structural properties of that expression. In few words, the longer the predicate *long*, the longer is considered to be the conference:

(14) If the conference is **looooooong**, I will leave.

To handle this pragmatic enrichments formally, Schlenker (2017) assumes that the utterance context makes available a relation of similarity between signs and their denotations. Consider (14). On its standard interpretation, such similarity relation is not available and thus the sign 'long' is interpreted normally. That is, it is interpreted as denoting a function that takes *the-*

conference as argument. However, in the enriched interpretation, the sign 'long'^{k} carries an iconic index k, and as a consequence its normal interpretation comes with an additional iconic requirement to the effect that its length (e.g., the amount of o's it has) correspond to the length denoted by the sign. To capture this, Schlenker (2017b) postulates a condition $sim^{c,w}$(*long'*, 'long'^{k}, g), which requires that in a world w, a similarity relation given by the context c should hold between the property of being long' as applied to g (in 14, *the-conference*) and the iconically interpreted sign 'long'^{k}.

In our proposal, degree interpretations of expletive adjectives also arise as pragmatic enrichments. In (15), the expletive adjective (e.g., *fucking*) enriches an utterance's logical form by specifying that the object it is applied to (e.g., Frege's degree of intelligence) should preserve a structural property of the denotation of *fucking*:

(15) Frege is **fucking** intelligent.

What structural property is that? In Chapter 3, we represented the indexical field of *fucking* as the distribution Pr($fucking|\alpha$), read as 'the likelihood that someone utters *fucking* given that the speaker is in the affective state α'. Now, the states α *fucking* is typically associated with are [A+] states, i.e., state that score high on the arousal dimension (e.g., JOY, ANGER). Now, the arousal dimension can be represented as a bi-polar scale [-1, 1], where negative (¡0) values represent low arousal states and positive (0¡) values represent high arousal states. Therefore, the indexical field of *fucking* ends up being associated with very high values in such scale. Therefore, the structural property of *fucking* that is contextually preserved by *intelligent* in (15) is the property of 'scoring very high' on the scale to which it is associated.

More precisely, inspired on Schlenker (2017), we assume that the utterance context makes available a relation of similarity between the denotation (i.e.,

indexical field) of *fucking* and the argument to which it applies, namely, *intelligent*. Consider (15). On its standard interpretation, such similarity relation is not available and thus *fucking intelligent* only gets its standard, affective interpretation. That is, it updates the context by expressing that the speaker is upset with Frege's intelligence (or happy with it, depending on what is previously assumed about the speaker's affective predispositions). However, in the enriched interpretation, $fucking^k$ carries an iconic index k, and in consequence its affective interpretation comes with an additional iconic requirement to the effect that the degree of arousal it is typically associated with (e.g., its high value in the bi-polar scale [-1, 1]) is similar to the degree in which its argument (i.e., *intelligent*) is instantiated by Frege. Therefore, we postulate a condition $sim^{c,w}$(*fucking'*, *intelligent*, g), which requires that in a world w, a similarity relation given by the context c should hold between the indexical field of $fucking^k$ (as applied to *intelligent*) and the degree in which *intelligent* is in turn instantiated by its argument g, *Frege*.

In this way, we can explain how degree interpretations are triggered, without appealing to ambiguous lexical items. Moreover, in this way it becomes clear why degree interpretations come as an *addition*, rather than in place of, to a still active affective interpretation, and why such phenomenon can be observed cross-linguistically.

4.3 Non-local readings

4.3.1 The data

In Potts (2004), it is observed that expressive's syntactic realization in a sentence sometimes seems to differ from their scope of semantic interpretation. For example, in (16), *damn* seems to apply to the embedded sentence rather than to *telephone*:

(16) Nowhere did the instructions say that the **damn** telephone didn't come with a plug!
→The speaker feels negatively about the telephone coming without a plug.

However, the cases in which these readings arise are not further explored until Frazier et al. (2014). In their study, the authors present experimental evidence in favor of the idea that individuals interpret expletive adjectives taking into account two pragmatic factors: i) the position of the expletive adjective in the host utterance and ii) what element of the host utterance can be construed as a causal agent. On the one hand, participants were more likely to interpret expletive adjectives as expressing emotions about those elements in the utterance which occur closer to them. For example, in (17), *fucking* is more likely to be interpreted with respect to the subject DP than with the object DP in virtue of its proximity to the former:

(17) The **fucking** dog was playing with the cat.

On the other hand, the study show that participants tend to interpret expletive adjectives as communicating emotions about those elements in the utterance which can be seen as causing the situation described by the utterance. In (18), *fucking* is more likely to be interpreted with respect to subject DP than with the object DP because the former, but not the latter, refers to an object that can be seen as causing the situation described:

(18) The dog was playing in the **fucking** couch.

However, Gutzmann (2019) argues that there are also grammatical constraints on the possible interpretations an expressive may receive. In his study, he presents experimental evidence showing that some readings of ex-

pletive adjectives are much less preferred than others. For example, when an expletive adjective such as *damn* occurs in the scope of a speech-report predicate such as *say*, participants judge it to indicate the speaker's feelings towards its syntactic sister (19a) or the embedded clause (19b), but not so much towards the subject of the matrix clause (19c) or the matrix clause itself (19d). The matrix clause interpretation, for example, was only chosen in 10-15% of the cases as the most probable interpretation of *damn*:

(19) Peter said that the dog ate the **damn** cake.

a. The speaker is upset with the cake.
b. The speaker is upset with the fact that the dog ate the cake.
c. ??The speaker is upset with Peter.
d. ??The speaker is upset with Peter saying to him that the dog ate the cake.

According to Gutzmann (2019), this implies that 'syntactical embeddings of an EA [expressive adjective] have a very strong blocking effect' (p. 103). However, Bross (2021)'s study shows that the readings in (19c,d) become more salient when more information about the utterance context is provided. For example, if the common ground entails that the speaker has a positive attitude towards referents appearing in the utterance except for the subject of the matrix clause, then the possibility that (19c) is the intended interpretation raises up to approximately 50%. Even though such results are not clear-cut, Bross (2021) argues, they show that syntactic embeddings don't constitute barriers blocking expressive's interpretations.

It is worth noting that, in both Frazier et al. (2014) and Gutzmann (2019) studies, participants are asked about the most probable interpretation of an expressive from a group of predefined options (i.e., are asked to answer 'What does the speaker most likely judge as negative?'). However, in this

kind of experimental setting, participant's responses do not necessarily mean that their preferred option are 'perfect', but that they are the best given the available options. Indeed, the type of question that the participants are asked presupposes that *damn* can only have one possible interpretation when it is uttered. Yet, that is not what we intuit: the interpretations in (19a-d) are not in competition with each other in terms of informativeness, so it seems possible that they may simultaneously arise with different degrees of probability. For example, if the common ground entails that the speaker has a negative attitude towards all the elements in the utterance except for the cake, the speaker can be interpreted as probably conveying all the readings in (19b-d), i.e., as expressing his frustration towards many elements at the same time.

Before moving to the next section, notice that, in the literature on expressives, non-local interpretations have been only theoretically and empirically studied in expletive adjectives like *damn*. However, non-local interpretation can also be attested in other kinds of expressive items. Observe (20). In standard situations, the Spanish diminutive suffix *-ito* expresses the speaker's affect towards (the object referred by) the noun modified, namely, the dog:

(20) El perr-**ito** está ladrandole al gato.
'The dog-DIM is barking at the cat.'

However, in embedded utterances, the suffix can also signal the speaker's affection towards the situation described by an embedded clause rather than the (object referred by) the noun it modifies. For example, in (21), the suffix can be interpreted as expressing the speaker's joy towards the event described by the embedded utterance:

(21) Nadie me dijo que las empanad-**itas** estaban a mitad de precio!

'Nobody told me that the empanadas-DIM were half-price.'

Moreover, following Frazier et al. (2014), it seems that the readings that the diminutive suffix end up receiving depend on the interplay between the position of the suffix and what element of the host utterance can be seen as causing the situation described. In (20), for example, the suffix can be interpreted as signalling an affective state towards the dog due to its proximity. Yet, in (22), it can be interpreted towards the cat because it can be seen as provoking the situation described:

(22) El gato está durmiendo en su cam-**ita**.
'The cat is sleeping at his bed-DIM.'

If these observations, which should be empirically tested in future work, are correct, then they would indicate that an explanation of expletive adjective's non-local readings need to be structurally similar to that required to understand non-local readings of affective suffixes. Before moving to (the basis) of such an explanation, let's see the theories that have been proposed to explain non-local readings.

4.3.2 Previous proposals

About cases such as (16), repeated below as (23), Potts (2004) briefly suggest that expressive adjectives denote functions that may ignore the syntax and thus take as argument any sub-part of the host utterance (p. 166). However, this idea merely echoes the observation that expressives may disobey the syntax rather than providing a way of explaining it. Another way to say this is: expressives can take as arguments many elements in the host utterance because they are syntactically flexible, but what explains such flexibility?

(23) Nowhere did the instructions say that the **damn** telephone didn't come with a plug!

In order to account for non-local interpretations, Frazier et al. (2014) argue that expletive adjectives are not compositionally integrated in the host utterance and thus can ignore the syntactic structure of the sentence in which they occur. According to this view, uttering expletive adjectives constitutes a separate speech-act that expresses the speaker's global feelings (e.g., that the speaker is upset in general). These feelings are then interpreted with respect to particular constituents of the host utterance via purely pragmatic mechanisms, such as the proximity and causal factors mentioned before. For example, in (24), *fucking* would merely indicate that the speaker is upset, but due to its proximity with *dog*, it can be conversationally interpreted as expressing that that the speaker is upset with the dog:

(24) The **damn** dog is sleeping.

In other terms, Frazier et al. (2014) consider that an utterance such as (24) is semantically equivalent to (25), where *damn* occurs as a stand-alone utterance:

(25) **Damn**! The dog is sleeping.

Even though Frazier et al. (2014)'s theory seems adequate to explain why expletive adjectives can receive many interpretations depending on contextual factors, it may be too extreme. The main reason being that expletive adjectives *are* compositionally integrated with the host utterance. That is, they are part of the same syntactic and, therefore, compositional structure of the host utterance, rather than independent speech-acts. For example, in Spanish and German, that morphologically mark DP internal agreement,

expletive adjectives partake in agreement phenomena just like any other adjective (Gutzmann, 2019, p. 74):

(26) a. El pinche perro.
'The damn dog'
b. Los pinche-**s** perros.
'The damn dogs'

(27) a. Der verdammt-**e** Hund
'The damn dog'
b. Ein- verdammt-**er** Hund
'A damn dog'
c. Die verdammt-**en** Hund-**e**
'The damn dogs'

Thus, it seems that, in order to account for non-local readings, we need i) to preserve the idea that expressive modifiers are syntactically integrated in the host utterance but at the same time ii) allow pragmatic factors (e.g., the position of the expressive) guide different interpretations in a context.

In order to meet these two requirements, Gutzmann (2019) proposes to analyse the interpretation of expletive adjectives as a syntactic rather than compositional phenomenon. Roughly speaking, he considers that expletive adjectives like *damn* carry a valued but uninterpretable expressive feature, whereas the head of the constituents which can be the target of the speaker's attitude come with an unvalued but interpretable expressive feature. What is interpreted is thus a syntactic expressivity feature that needs to be in a c-command relation with the expletive adjective *damn* (within a single CP). As a result, some readings of *damn* are ruled out, namely, those in which the interpretable and uninterpretable feature do not stand in a c-command relation. However, as mentioned above, Bross (2021)'s study cast doubt on

whether some readings are really syntactically blocked, so Gutzmann (2019) account needs to explain why the syntactic blocking appear to be relaxed once contextual information is provided.

As it can be observed, in trying to account for the non-local interpretations of emotive expressives, theorists depart from the common assumption that expressive adjectives are interpreted through a mechanism of functional application. In the case of the speech-act view, expressives are not syntactically integrated in the host utterance, so no application can take place. In the case of the syntactic view, expressives are syntactically integrated but receive their interpretations through a mechanism of upward agreement instead of a mechanism of functional application.

4.3.3 Towards an explanation

The following proposal to understand expletive's non-local interpretations derives from the idea that, even though events cause emotions and emotions cause affective expressions, reasoning can flow in various directions. First, we saw in Chapter 1.3.2 that lay individuals infer an agent's emotional state on the basis of different affective cues. Then, in Chapter 3, we assumed that expressives like *damn* are interpreted on the basis of what it is known about the speaker's emotional state α ($\mathrm{P}(m_{damn}|\alpha)$). In other terms, following McCready (2012), we proposed that expressive items are interpreted by reasoning 'forward' from emotions to (verbal) emotional expressions:

Now, to explain why expressive adjectives receive non-local interpretations, I propose that these readings require interpreters to reason 'backwards' from an agent's emotional state to the potential events e that caused that state ($\mathrm{P}(e|\alpha)$). In other terms, once the listener has knowledge about the speaker's emotional states, he can then reason about the object(s) towards which such state is directed. As Frazier et al. (2014) observed, such reasoning would take into account i) the proximity of the expressive to a constituent of the

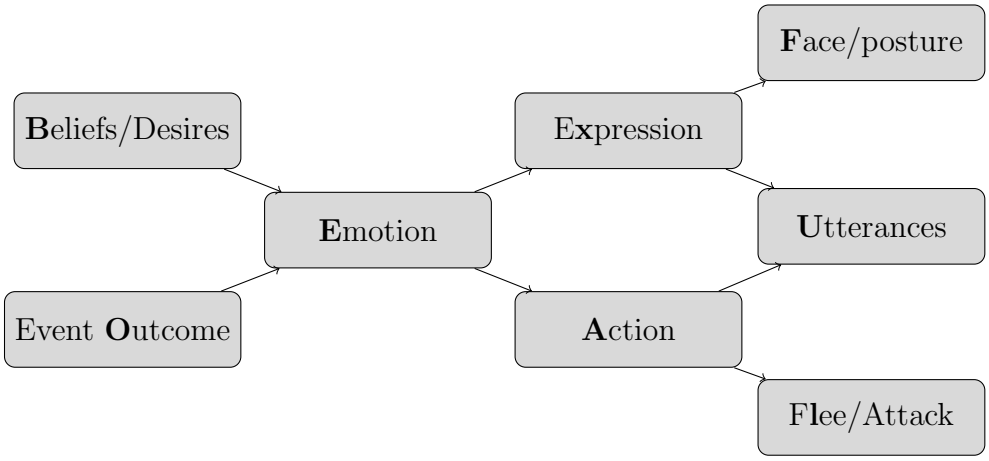


Figure 4.1: Lay theories of emotions (based on Ong et al. 2020)

utterance and ii) whether the constituent can be conceptualized as a causal agent or not.

This hypothesis can be tested with an eye-tracking, Visual World Paradigm. In a recent study, Ronderos and Domaneschi (2022) used the VWP to test whether listeners use expletive adjectives to anticipate an upcoming referent once they know which are the speaker's emotions. In this study, participants were presented with discourse contexts including i) a supporting statement indicating that the speaker holds a negative attitude towards a target referent (e.g., 'Elena hates the hat') and ii) a neutral statement (e.g., 'Elena likes shopping'). Then, participants listened to an utterance where the expressive applied to the target referent *hat* (e.g., 'The delivery man brought us the fucking hat', called 'in-situ' conditions) or to other element (e.g., 'The fucking delivery man brought us the hat', called 'ex-situ' condition) while visualizing four images (one corresponding to the hat and the other to competitor referents). Participants were then asked to choose the correct visual referent.

The results of this experiment show that, when participants know that Elena

hates the hat and then listen to an utterance including an expressive such as *fucking*, they look at the image of the target referent (i.e., the hat) in higher proportion than to the competitor images even before encountering the disambiguating word *hat* in the sentence. Moreover, the results show that this anticipatory effect occurs both for utterances where *fucking* takes *hat* as argument and utterances where it doesn't (i.e., in-situ vs. ex-situ conditions). This supports our general hypothesis that expletive's interpretation can occur by reasoning backwards, i.e., from the speaker's emotional state to the potential objects or causes of such state.

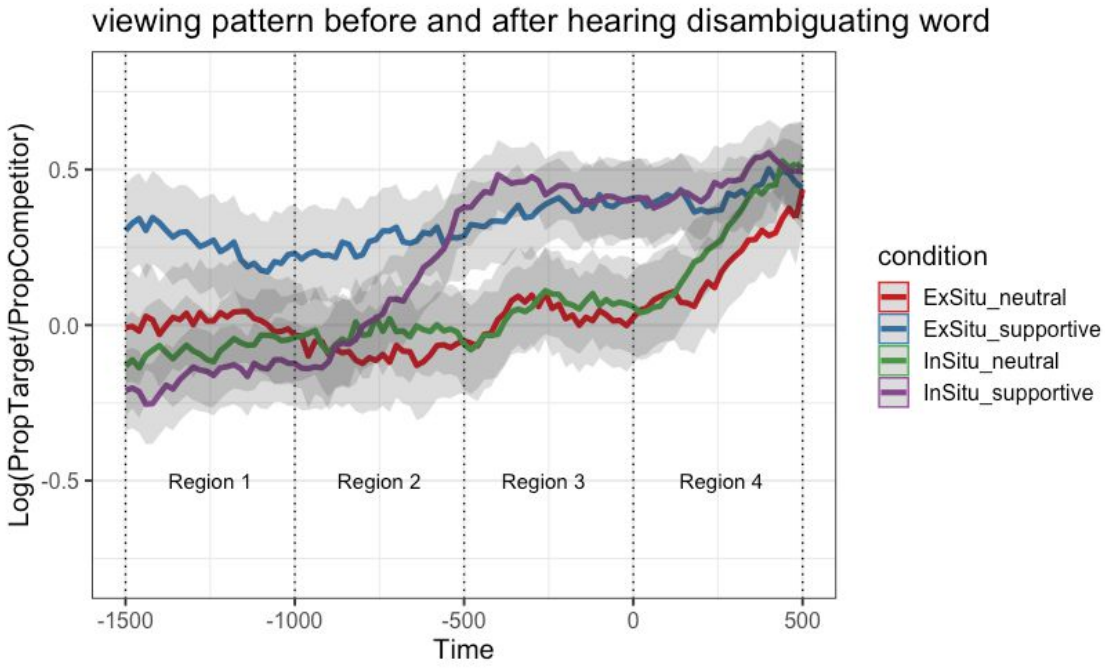


Figure 4.2: Log-gaze probability ratios of looks to target to looks to competitor, time-locked to the disambiguating word 'hat'. Values above zero signify a preference for the target picture. Gray ribbons are confidence intervals. (Ronderos and Domaneschi 2022).

Ronderos and Domaneschi (2022) VWP study can be extended to test the other hypotheses about expressive's non-local interpretations that we have

analysed in this section. For example, we could add a new type of ex-situ condition where the expressive occurs in the scope of a speech report predicate (e.g., *say*), and where the target referent (e.g., *John*) occurs instead in the matrix clause (e.g., 'John says that the dog is in the fucking couch'). If the speaker is presented with a discourse context indicating that the speaker feels negatively towards John, but nonetheless ends up looking in higher proportion at the competitor images while hearing the sentence, then Gutzmann (2019) idea that 'syntactical embeddings of an EA [expressive adjective] have a very strong blocking effect' (p. 103) would be corroborated. Otherwise, i.e., if the speaker looks in higher proportion to an image depicting John after hearing *fucking*, when *fucking* is in the syntactic scope of the speech-report predicate, then Frazier et al. (2014) pragmatic proposal would be corroborated.

Another hypothesis that can be tested in future work is whether *fucking* can signal an attitude towards more than one element at the same time. For example, we could add a new type of discourse context indicating that the speaker holds a negative attitude towards two target referents (e.g., a hat and a bag) and a neutral statement. Then, participants would listen to two types of utterances: i) one where the expressive is applied to a third element (e.g., a shirt, 'The fucking shirt is behind the hat and the bag'), and ii) one where the expressive is applied to either the hat or the bag (e.g., 'The shirt is behind the hat and the fucking bag') while visualizing images of the two targets and other competitor elements. If, while hearing the sentence (i) and (ii), participants look in equal proportion to the two target referents, then we can infer that listener's interpret *fucking* as expressing a 'raw' emotion directed towards both the hat and the bag.

4.4 Conclusion

In this Chapter, I briefly analyzed two properties that expletive adjectives display: degree uses and non-local readings. It was observed that these properties are not unique to them: iconic enrichments and prosodic stress also tend to give rise to degree interpretations, and diminutive suffixes in Spanish can also give rise to non-local interpretation. Thus, in both cases a solution that is structurally similar, and not exclusive to expletive adjectives, has been proposed. In order to explain degree interpretations, I assume that the contexts makes available a relation of similarity between the denotation of the expletive adjective (e.g., *fucking*) and the argument to which it applies (e.g., *intelligent*). To understand non-local interpretations, I appeal to the idea that expressives' interpretation not only consist in using our assumptions about the speaker's emotional states to reason about the expressive's intended interpretation, but also in using our hypothesis about the speaker's emotions to reason about the causes or objects of such emotions.

Chapter 5

Particularistic insults

5.1 Introduction

Particularistic insults (PIs) are expressions such as *bastard* or *wimp* that derogate individuals on the basis of personal or behavioral traits.[1] It has been standardly assumed that PIs convey the speaker's negative attitudes with a high degree of affect as part of their conventional meaning (Potts, 2004; Gutzmann, 2015). Whereas multidimensional approaches classify PIs alongside expressive adjectives such as *damn* or *bloody* (Potts, 2004), unidimensional approaches classify PIs alongside predicates of personal taste such as *tasty* or *boring* (Beller, 2013). However, as mentioned in previous chapters, there are difficulties to assign a single and stable affective interpretation to PIs. On the one hand, the affective information they may display varies considerably across PIs. For example, even though both predicates in (1) can be used to describe John as being unintelligent, (1b) expresses a

[1]Particularistic insults are also known as 'epithets', 'pejoratives', 'derogatives', etc. in the literature. However, I will follow Saka (2007b)'s terminology given that these other expressions have also been used in reference to slurs, which which will be the subject of Chapter 6.

negative attitude with a greater degree of strength than (1a):

(1) a. John is silly.
b. John is a fuckhead.

On the other hand, PIs' affective interpretation also varies within different uses of the same PI. Even though PIs are typically used to convey negative affective states, they can also be felicitously used in contexts where the speaker is not displaying a negative emotion towards the target. For example, in situations where the speaker and interlocutor are close friends, the utterances in (2) may be interpreted as expressing affect rather than contempt towards the addressee:

(2) a. Hey, dumbass, what are you doing?
b. Here's To You, Ya Bastard! You've been such a good friend to me through the years.

Thus, PIs express not a unique but a wide array of different emotions depending on the utterance context. In order to accommodate these and other types of affective variation (to be analyzed shortly) in a unified theory, I propose that PIs' affective meaning is indexical (i.e., multilayered and associative) (Beltrama, 2020) rather than conventional. In other terms, I propose that PI such as *bastard* are indexically associated with a range of different affective qualities (derived, in turn, from the pleasure and arousal dimensions), anyone of which may potentially emerge to different degrees depending on interpreter's prior assumptions about the speaker's general affective states and/or his relation with the target of the insult.

In Section 5.2, I analyze different types of PIs according to i) how descriptive they are (thin vs. thick), ii) how strong they are (soft vs. strong), iii) how

positive or negative they are, and iv) how they may be used in a sentence (as predicates vs. modifiers). On the basis of such analysis, in Section 5.3 I argue that PIs' lexical meaning is neither expressive nor evaluative, among other reasons, because their affective impact is absent in many contexts (e.g., saying 'John is not a bastard' may, but need not to, express the speaker's emotions). Then, in Section 5.4 I analyze PIs affective meaning as indexical, that is, as given by the set of affective states that the PI has the potential to express at a given context. Thus, inspired on Burnett (2017, 2019)'s model of social meaning, I analyze PIs interpretation as based on a PI's stereotypical indexical meaning and the listener's prior assumptions about the speaker's affective stance. Section 5.5 concludes.

5.2 The empirical landscape

Despite the increasing interest on the semantics and pragmatics of derogatory expressions, PIs have typically not attracted much attention. In the linguistic literature, PIs appear in discussions of gradability (Bolinger, 1972; Morzycki, 2009) and epithets, in which PIs tend to be the primary component (Jackendoff, 1972; Umbach, 2001; Schlenker, 2005; Potts, 2004). In the philosophical literature, PIs figure in discussions of slurs, with which they are often compared and contrasted: whereas PIs target individuals in virtue of their personal traits, slurs target in virtue of their belonging to a relevant social group. In recent years, however, researchers have started to focus on how PIs convey affective states, and what restrictions they impose on the context, from theoretical and empirical perspectives (Cepollaro et al., 2021a). In this section I follow this recent trend, by exploring the diversity of (uses of) PIs that can be identified according to their descriptiveness (Section 5.2.1), their strength (Section 5.2.2) and their valence (Section 5.2.3). Given that PIs have been often studied in connection to epithets, I will end this discussion

by comparing and contrasting them (Section 5.2.4).[2]

5.2.1 Thin vs. thick

As observed in the Introduction, PIs are typically negatively valenced. Since Williams (2006), valenced expressions (and concepts) are divided into 'thin' and 'thick'. On the one hand, thin ethical predicates assess the object of predication as good or bad, yet they don't specify in which way the object is good or bad. For example, qualifying a carnival as impermissible qualifies it as bad but doesn't indicate in virtue of which property it is bad. Thick ethical terms, on the other hand, assess the object of predication as good or bad and, additionally, specify the way in which it is good or bad. Qualifying a carnival as lewd qualifies it negatively in virtue of being sexually explicit.

Does the thin-thick distinction applies to PIs? PIs such as *bastard*, *asshole* or *jerk* (also called 'all-purpose pejoratives') express the speaker's negative evaluations about an individual but are 'too lean' to indicate the descriptive dimension along which the target is so evaluated (Jeshion, 2020). In contrast, PIs such as *wimp*, *crook* or *nutter* express the speaker's negative evaluations and, in addition, indicate the descriptive basis of such reaction -roughly speaking, being fearful, dishonest or eccentric, respectively. Therefore, even though all PIs express negative evaluations of individuals *qua* individuals (and not as a result of their membership in a socially relevant group, like slurs), we can distinguish them in thin and thick.

However, notice that the thin-thick distinction doesn't straightforwardly apply to all PIs. First, various PIs can have thin or thick interpretations depending on the utterance context. In some situations, the use of *stupid* may merely express the speaker's negative evaluation of the target. In others,

[2]In what follows I only focus on adjectives (e.g., *stupid*) and nouns (e.g., *bastard*). However, it is important to note that PIs also include verbs (e.g., *to fuck up*), for which there has been even less interest in the literature. Some exceptions are Gutzmann and McCready (2016) and Hom (2012)

however, *stupid* may also describe the target as being unintelligent or naive. The former type of interpretation can be more clearly observed in cases where *stupid* applies to an object which can't be described as (un)intelligent, as in 'I don't like this stupid weather', and the latter in cases where it can, as in 'The stupid student failed the exam'. Therefore, the thickness of some PIs may vary with respect to the object of the predication and other contextual factors.

Second, various PIs acquire their evaluative effect based on the metaphoric association between the target and things that are stereo-typically seen as negative. For example, some PIs associate individuals to a genealogy (*bastard*, *son-of-a-bitch*), body-parts (*asshole*, *dick*), animals (*chicken*, *cockroach*), etc. Some of these metaphoric associations, through repeated circulation and use, may undergo a certain degree of conventionalization within a linguistic community. When the associations are highly conventionalized, a PI can be classified as thin or thick without difficulties. For example, derogatory uses of *chicken* incorporate a stable descriptive component (i.e., being fearful) based on stereotypical assumptions about chicken's behaviors, and can thus be classified as thick. However, when such associations are still fluid and open-ended, the thin-thick distinction cannot be profitably applied to a PI. For example, derogatory uses of *cockroach* may be interpreted as thin (e.g., as indicating that the target is just bad) or thick (e.g., as indicating that the target is physically unattractive) depending on how the speaker recruits its different connotations.

Now, the thin-thick distinction qualifies predicates according to how descriptive they are. However, PIs not only vary in terms of descriptiveness, but also in terms of how 'strongly' they evaluate their targets. To wit, judging someone as a fuckhead may express or elicit a stronger emotional impact than judging him as fool, even though both expression can be used to describe the same type of person, namely, one with poor intellectual skills. In

contrast, non-insulting valenced expressions that describe the same state of affairs don't seem to vary in terms of force: judging a concert as *lewd* doesn't seem to express stronger or weaker emotions with respect to judging it as *lascivious* or *obscene*. What explains the different degrees of force associated with PIs?

5.2.2 Soft vs. strong

PIs not only vary with respect to how descriptive they are, but also with respect to how 'strongly' they convey the speaker's attitudes. One way to explain this variability is to assume that different PIs incorporate various degrees of valence: to wit, some things are not just bad, but very or extremely bad. However, Janschewitz (2008) has provided empirical evidence that, in average, taboo words (e.g., *bastard*) and evaluative words (e.g., *rude*) don't differ in their valence ratings. In particular, this study found that taboo words (including PIs) typically yield higher arousal ratings than evaluative words, thus providing evidence supporting the idea that PIs' 'strong emotionality' comes from arousal rather than valence (Jay, 2000). Therefore, even though differences in valence among PI expressions may have an impact on how powerful PIs are perceived, we will distinguish soft from strong PIs by appealing to the degrees of arousal they typically express and/or elicit in others.

Variation in force comes in at least two types: there is variation in force across different PIs (henceforth 'type I variation') and across uses of the same PI (henceforth 'type II variation'). Type I variation is exemplified by PIs that are to some extent truth-conditionally equivalent, but differ in the degree of excitement or energy they typically express and/or elicit in others. As mentioned above, while both *fuckhead* and *silly* can be used to describe the target as unintelligent, the former is correlated with a greater degree of arousal than the latter.

Type II variation, in turn, is related to how powerful the same word can be interpreted depending on contextual factors. For example, *fuckhead* can have a stronger or weaker impact depending on whether the user employs a contemptuous intonation or not, or depending on whether user and target are know each other or not. To have a rough grasp of how these two factors interact with each other, observe the following examples, inspired on Popa-Wyatt and Wyatt (2017) (the superscripts F and C refer to a friendly and contemptuous intonation, respectively):

(3) a. Speaker to unknown person: 'You are a total [fuckhead]^{C}!'
b. Speaker to unknown person: 'Hey, [fuckhead]^{F}, let me pay you a beer!'
c. Speaker to a friend: 'You are a total [fuckhead]^{C}!'
d. Speaker to a friend: 'Hey, [fuckhead]^{F}, let me pay you a beer!'

With respect to intonation, the contemptuous uses in (3a) an (3c) are clearly more arousing (e.g., offensive, harmful, etc.) than the friendly uses in (3b) and (3d): through the contemptuous intonation and content asserted, the speaker makes overt his intention to derogate the target. However, with respect to the relation between speaker and addressee, the variation is more complex: all other things being equal, (3a) can be seen as slightly more arousing than (3c), since being insulted by a stranger may carry a clearer threat of violence than being insulted by a close acquaintance. Similarly, despite their friendly character, (3b) seems potentially more arousing than (3d), arguably because friendly uses of a PI performed by strangers can be easily interpreted as covert forms of aggression or harassment. Even though these observations need to be empirically tested in future work, they illustrate how the same PI (e.g., *fuckhead*) can be associated with different degrees of arousal depending on how they are used and by who.

Interestingly, the use of *fuckhead* in (3d), namely, its use among friends in contexts where there are various signals that the speaker is well intended (e.g., 'let me pay you a beer'), can make the PI express positive, endearing or friendly emotions rather than negative ones. This would show that PIs not only vary in their degree of descriptiveness, nor the degree of arousal they can express, but also that, in certain contexts, they can also completely shift their valence. The next section focuses on this type of variation.

5.2.3 Negative vs. positive

So far, we have observed that PIs trigger complex affective inferences. As illustrated in (4), by uttering *stupid*, the speaker typically intends to express a negative evaluation with an additional high degree of arousal:

(4) Hey, stupid, what are you doing?

However, even though *stupid* is typically associated with negatively valenced affective states (e.g., anger), it can also be interpreted as friendly, playful, non-face threatening in certain contexts. To wit, in a context where speaker and addressee know each other, and where there are various cues that the speaker is well intended (e.g., intonation, gestures, etc.), (4) can flip to a positive interpretation. How can we explain this phenomenon?

It may be argued that PIs are in fact ambiguous (or polysemous) between negative and positive interpretations. Indeed, it seems to be the case that listeners need to reason about the speaker's intentions and other contextual cues in order to recover the affective message of a PI. However, ambiguity fails to explain why positive interpretations only arise in situations where speaker and listener are close acquaintances, or at least presume to be so. To wit, ambiguous expressions (e.g., *bank*) don't impose constraints on who can use them to express one or other of their contents.

Now, it might be also considered that positive interpretations involve a type of simulation. Since the notion of simulation covers many different types of abilities, this view can be understood in at least two different ways. First, according to the 'irony' approach, positive uses of *stupid* arise when the speaker makes manifest that he wants to convey a message that is the opposite of what he literally said: saying 'What a terrible cake!' after devouring it can be interpreted as praising rather than criticising it. However, the main problem with an approach in terms of irony is that positive uses of *stupid* don't necessarily convey a message that is the opposite of what is literally said: by uttering (4), the speaker is not trying to convey that the addressee is intelligent. In other words, in positive uses of (4), the negative description associated with *stupid* survives, but nonetheless doesn't express a negatively valenced state.

Second, according to the 'echoic' approach, positive uses of *stupid* would arise in situations where the speaker makes manifest that he is attributing the responsibility of the utterance to some other agent, actual or potential, in order to express a critical attitude towards it: if the speaker refers to a soldier as a *freedom fighter* after making it clear that he doesn't agree with the positive connotations of such label, he may be interpreted as expressing a negative (e.g., mocking) attitude towards its referent. Thus, the echoic approach would be able to explain why (4) expresses a positive attitude without necessarily conveying that the target is intelligent. However, the same problem looms: the echoic account doesn't explain why positive interpretations of PIs require speaker and addressee to be close or at least presume to be so: in contrast to the echoic uses of *freedom fighter*, echoic uses of *stupid* impose a restriction on who can use it.

Positive uses of PIs do involve some form of simulation which is not ironical nor echoic. Instead, we should understand positive uses of PIs as a form of 'mock aggression'. Mock aggression is an activity that is structurally similar

to serious aggression but that lacks its harmful effects. Instead, mock aggression has been observed to induce positive emotional states, relieve negative emotions (e.g., stress), and even to have developmental outcomes, namely, enhance affiliation and social skills (Smith and Boulton, 1990). Importantly, mock aggression typically happens in conjunction with other affective cues that signal the lack of harmful intent (e.g., positive facial expressions such as smiling) (Driver and Gottman, 2004) but has a risky character. To wit, it often occurs among people that is already close, and have thus developed a 'script' or 'insider knowledge' for mock aggressive behavior (Ballard et al., 2003). Otherwise, even in the presence of positive cues such as laughing and smiling, mock aggression is likely to slip into serious aggression.

Thus, the positive use of (4) can be understood as a form of mock aggression. Even though positive interpretation of (4) typically co-occur with cues indicating the speaker's lack of harmful intent (e.g., positive gestures, friendly intonation, etc.), this interpretation requires speaker and addressee to have developed a relationship in which the use of a PI doesn't has a harmful effect without losing its derisory descriptive meaning. In that sense, positive uses of PIs arise in those contexts where speaker and target are psychologically close or presume to be so, so it can be viewed as testifying to 'real friendliness' (Radcliffe-Brown, 1940).

Now, before moving to the next section, I will end this review by comparing PIs with epithets, given that they have been typically studied in conjunction in the semantic and pragmatic literature.

5.2.4 PIs vs. epithets

PIs are often referred to as 'epithets'. However, 'epithet' may refer to two different phenomena (cf. the Merriam-Webster Dictionary):

A: a characterizing word or phrase accompanying or occurring in place of

the name of a person or thing.

B: a disparaging or abusive word or phrase.

A picks up expressions such as *Alexander the Great*, *Earvin Magic Johnson*, *Obummer*, etc. Thus, A's definition is closer to the one corresponding to nicknames, that is, expressions that modify or replace a standard name and which doesn't need to be negatively valenced or derogatory. In contrast, B picks up particularistic insults such as *bastard* and slurs such as *Wop*, independently of whether they accompany or replace the name of a person or thing.

While discussing the syntactic and semantic features of PIs, researchers have primarily focused on constructions where a PI occurs as an adnominal modifier or replaces a noun (Kaplan, 1998). That is, on utterances such as (5), where a disparaging word (e.g., *bastard*) accompanies or replaces a given name (e.g., *John*) :

(5) a. That {bastard, jerk, asshole} John got promoted.
b. The {bastard, jerk, asshole} got promoted.

However, focusing on these constructions (henceforth 'A+B constructions') risks of making unclear which properties of a PI derive from its lexical meaning and which are contingent on the syntactic environment in which the PI occurs. For example, by analyzing A+B constructions, it has been argued that PIs exclusively contribute expressive content, that is, content about the speaker's affective states that have no truth-conditional impact. To wit, PIs that occur as modifiers can be omitted without altering the utterance's truth-conditions:

(6) That {bastard, jerk, asshole} John got promoted.

$\cong$ That bastard got promoted.

However, it is worth noting that, even though PIs that occur as adnominal modifiers have a strong preference to receive truth-conditionally irrelevant interpretations, they can also contribute to the host's utterance truth-conditions. To wit, in those contexts where a PI i) introduces rhetorical information or ii) answers the current Question Under Discussion, a PI typically restricts the denotation of the noun it modifies, thus becoming truth-conditionally relevant (Martin, 2014). In (6a), *jerk* picks up one among Alex ex-boyfriends in order to answer the QUD, and thus can't be omitted without altering the host utterance's truth-conditions (and grammaticality). In (6b), *bastard* picks up a subset of Alex's Facebook contacts in order to explain why they were eliminated, and thus can't be omitted as well:

(7) a. A: Which one of Alex's ex-boyfriends did you see?
B: The jerk one.

b. Alex eliminated all her bastard contacts from Facebook.
$\rightarrow$Alex eliminated those contacts *because* they where bastards.

Therefore, we should be careful in distinguishing the semantic contribution of a PI from the contextual factors that make it truth-conditionally relevant or not. What is more, the fact that PIs can contribute truth-conditional and not only expressive information can be clearly attested in the case of thick PIs such as *wimp* or *crook*, which have a clear descriptive component. In sum, to understand the semantics and pragmatics of PIs, we should observe how they behave in different syntactic environments and not only as epithets (Cepollaro et al., 2021a; Stojanovic, 2021).

5.3 Previous accounts

5.3.1 The expressive view

According to Kaplan (1998), thin and thick PIs contribute different types of content. On the one hand, thin PIs such as *bastard* contribute expressive content, that is, only display the speaker's heightened emotions about an individual. On the other, thick PIs such as *wimp* are descriptive, 'though descriptive of properties that are generally seen as personal filings' (p. 25). Thus, the denotation of *bastard* can be roughly specified by the set of contexts in which its use is felicitous, whereas the denotation of *wimp* by the set of worlds in which it is true (following Gutzmann (2015) we use the interpretations functions t and e to distinguish truth-conditional and expressive content, respectively):

(8) a. $[\![\text{bastard(x)}]\!]^e$: {c: c_s is upset with x in c_w}
b. $[\![\text{wimp(x)}]\!]^t$: {w: x is fearful in w}

Kaplan's thesis has been later defended by appealing to the 'descriptive ineffability' of thin PIs. According to Blakemore (2011), when individuals are asked about the conceptual meaning of a term like *bastard*, they tend to illustrate the contexts where uttering *bastard* would be felicitous, rather than provide a conceptual definition (Blakemore, 2002; Potts and Roeper, 2006). However, two issues arise. First, as Geurts (2007) points out, descriptive terms such as *the* or *green* can also be hard to define in conceptual terms, so ineffability does not correctly distinguish expressive from descriptive denotations. Second, Hyatt et al. (2017) study presents extensive evidence that individuals do in fact associate descriptive properties with those they consider stereotypical assholes or dicks. Namely, being arrogant, distrustful, selfish or manipulative (that is, personality traits that score low on agreeableness).

Now, even if we still assume that thin PIs are only associated with use-conditional content, how do their formal treatment would look like? In Potts (2004), thin PIs (e.g., *bastard*) are considered to denote functions that take a descriptive argument (e.g., *John*) and return i) the same descriptive argument unmodified and ii) an expressive proposition of the form 'the speaker feels upset towards John'. In proof-style notation, the DP 'that bastard John' can be roughly translated as follows (the metalogical bullet '•' isolates the descriptive and expressive contributions of the thin PI at the same line of the proof, and **bad** is a function that says, roughly, 'the speaker is in a heightened emotional state regarding x'):

$$\text{(9)} \qquad \text{that bastard John} = \frac{\dfrac{\text{bastard}}{\text{bastard} : \langle e, \varepsilon \rangle} + \dfrac{\text{John}}{\text{John} : e}}{\text{John} : e \bullet \text{bad(John)} : \varepsilon}.$$

A problem with this view is that it only applies to thin PIs that occur as adnominal modifiers -and, among these, only on those that have a non-restrictive interpretation (Schlenker, 2007; Hom, 2012) (see also Section 2.3.1). To wit, Potts (2004)' bi-dimensional treatment predicts that the expressive inference triggered by thin PIs is projective, i.e., that it always survives as an utterance implication when the PI occurs under the syntactic scope of an entailment cancelling operator (e.g., negations, if-clauses, modal operators) (Simons et al., 2010). However, that is not what we observe. As (10 illustrates), when *bastard* occurs as a predicate, it no longer triggers the projective inference that the speaker evaluates negatively the target or that he is in a heightened emotional state:

(10) John is a bastard.

a. John is not a bastard.

b. If John is a bastard, then I won't call him.

c. John might be a bastard.
   $\not\rightarrow$The speaker is upset with John.

In sum, we have observed that the semantic expressive account faces two important challenges: i) it proposes radically different treatments for thin and thick PIs, which otherwise only seem to differ in their degree of descriptiveness ii) its formal treatment of thin PIs only applies to PIs that appear as (non-restrictive) adnominal modifiers, but not in other syntactic environments. To these two issues, we may add the fact that the expressive semantic account remains silent about the variation in force across PIs (i.e., the fact that some PIs express more intense emotions than others) and about the fact that some PIs can receive positively valenced interpretations.

5.3.2 The evaluative view

According to Beller (2013), thin PIs such as *jerk* are evaluative. Evaluatives constitute a class of expressions that carry appraisal and figure in value judgments (Vayrynen, 2013, p. 29). It includes, among others types of terms, predicates of personal taste (henceforth PPTs) such as *delicious* or *fun*. Now, it is standardly agreed that evaluatives give rise to 'faultless disagreements', that is, situations where two parties appear to disagree but where there doesn't seem to be any objective way of telling who is right (Kolbel, 2002). This seems also to be the case for PIs. For example, in the two following exchanges, none of the parties seem to have the last word about whether John is fun/a jerk:

(11) a. A: John is fun.
        B: No, he isn't.
     b. A: John is a jerk.
        B: No, he isn't.

An immediate worry that arises is that, as Sundell (2016) points out, non-evaluative gradable terms such as *bald* or *sharp* can also give rise to faultless disagreements. For example, in discussing whether John is bald or not, A and B may agree in which the facts are (i.e., what's the amount of hair John has) but still disagree, e.g., about what the threshold for *bald* should be in the context, or which dimensions should be relevant to qualify someone as bald (e.g., total amount of hair, its distribution, its thickness, etc.). Therefore, the observation that a term gives rise to faultless disagreements doesn't warrant that it should be classified alongside PPTs such as *fun*.

(12) A: John is bold.
B: No, he is not.

Yet, if we assume that thin PIs are evaluative, how would their formal treatment look like? Beller (2013) proposes to capture the evaluative character of thin PIs in a relativist semantic framework. Under that approach, statements containing an evaluative predicate are true not only relative to a world and time of utterance, but also relative to a judge parameter j. That is, to an agent whose taste or opinion dictates whether the evaluative judgment is true or false. This move allows to account for the intuition that, in discussion such as (11b), what A asserts is true with respect to his own standards and false with respect to B's standards (and vice-versa) thus explaining why A and B's disagreement is faultless:[3]

(13) $[\![\text{John is a jerk}]\!]^{c,w,j} = 1$ iff John is a jerk in w according to j.

The main advantage of the evaluative view is that, in contrast to the ex-

[3]Beller (2013) also analyzes two other properties of PIs, namely, their behavior-dependence and gradability. However, I won't discuss such properties in what follows, as they to some extent orthogonal to the issue of whether PIs have a relativistic semantics or not.

pressive view, it can be applied to thin PIs independently of their syntactic position. However, an issue that arises is that it doesn't seem easy to apply to thick PIs such as *crook*. Intuitively, whether 'John is a crook' is true or false doesn't depend on how an agent perceives John, but on how John actually behaves (i.e., as a criminal). Moreover, Cepollaro et al. (2021a) have presented evidence that even thin PIs such as *jerk* impose certain constraints on the common ground. To wit, in a context where it is not known whether John actually deserves to be held in low opinion, the acceptability of calling him jerk decreases considerably (compared to, e.g., calling him Italian in a context where it is not known whether he is Italian). Therefore, whether 'John is a jerk' is true or false does not merely depends on the dictates of a judge, but on how John actually behaves.[4]

Yet, perhaps the main issue with an evaluative account is that it doesn't capture the distinctive affective character of PI's. As we saw in Section 5.2.2, PIs' strong emotional impact doesn't come from their evaluative component, but from the degree of arousal to which they are associated (Jay, 2000). To wit, calling someone a jerk not only expresses the speaker's negative evaluation of the target, but also his heightened or aroused state. Now, it might be argued that the arousal inference can be accounted for by appealing to pragmatic mechanisms. For example, it may be argued that, by using *jerk*, the speaker flouts politeness standards and thus, in virtue of such act, indicates that she is in a heightened emotional state. However, if the arousal component is accounted for pragmatically, then it would seem unclear why the evaluative component would have to be semantic in the first place.[5]

[4]This criticism applies to both to contextualist and relativist theories, which share the intuition that subjectivity should be incorporated in the semantic apparatus. The differences between the two amount to how they incorporate such perspective-dependence: whereas contextualist specify it as a judge in the content of the utterances containing evaluative predicates, relativist consider it a value of a parameter in the circumstances of evaluation with respect to which the evaluative utterance is interpreted.

[5]For example, inspired on Vayrynen (2013), we could consider that PIs trigger evaluative inferences in virtue of general conversational mechanism: assuming that it is common

In sum, the main issues with the evaluative approach are two: i) that it is unclear to what extent they pattern with PPTs such as *fun* or *tasty* and ii) that it doesn't account for the arousal component associated with PIs.

5.4 The proposal

As we observed in this chapter, PIs' interpretation vary with respect i) to their descriptive component: some PIs are more descriptive than others; ii) to their arousal properties: PIs indicate a degree of arousal that varies depending on contextual factors; and iii) to their evaluative properties: PIs typically express negative evaluations of the target, but such evaluation can flip to positive in certain contexts. A satisfactory theory of PIs must be capable of bringing together these different types of variation in a compact way, and allow us understand how they arise during the use and interpretation of a PI within a context. In what follows I propose a framework that can serve as the foundation of such a theory.

5.4.1 The semantics

I contend that, at the semantic level, PIs are descriptive predicates. That is, that both thin and thick PIs are descriptive of personal traits that are often considered impairing for the self and inter-personal relationships across contexts; for example, *bastard* describes the selfish, manipulative, arrogant (clustered as 'disagreeable') and *wimp* the fearful. Moreover, PIs are also gradable: they refer to personal traits that come in degrees. Some individuals are wimpier than others, some less bastard than others. Therefore, to evaluate whether 'John is a wimp/bastard' for a truth-value, we need to es-

knowledge i) that *jerk* refers to personality traits that are evaluated negatively and ii) that utterances that include those terms reflect those negative evaluations, then qualifying someone as a jerk will typically express the speaker's own attitudes about the individual target.

tablish a scale that can be correlated with degrees in which the property of being a wimp/bastard may hold at context c. Once the scale is established, we fix a threshold t on that scale, so that 'John is a wimp/bastard' is true iff the degree d to which John is mapped is higher than the threshold t with respect to context c:

(14) a. $[\![bastard]\!]^c = \lambda d_d.\lambda x_e.disagreeable'(x)$ to degree $d \wedge d \geqslant t$ at c.
b. $[\![wimp]\!]^c = \lambda d_d.\lambda x_e.fearful'(x)$ to degree $d \wedge d \geqslant t$ at c.

How can we distinguish between thin and thick PIs in this framework? I propose that the difference arises at the 'meta-semantic level' (Glanzberg, 2007). To wit, when a thin PI is applied to an individual, as in 'John is a bastard', the sentence may have different interpretations depending on the dimensions that are considered relevant in the context. In some context, such statement can be interpreted as saying that John is arrogant. In others, in contrast, it can be interpreted as saying that John is distrustful, or yet that John is both arrogant and distrustful at the same time. Thus, for each situation, there will be various candidate descriptive qualities that can be used to establish a single scale in which individuals can be mapped to degrees of bastardness. In contrast, statements involving thick PIs, like 'John is a wimp', only have a limited number of interpretations across contexts. Even though the sentence can be interpreted as saying that John is anxious or avoids responsibilities, these are particular ways of instantiating the same property, namely, being fearful. In other words, whereas there are many ways of being a bastard, a wimp is always someone considered to manifest a certain degree of fear. Therefore, both thin and thick PIs are descriptive (at the semantic level), but differ in the dimensions the context can make available to establish a common scale for their attribution (at the meta-semantic level).

And what about the affective components associated with PIs? By saying 'John is a bastard', the speaker no only describes John in a certain way, but also signals his negative and aroused attitude towards him. It may be tempting to assume that such affective components are part of the semantic meaning of *bastard*, either as part of its truth-conditional content (as in an unidimensional view) or as part of its expressive content (in a multidimensional view). However, there are two problems with a semantic approach: first, as mentioned in Section 5.4.1, there are situations where the use of a PI does not trigger any inference about the speaker's affective states. By uttering 'I don't think John is a bastard/a wimp' the speaker does not signal any particular affective state towards John.[6]. Second, even it we assume that PIs trigger an inference about the speaker's affective states in virtue of their semantics, then it would remain unclear how to account for the variation in arousal and valence that we observed in Section 2. A semantic account would probably have to appeal to different pragmatic mechanisms to explain why various types of variation arises. In the following section, I propose a more compact explanation.

5.4.2 The pragmatics

How can we account for the typically negative affective inferences triggered by PIs while at the same time leaving open the possibility of different affective interpretations? In what follows, I argue that, at the pragmatic level, PIs such as *bastard* or *wimp* trigger probabilistic inferences based on i) what is previously known about the speaker's affective relation with the target of the PI, and ii) the 'normal' (stereotypical) interpretation of the PI.

[6]It may be argued, though, that by uttering 'I don't think John is a bastard/a wimp' the speaker nevertheless expresses a global negative evaluation of individuals that are disagreeable/fearful (because of being disagreeable/fearful). However, even if this were the case that embedded PIs trigger global evaluations, then it would remain unclear whether such inference arises in virtue of the PI's semantic content, or instead as a conversational implicature.

PIs and indexical fields

PIs trigger affective inferences. By uttering (15), the speaker (henceforth 'Alex') is likely to be interpreted as expressing a negative evaluation with a high degree of arousal towards the addressee (henceforth 'John'):

(15) Hey, bastard, what are you doing?

Therefore, *bastard* is associated with two affective dimensions simultaneously, pleasure (also called 'valence') and arousal. As we saw in Chapter 1.3, in Mehrabian and Russell (1974)'s multidimensional theory of emotions, these dimensions are defined as follows:

- PLEASURE: this dimension serves to measure the pleasure experienced by the subject during an emotional episode. Thus, it corresponds to a scale including negative ([P-]), neutral ([P±]) and positive ([P+]) affective states. It is the evaluative component.

- AROUSAL: this dimension measures the energy of an emotion as provoked by something, that is, how 'excited' or 'heightened' the individual feels upon perceiving a stimulus. Thus, it corresponds to a scale ranging from calm ([A-]) to aroused ([A+]) affective states, where there is no qualitatively 'neutral' arousal ([A±]), which instead can be represented as the absence of an emotion.[7]

Now, the first step to build our model is to use these five affective qualities ([A-], [P+], etc.) to formally define types of affective states. Inspired on Burnett (2017, 2019), we assume a structure $\langle \mathbb{Q}, > \rangle$, where $\mathbb{Q}$ is the set of

[7] In this chapter, I won't consider a third dimension, Dominance. This dimension serves to measure how the subject feels with respect to her environment (e.g., being anxious vs. being relaxed). The reason to set aside dominance is that, by qualifying someone as a *bastard*, it is unclear whether such statement expresses that the speaker feels dominated or dominant towards the person addressed. However, we will use this scale to analyze slurs in Chapter 6.

affective qualities and $>$ encodes relations of compatibility between them (e.g., that an individual cannot be in a [P-] and [P+] state simultaneously):

(16) $\mathbb{Q} = \{[P+], [P-], [P\pm], [A-], [A+]\}$

a. $[P+] > [P-]$
b. $[P+] > [P\pm]$
c. $[P-] > [P\pm]$
d. $[A-] > [A+]$
e. $[A-] > [P\pm]$
f. $[A+] > [P\pm]$[8]

Second, based on this structure, we derive types of affective states α: the [P-, A+] affective state, which we label CONTEMPT, the [P+, A+] state, which we label FRIENDLINESS, etc. Notice that we use these labels instead of those used in Chapter 3 (i.e., ANGER, JOY, etc.) in order to foreground the binding (i.e., social, relational) character of the affective states expressed by particularistic insults. To wit, by uttering (15), Alex does not merely expresses anger (i.e., a strong negative reaction), but contempt towards John (i.e., that John is lowly regarded by him):[9]

(17) Possible affective states α:

Label	CORDIALITY	FRIENDLINESS	DISDAIN	CONTEMPT
α	[P+, A-]	[P+, A+]	[P-, A-]	[P-, A+]

[8]We add the constraint in (16e) because it corresponds to states which can be hardly characterized as affective, such as being sleepy, and the constraint in (16f) because it corresponds to a 'pre-affective' state, i.e., excitement.

[9]However, as mentioned in Chapter 3, we should keep in mind that these labels assemble different types of affective phenomena. For example, CONTEMPT represents [P-, A+] states in general (e.g., anger, rage, hostility, etc.), and not only contempt.

Third, how can we characterize the association between a PI (e.g., *bastard*) and the affective states *alpha* it has the potential to express? As we have observed through this chapter, we cannot assign a single and stable affective meaning to a given PI. To wit, the use of *bastard* may signal that the speaker feels excited ([A+]) but negatively ([P-]) about the target in some situations, but that he feels positively ([P+]) but not necessarily heightened ([A-]) about the target in other situations. For this reason, I will assume that PIs' link to affective states is 'indexical' (Silverstein, 1976; Eckert, 2008; Podesva et al., 2015) rather than conventional. That is, that such relation is grounded on the typical co-occurrence between the use of a PI and a range of affective states, anyone of which could become relevant in a particular context of interaction.

How can we characterize such indexical association? In Jay (1992), it is observed that individuals use curse words, including PIs, approximately 0.7% of the time motivated by negative emotions such as anger, and only marginally motivated by joy or surprise. Thus, we need to capture not only the fact that a term like *bastard* is associated with a range of different affective states, but that it is more strongly associated with some states rather than others. Therefore, in order to characterize a PIs' indexical meaning, I use the notion of 'probabilistic indexical field' proposed in Chapter 3. Namely, I associate a PI such as *bastard* with the probability distribution $\Pr(m|\alpha)$, read as 'the likelihood of uttering m given an affective state α'). This distribution captures which emotions people usually express by using a PI like *bastard*:

(18) Probabilistic field associated with *bastard*:

AFF	CORDIALITY	FRIENDLINESS	DISDAIN	CONTEMPT
$\Pr(bastard\|\alpha)$	0.4	0.5	0.6	0.7

What is the alternative of *bastard*? As we observed in Section 5.4.1, *bastard* denotes a cluster of properties that we have labelled 'disagreeable' (i.e., dis-

honest, manipulative, etc.), so we can assume that speaker's have the option to choose a PI or its non-colloquial counterpart *disagreeable* when describing an individual. As we observe in (19), Pr(disagreeable|α) = 1 - Pr(bastard|α). That is, we assign a low value to *disagreeable* displaying CONTEMPT, and a slightly higher value to *disagreeable* displaying CORDIALITY (due to its more 'polite' or 'formal' character):

(19) Indexical field associated with *disagreeable*:

AFF	CORDIALITY	FRIENDLINESS	DISDAIN	CONTEMPT
Pr($disag|\alpha$)	0.6	0.5	0.4	0.3

Even though these estimations can be made more precise by using psychological or ethnographic studies about the typical use of PIs within a linguistic community, they will be enough for our present purposes.[10]

Fourth, I assume that PIs are interpreted based on what it is assumed about the speaker's affective disposition of the target of the insult. For example, in (15), *bastard* is interpreted by John relative to what he assumes about Alex's affective relation with him. Thus, inspired on Burnett (2017, 2019), I represent the listener's prior beliefs as the relativized probability distribution 'Pr(α)', read as 'the probability distribution that the speaker s feels α with respect to target x'. In a context where the listener doesn't know the speaker, and thus where he has no prior expectations about the speaker's affective relation with the the target (who might be the listener himself or other person), we represent Pr(α) as a uniform distribution over affective states:

(20) Listener's prior beliefs about S's affective relation with the target x:

[10] As mentioned in Chapter 3, the reason not to give 0% to a PI expressing EASE is that, intuitively, the utterance of a PI doesn't completely eliminate the possibility that the speaker may be in one of such states later in the conversation. By uttering a PI one expresses an emotion or mood that can always change with the evolution of a conversation.

AFF	CORDIALITY	FRIENDLINESS	INDIFFERENCE	CONTEMPT
$\Pr(\alpha)$	0.25	0.25	0.25	0.25

Which factors determine the listener's prior beliefs about the speaker's affective states? As we will see in the next section, in the case of PIs, listener's priors are mainly determined by the following two types of factors:

- The speaker's general emotional state (which we can call its 'mood'): the speaker's mood represents the publicly accessible emotional aspect of the speaker, which arises by publicly visible actions (e.g. his intonation, facial gestures, body posture, etc.).
- The speaker's psychological closeness with the target: this factors is determined by their past experiences and interactions (e.g., the frequency with which speaker and target engage in mock aggressive behavior).

In a context, both factors will influence the listener's expectations about the speaker's affective stance towards the PI's target. However, as mentioned in Chapter 3, I remain neutral about the epistemological problem of deciding whether an agent's prior beliefs are constrained by rational principles (as objective Bayesians claim) or by non-rational processes such as socialization, evolution or free choice (as subjective Bayesian claim) (Talbott, 2001). Instead, the aim of our model will be explaining why, even though PIs are typically interpreted negatively, they can have multiple interpretations depending on the context.

Finally, once the speaker s utters an insult m directed at x, the listener's prior beliefs are updated by conditioning $\Pr(\alpha)$ on the m's indexical field, $\Pr(m|\alpha)$. In other terms, the interpretation proceeds by i) combining the likelihood of m's signalling an affective state α with the listener's prior beliefs about the speaker's affective relation with x, and then ii) readjusting the resulting

measure with a normalizing constant, i.e., the sum of these terms computed for all affective states α:

$$\Pr(\alpha|m) = \frac{\Pr(\alpha) \times \Pr(m|\alpha)}{\sum_{\alpha} \Pr(\alpha) \times \Pr(m|\alpha)} \tag{21}$$

The model set out in this section lead us to state the main prediction that our model makes about the affective information expressed by a PI. That is, that the affective information perceived by an audience member in relation to the use of a PI is constrained by the perceived affective relation -according to that audience member- between the speaker and the target of the PI and by how the PI is typically interpreted in the linguistic community where the PI is used (in contrast to its non-insulting counterpart). In the next section we will see some examples of how this works.

Explaining PIs' affective variability

In this section, I model four types of situations using the proposal sketched above: first, a situation in which speaker and target don't know each other; second, a situation which is similar to the first except that the speaker expresses his heightened emotions using additional non-verbal cues; third, a situation where speaker and target are close friends; and, fourth, a situation where speaker and target don't know each other but where the speaker intends to convey a positive message.

First, in a situation where John hears that Alex, an individual he hasn't met before, address him by uttering (22), the PI *bastard* receives its normal (negative and heightened) interpretation:

(22) Hey, bastard, what are you doing?

How can we explain this? In this case, we assume that John doesn't have any prior expectation about Alex's affective relation with him due to their lack of familiarity. Thus, we plug the uniform distribution in (20) and the probabilistic indexical field associated with *bastard* in the formula in (21). As a result, we obtain that John's posterior beliefs about Alex's feelings qualify as higher the probability that Alex is expressing contempt towards him rather than *cordiality* or *friendliness* (cf. the fourth row):

(23) John's beliefs after hearing (22) (situation 1):

AFF	CORDIALITY	FRIENDLINESS	INDIFFERENCE	CONTEMPT
$\Pr(\alpha)$	0.25	0.25	0.25	0.25
$\Pr(bastard\|\alpha)$	0.4	0.5	0.6	0.7
$\Pr(\alpha) \cdot \Pr(bastard\|\alpha)$	0.1	0.125	0.150	0.175
$\Pr(\alpha\|bastard)$	0.181	0.227	0.272	0.318

Second, in a situation that is similar the first one but where John perceives that Alex is emitting non-verbal signals indicating that he is in a heightened emotional state (e.g., a louder voice, faster speech rate, high pitch, etc.) (Bänziger and Scherer, 2005), then we assume that he expects Alex' to be, in general, more energetic or aroused at the utterance's context. In this case, John's appraisal of Alex's global emotional state -given Alex's publicly visible actions before he talks- have an impact on John's prior beliefs about Alex's affective stance towards him.[11] Thus, we plug a distribution which favours [A+] states (CONTEMPT and FRIENDLINESS), and the probabilistic indexical field associated with *bastard* in the formula in (21). This time, we obtain that John's posterior beliefs about Alex's feelings still favour CONTEMPT over the other affective states (similar to what occurs in situation 1). However, we also obtain that the probability that Alex is expressing CONTEMPT is higher

[11]It is worth noting that moods often crystallize in emotions. If we see someone as grumpy or agitated, its more likely that we will see that person as angry towards things surrounding him, including other individuals.

than before. This explains what we observed in Section 2.2, namely, that PIs such as *bastard* can have a more or less 'strong' impact on the audience across different uses depending on how they are used:

(24) John's beliefs after hearing (22) (situation 2):

AFF	CORDIALITY	FRIENDLINESS	INDIFFERENCE	CONTEMPT
$\Pr(\alpha)$	0.20	0.30	0.20	0.30
$\Pr(bastard\|\alpha)$	0.4	0.5	0.6	0.7
$\Pr(\alpha) \cdot \Pr(bastard\|\alpha)$	0.1	0.125	0.150	0.175
$\Pr(\alpha\|bastard)$	0.142	0.267	0.214	0.375

Third, in a situation where John and Alex are close friends, or at least where John presumes that Alex is a close friend, we assume that John expects Alex to express positive [A+] rather than negative [A-] affective states towards him. Characterizing the factors that make John perceive Alex as close or intimate is difficult, however. For example, approaches that co-relate close relationships with positive affect have been challenged by observations that close relationships include both positive and negative affects (e.g., loyalty and rivalry) (Furman, 1989). Instead, we will assume that John's prior beliefs are determined by whether he and Alex have developed a 'script' or 'insider knowledge' about their mock aggressive behavior (Ballard et al., 2003). In this case, even in the absence of non-verbal positive cues (e.g., laughing, smiling, etc.), John will presume that Alex's affective relation with him is nonetheless positive. Thus, we plug a distribution that favours [P+] states (FRIENDLINESS and CORDIALITY), and the probabilistic indexical field associated with *bastard* in the formula in (21). As a result, we obtain that, this time, John will interpret Alex as more probably displaying FRIENDLINESS rather than CONTEMPT:

(25) John's beliefs after hearing (22) (situation 3):

AFF	CORDIALITY	FRIENDLINESS	INDIFFERENCE	CONTEMPT
$\Pr(\alpha)$	0.40	0.40	0.10	0.10
$\Pr(bastard\|\alpha)$	0.4	0.5	0.6	0.7
$\Pr(\alpha) \cdot \Pr(bastard\|\alpha)$	0.16	0.20	0.06	0.07
$\Pr(\alpha\|bastard)$	0.326	0.408	0.122	0.142

However, notice that, unless John is certain that Alex is well intentioned, the risk of Alex's being misinterpreted as expressing CONTEMPT cannot be neglected.[12] In that sense, we can analyze positive uses of a PI as a 'test' of the speaker's relation with the target: if the target takes offense, or takes the utterance as warranting offense, then the speaker's presumption of interpersonal closeness is proven erroneous. Otherwise, i.e., if the listener interprets the PI as endearing (or at least as warranting endearment), the speaker's presumption of closeness is confirmed and thus the PI reinforces their relation. In other terms, the speaker's positive use of the PI testifies of his 'real friendliness' with the target.

Fourth, what happens in contexts where speaker and listener don't know each other, but where the speaker intends to use a PI positively? For example, in order to advertise a product, a company may use the following utterance:

(26) Here's To You, Ya Bastard! You've been such a good friend to me through the years.

In such cases, we may consider that the company's presumption of closeness required to interpret *bastard* positively needs to be accommodated in the context. Accommodation, broadly speaking, is the process by which the context is adjusted in order to make the utterance of a sentence (that im-

[12]Notice that positive uses of PIs more often occur with thin rather than thick PIs, that is, with terms such as *bastard* rather than terms such as *wimp*. This can be explained by observing that a positive use of a PI still describes the target as instantiating a negative property (e.g., being disagreeable), so a more precise or thicker description may increase the risk of the addressee feeling insulted.

poses certain requirements on the context in which it is uttered) acceptable or 'correct play' in such context (Lewis, 1979). Thus, if the listener is willing to accommodate the presumption of closeness without fuss, the company will implicitly make it the case in the common ground that they are friendly, which seems a more risky, but also effective way of establish some sort of affiliation (which is then explicitly reinforced by the follow-up sentence 'You have been such a good friend...'). In this case, the calculation of the probabilities would be similar to that in (23) or (24), in case the listener doesn't accommodate the presumption of closeness, or as (25), in the case the listener does accommodate it.

Before ending this section, it is worth noting that our framework shows some of the similarities and differences between PIs and other emotional expressions, such as expletive adjectives (e.g., *damn*): both types of expressions are associated with an indexical field derived from the affective dimensions valence and arousal, and both typically express [P-, A+] states. However, the main difference is that PIs' function is not merely expressive, but also descriptive of behavioral or idiosyncratic traits that are considered impairing for social relations (e.g., being selfish, being fearful, etc.).

5.5 Conclusion

The affective meaning associated with PIs is multidimensional, that is, linked to different affective qualities derived from the dimensions pleasure (also called 'valence') and arousal. In this chapter, I've defended an indexical view of PIs, where different affective qualities can emerge during the process of interpretation depending, on the one hand, on how a PI is typically interpreted within a linguistic community (when contrasted with a non-insulting alternative) and, on the other hand, on the listener's prior beliefs about the speaker's affective states and/or relation with the target of the PI. This proposal has been used to explain different kinds of variation in a PIs affective

interpretation. In future work, empirical observations about speaker's choice of a PI depending on what he assumes that others assume about him should be integrated to make this model more precise. This would allow us model not only the interpretation, but also the use of PIs in the case of real agents using tools from game-theory.

Chapter 6

Slurring terms

6.1 Introduction

Slurs are expressions such as *Boche* or *Spic* that derogate individuals on the basis of their membership to a social group. However, the class of slurs don't classify just any expression that derogates individuals on the basis of their group membership. As Nunberg (2018) points out, labelling something as a slur assigns to it 'moral or political tenor to the offense it gives and the offense one commits in uttering it', attested by the fact that, in using those terms, an individual performs a speech act in which institutions and the law may take an official interest' (p. 239).

One of the reasons why the harm caused by slurs is different from the harm caused by other expressions (e.g., particularistic insults such as *bastard*) is that, as Jeshion (2020) points out, the 'distinctive scornful denigration inherent to slurs (...) manifest that the target is *lesser*' (p. 11). In other terms, by using a slur, speakers express that the target is unworthy of respect, undeserving the same treatment as the speaker, which thereby presents himself as superior.

However, slurs manifest the same kind of variability that we have observed in other curse words in previous chapters. For example, even though both *fairy* and *faggot* are used to describe an individual as effeminate or homosexual, the latter slur expresses contempt with a greater degree of strength. Moreover, even though *faggot* is typically used to express contempt, 'reclaimed' uses of this expression (or of its shortened version *fag*) among members of the group derogated) are typically described as expressing pride or solidarity.

Thus, slurs pose two problems: i) how can we distinguish the negative attitudes expressed by, e.g., insults like *bastard*, from the harmful attitude expressed by slurs? and ii) how can we explain the affective variation observed in slur's use, which allow them dehumanize individuals in some contexts but to establish affective and affiliating relations in others?

In order to explain slur's distinctive scornful denigration and its high sensitivity to the utterance context (to be analyzed shortly) in a compact way, I propose that slurs' affective meaning is indexical (i.e., multilayered and associative) rather than conventional. That is, I propose that slurs such as *Boche* are indexically associated with a range of different affective qualities. However, in contrast to other types of insults, which are associated with qualities derived from the pleasure and arousal dimensions, I will contend that slur's are associated with affective qualities derived from the pleasure and dominance dimensions. In few words, I will argue that weapon uses of slurs express affective states that score low on pleasure (i.e., that express a negative evaluation of the target group), and that, in addition, score high on dominance (i.e., that signal that the speaker feels dominant with respect to the target group).

In Section 6.2, I i) distinguish between the affective states expressed by slurs and those elicited in others by their use ii) argue that slur's offensiveness doesn't derive from the fact that they target socially relevant groups (i.e., nationality, gender, religion) but from the fact that they portray the tar-

get as being lesser and iii) analyse the different types of affective variation that can be found in their interpretation depending on the context and the speaker. Then, in Section 6.3, I discuss three previous proposals on slurs: a semantic truth-conditional view, a semantic expressivist view, and a pragmatic prohibitionist view. In Section 6.4, based on the observations made in previous sections, I analyze slur's affective meaning as indexical. That is, as associated with different affective qualities that might emerge during its interpretation depending on previous assumptions about the speaker's affective stance. Section 6.5 concludes.

6.2 The empirical landscape

6.2.1 Affectiveness vs. offensiveness

In studying different types of emotive expressives (e.g., curse words like *damn*), researchers mainly focus on what is *expressed* by them at a certain utterance context (e.g., the speaker's exasperation). However, in the case of slurs, research not only take into account what affective states slurs may express (e.g., contempt), but also what affective states slurs may *elicit* in the audience (e.g., offensiveness), including those who are not the direct target(s) of the slur. This distinction raises the following question: should a theory of what slurs express (i.e., their 'content') also explain what slurs elicit in others (i.e., their perlocutionary effects)?

Let have a closer look at the express/elicit distinction in relation to slurs. On the one hand, slurs typically *express* negatively valenced states. As Jeshion (2013) observes, slurs express contempt towards the members of a social group G on account of their belonging to G. Importantly, she adds, such contempt incorporates a negative evaluation: the use of a slur expresses that the speaker ranks the members of G as low in worth, i.e., as being 'beneath the rest' (Jeshion, 2016). For example, by uttering *Kike*, the speaker manifest

that he sees Jews as ranking low in worth *qua* persons on the basis of being Jew.

On the other hand, slurs typically *elicit* negatively valenced states. As Bolinger (2015) points out, slurs are typically offensive. More precisely, the utterance of a slur *warrants* offense: even though a hearer (including the direct target of the slur) may not be actually feel offended by the use of a slur, he would be morally justified in taking offense (2015, p. 3). That is, the slur's utterance would give the hearer a valid reason to feel angry at the speaker. For example, an utterance of *Kike* warrants offense of Jewish people and those who find it detrimental to society to discriminate Jewish people; yet, the utterance may fail to actually generate offense in the hearers (because there are no hearers at the utterance's context, because the hearer shares the negatively evaluation expressed or doesn't take it seriously, etc.).

Now, it may be assumed that slurs elicit negatively valenced states (i.e., offensiveness) in part *because* they express negatively valenced states (i.e., contempt). However, the relation between what slurs express and what they elicit is more complex. As Nunberg (2018) observes, slurs are often used by a speaker 'to provide pleasure and gratification to their friends' (p. 25) rather than to express contempt towards the members of the social group derogated, which may be absent from the conversation. For example, an individual may use the word *Kike* to express his amusement with respect to Jews, even though he hasn't meet a member of the Jewish community in his life. In this cases, humour conceals contempt: slurs can express the speaker's amusement precisely because -for those speakers- using slurs cause harm to their targets. In this case, slurs elicit negatively valenced states (i.e., offense) while expressing positively valenced states (i.e., the speaker's amusement in derogating others).

In the case of 'reclaimed' uses of slurs, that is, cases where members of the group derogated by a slur (and perhaps others with 'insider' status) use the

slur to express a non-derogatory attitude, the same phenomenon occurs. To wit, reclaimed uses of slurs are often described as cases where the speaker expresses positively valenced affective states (e.g., pride, solidarity, friendliness, etc.), and thus as cases where the slur is used non-offensively. However, reclaimed slurs can also be used to express affective states which would be difficult to categorize as positively valenced (e.g., concern, disgust, or even animosity). For example, during a confrontational situation, a member of the Black community in the U.S can use the n-word in reference to another member of the same community to express annoyance, without thereby being offensive towards the entire Black community. Thus, reclaimed slur's may fail to elicit offense despite expressing the speaker's negatively valenced states.

Therefore, even though slurs are often offensive because they express negatively valenced states, or are non-offensive (among members of the group derogated) because they may express positively valenced states in certain situations, slur's affective effects on the audience are nearly orthogonal to the pleasure dimension, that is, to the type of evaluation they are intended to express. Does this mean that we should explain slur's offensiveness without taking into account the affective states slurs express? In section 6.4, I propose that slur's offensiveness doesn't derive from their relation to a particular (negative) evaluation, but from their indexical association to a different affective dimension, namely, 'dominance'.

Before moving to the next section, it is important to note that slur's offensiveness is not only understood in psychological terms (e.g., the feeling of anger that the use of a slur may warrant), but also in moral terms. To wit, researchers take slurs not only to be offensive, but also to enable and reinforce a system of oppression that places some groups below others in a social hierarchy and thus dehumanize their members. Indeed, it is this moral component that distinguish slurs from other types of derogatory expressions

(Nunberg, 2018). To wit, slurs are distinguished from those expressions that derogate dominant groups (e.g., the rich or the powerful) because these latter express derogation but fail to dehumanize their targets (Popa-Wyatt and Wyatt, 2017). To wit, calling a white person *Cracker* express contempt but cannot place white people low in a social hierarchy.

Thus, a theory of slurs should be able to shed light not just on why their linguistic properties make them psychologically offensive, but also morally offensive. Now, it may be assumed that slurs are morally offensive because they target members of a group on the basis of relevant social categories such as ethnicity, gender or religion. However, in the next section I will argue that slur's moral offensiveness does not derive from the fact that they derogate relevant social groups, but from the fact that they constitute a particular form a violence that express that the target are *lesser*.

6.2.2 Slurring groups vs. slurring individuals

The class of slurs typically includes those expressions that refer and derogate social groups (e.g., *Kike*, which refers to Jewish people) or that appeal to ideologies that oppress a given social group (e.g., *bitch*, which don't refer to women but reinforce ideologies that police women behaviors). In other terms, researchers have standardly focused on how slurs' offensiveness enables and/or reinforces social oppression, that is, a structural phenomenon that positions certain groups as disadvantaged in relation to others with respect to social categories such as religion or gender (Frye, 1983). However, does the offensiveness of slurs derives from the fact that they appeal to social categories? As we will see in this section, slurs and pejorative nicknames display a similar offensive profile: both types of expressions convey that their target is *lesser* and, in consequence, behave similarly along various dimensions. This parallelism will show that the source of slurs' offensiveness doesn't derive from the fact that they target social groups.

What are pejorative nicknames? Pejorative nicknames are expressions that modify or replace the 'standard' name of an individual. In ethnographic studies of nicknaming practices within small communities, it has been observed that (pejorative) nicknames come in different sub-types, such as the following:

- Descriptive: these nicknames appeal to a distinctive trait of the bearer, such as his appearance (e.g., *Dumbo*), personality (*Crazy Tom*) or behavior (*Eats-a-lot*). Various ethnographic reports indicate that nickname's offensiveness is often orthogonal to their descriptive component: an explicitly disparaging nickname (e.g., *Dumbo*) can be used to express affect, and, conversely, an apparently neutral nickname (e.g., 'The Chinese') can be considered deeply offensive by the bearer (Dorian, 1970).

- Non-sensical: in this case, there is little or no agreement among users about the meaning or etymology of a nickname (e.g., *Matruco*, reported in Gilmore (1982)). Yet, despite the absence of descriptive associations, non-sensical nicknames may be as offensive as descriptive nicknames. Gilmore (1982) suggest that, in these cases, their pejorative flavor derives from their phonetic rather than semantic features. That is, the phonetics of the expression might suggest, e.g., stupidity, which will be then associated with the bearer of the nickname by metonymy.

- Gendered: these nicknames are applied on the basis of social norms that police the behaviors of individuals. de Klerk and Bosch (1996) observes that gendered nicknames include sex reversing nicknames (e.g., *Johnny Lassie*) and objectifying nicknames (e.g., *Sexy ankles*), among others sub-types. Like gendered slurs such as *slut*, gendered nicknames also reinforce social attitudes by perpetuating false expectations based on gender-role stereotypes.

To what extent are nicknames' and slurs' offensiveness similar? First, slurs

display 'derogatory autonomy', that is, they are offensive independently of whether the speaker feels any contempt or ill-will towards the target group (Hom, 2008). Similarly, ethnographers offer rich descriptions about speaker's 'lack of innocence' in using pejorative nicknames without the intention to derogate. In her study of nicknaming practices among Gaelic communities, Dorian (1970) observes that speaker's noticeable ignorance of the offensiveness of a nickname doesn't block its harmful effects. Moreover, during their investigations, ethnographers themselves are often recommended 'never to use the names, either in reference or address, lest I provoke insult, mortification and ill will' (Brandes, 1975, p. 141) or are 'warned against using the nicknames openly because most people take offense' (Gilmore, 1982, p. 693). This indicates that, similarly to slurs, pejorative nickname's offensiveness is autonomous from the speaker's non-derogatory intentions.

Second, slurs display an 'hyper-projective' character. That is, slur's offensiveness scopes out from truth-conditional operators (e.g., negations, conditionals, etc.) and even from quotation marks: for example, saying 'John is *not* a Kike' offends the Jewish community at least as much as saying 'John is a Kike'. Does this property applies to pejorative nicknames? Ethnographic studies of nicknames have focused more on understanding their role in the communities that devise and use them, rather than on their scopal properties. However, from their reports it is possible to extract information about nickname's interaction with entailment-cancelling operators. To wit, researchers notice that the mere pronouncement of a nickname can elicit emotional responses (Gilmore, 1982). In other terms, nickname's offensiveness is associated with the sheer presence of their tokens, rather than to their semantic properties. Using Gilmore (1982)'s case, uttering 'Who is 'Matruco'?' or 'Is 'Matruco' offensive?' is likely to offend the bearer of the nickname despite the presence of quotation marks. Therefore, the offensiveness of pejorative nicknames also displays a hyper-projective character.

Finally, it has been observed that slurs can be 'reclaimed'. As mentioned in Section 6.2.1, slurs may eventually come to be neutralized through social processes of appropriation, and may even reverse their standard negative valence to positive in order to promote pride or solidarity among members of the target community. For example, after a process of reclamation, a term such as *queer* has now non-pejorative uses that can be attested in expressions such as *queer festival* or *queer cinema* (Cepollaro et al., 2021b). Similarly, ethnographers have observed that pejorative nicknames sometimes end up being embraced by their bearers. de Klerk and Bosch (1996) observe that the appropriation of nicknames often aims at 'underlining popularity' within the target's community. In the same vein, Gilmore (1982) hypothesizes that the uniqueness of a pejorative nickname can be an important factor behind its acceptance, such that 'the motivation to be distinctive may be at times stronger than the one 'to put one's best foot forward" (Seeman (1976) quoted in Gilmore (1982)). Thus, both slurs and pejorative nicknames can be re-appropriated in order to enhance pride and build a stronger identity.

As we can observe, slurs, i.e., expressions that derogate groups in virtue of relevant social categories, and pejorative nicknames, i.e., expressions that derogate individuals *qua* individuals, behave similarly within the linguistic communities in which they are active. We can hypothetize that the reason why slurs and pejorative nicknames behavior is similar is that both express the speaker's dominance over the target. On the one hand, slurs are used to express that certain groups of individuals are lesser *qua* members of such group, and thus enable and reinforce social hierarchies among groups. On the other, certain pejorative nicknames (or 'slurs for individuals') are used to express that their individual targets are lesser *qua* individuals, and thus enable local hierarchies among the members of the target's community (e.g., family, school, work, etc.), often related to physical violence. Therefore, slur's moral offensiveness' is located in the fact that they express that the target has a lesser standing, independently of whether the target is a social group

or just an individual. A theory of slurs should be able explain how their linguistic properties make them morally offensive in virtue of expressing that some individuals have a lower standing as human beings.

In what follows of this chapter, I will continue focusing on slurs for groups rather than slurs for individuals. Yet, it will be assumed that a theory of slur's offensive profile has to be structurally similar or at least compatible with a theory of pejorative nicknames' offensive profile.

6.2.3 User vs. interpreter offense variation

It has been observed that slurs' offensiveness comes in different degrees of strength (e.g., the n-word is considered more offensive than *Chink*) (Jeshion, 2013) and that different uses of the same slur are offensive to different degrees depending on various contextual factors (e.g., the use of *faggot* with a contemptuous intonation is more offensive than its use with a friendly intonation) (Popa-Wyatt and Wyatt, 2017). Moreover, it has also been observed that slur's offensiveness also varies with respect to language-internal factors such as, for example, the grammatical environment in which the slur appears[1]. In particular, Cepollaro et al. (2019) presents empirical evidence that slur's are, on average, perceived as more offensive when used in atomic sentences (e.g., 'John is a Spic') than when uttered in speech-reports (e.g., 'Alex says John is a Spic'), thus showing that quotational environments (e.g., speech reports) mitigate (but don't completely block) the offensive effects of a slur.

However, slurs' offensiveness not only varies with respect to how they are used or who uses them, but also with respect to who interprets them -more precisely, with respect to the particular ideological stance of the interpreter.

[1]The distinction between internal (e.g., grammatical) vs. external (e.g., social) factors has been mainly studied in the variationist sociolinguistics framework. See, e.g., Labov (1966).

Arguably, some slur's coiners and users often don't see slurs as harmful, but as merely referential devices. For example, Nunberg (2018) reports that users of the term *Redskin* maintain the the term is laudatory of the 'toughness, bravery and perseverance' of Indian people (p. 28), even though members of the group referred by such term explicitly qualify it as offensive. In the same vein, slur's perceived degree of offensiveness may also vary with respect to how much the interpreter has normalized (e.g., is blind to) social hierarchies. For example, individuals that believe in 'inverse racism', i.e., that all groups individuated by their race can be object of discrimination or oppression, will probably consider that slurs that target, e.g., white people (e.g., *cracker*) and black people (e.g., *spade*), don't differ much with respect to their degree of offensiveness. Even though these hypotheses need to be empirically tested in future work, it seems the case that the degree of offensiveness attributed to slurs is mediated by the ideological orientation (e.g., social values) of the interpreter.

6.3 Previous accounts

There are numerous accounts of slurs. On the one hand, content-based accounts consider that slurs encode derogation, either at the truth-conditional (Hom, 2008; Hom and May, 2014) or non-truth-conditional dimensions (**?**Cepollaro and Stojanovic, 2016). On the other, non-content appeal to taboos (Anderson, 2014) or conversational implicatures (Bolinger, 2015; **?**). In this section, I briefly analyze three representative theories, and focus on how they account for the offensive profile of slurs and, in particular, for the wide array of emotions they can display.

First, Hom (2008) argues that slurs are offensive because they ascribe negative properties to their targets. In this framework, a slur like *Chink* is analyzed as a socially constructed property such as 'ought to be subject to higher college admissions standards, and ought to be subject to exclusion from ad-

vancement to managerial positions, and ..., because of being slanty-eyed, and devious, and good-at-laundering, and ..., all because of being Chinese'. Since it is false that an individual ought to be subject to any kind of discrimination on account of their race, ethnicity, etc., the truth-conditional account predicts that *Chink* has a null extension.

How does the truth-conditional account explains offense variation? Hom (2008) claims that the degree of offensiveness of a slur varies depending on the negative stereotypes attributed to the group it is functionally associated with. In this framework, *Kike* and *Guido* offend to different degrees because the stereotypes associated with the former group are more negative than those associated with the latter. However, a problem with this solution is that it is difficult to extend to other kinds of variation. For example, *Kike* is considered more offensive than *Yid*, even though both expressions refer to the same community, namely, to Jewish people. Yet, to explain this difference in offensiveness, we would need to assume the co-existence of two different ideologies against Jews within the same community, which seems implausible (Anderson, 2014; Popa-Wyatt and Wyatt, 2017).[2]

Second, Jeshion (2013) argues that slurs are offensive because they conventionally express the speaker's contempt towards the group referred by the slur. More precisely, slurs express that the members of the group referred by the slur are *lesser*. Thus, while the expressive account analyses slurs as *expressing* that the target of the slur is 'beneath the rest, possessing lower status along the moral dimension, broadly construed' (Jeshion, 2016), the

[2]An additional problem with an analysis of slurs as complex properties is that the class of slurs is not grammatically uniform. To wit, slurs not only include nouns (e.g., *Chink*) or adjectives (e.g., *Bitch*) but also denominalized verbs (e.g., *to jew*, *to bitch*, *to gyp*) (Sennet and Copp, 2019). For example, *to jew* is an expression that derogates Jewish people and which roughy means 'to cheat someone'. Which complex property could be assigned to *to jew* as its truth-conditional content? Clearly, *to jew* is not used to describe Jews as being subject to a certain discriminatory treatment because of having certain properties all because of being Jew, but to describe a particular type of action.

truth-conditional account analyses slurs as *describing* the target as being beneath the rest. Moreover, Jeshion (2013) maintains that this expression of contempt is uniform across all slurs, that is, that a common core attitude is lexically encoded by all slurs equally.

How does the expressivist view accounts for offense variation? Jeshion (2013) proposes that offense variation is a pragmatic phenomenon. In particular, various pragmatic effects account for derogatory variation. For example, the pragmatic activation of stereotypes, the offense caused by breaking prohibitions of varying strength, etc. Yet, a problem with this pragmatic explanation is that, if slurs lexically encode contempt, it is unclear which pragmatic mechanisms make it the case that uses of a slur by members of the group derogated no longer express such contempt. The expressivist approach might argue that appropriated slurs are ambiguous between derogatory and non-derogatory content. Yet, it would remain unclear why appropriated slurs impose restrictions on who can use them to express a positive attitude: not just any speaker can use the n-word or *bitch* to mean something friendly or positive Ritchie (2017).

Another problem with an explanation of appropriated uses in terms of ambiguity is that, as we saw in Section 6.2.1., derogatory uses of slurs not only express negatively valenced states (e.g., the speaker's hostility towards the group referred) but also positively valenced states (e.g., the speaker's amusement or joy in dehumanizing the group referred, the speaker's condescending or patronizing attitudes towards the target group, etc.). Therefore, the fact that derogatory uses of slurs typically *elicit* negatively valenced states in their targets (e.g., offensiveness) doesn't imply that slurs necessarily *express* negatively valenced states in those situations. Conversely, non-derogatory uses of a slur, i.e., uses where a member of the group targeted by the slur use it non-offensively, can also express a wide array of emotions, including negatively valenced states (e.g., anxiety, annoyance, etc.). Therefore, even though

reclaimed uses of a slur among members of the community derogated may *elicit* positively valenced states (e.g., affiliation, pride), they don't necessarily do so because they semantically *express* positively valenced states.

Finally, Anderson (2014) claims that the literal content of slurs is identical to the content of their neutral counterparts so, under this approach, slurs' offensiveness is not a result of any semantic mechanism. Instead, slurs are offensive because they are taboo: there are social norms prohibiting the use of slurring terms, so their use is offensive because it constitutes a violation of such norms. This would explain why, as observed above, slurs are offensive not only when they are used but also when they are merely pronounced.

How does the prohibitionist view account for offense variation? Anderson (2014) claim that taboos are flexible. For example, in non-offensive uses, the taboos associated with slurs might be suspended or alleviated according to different contextual factors: whether the speaker is member of the group derogated, whether the slur is used in an academic setting, etc. Yet, a problem with this solution is that it lacks explanatory power Popa-Wyatt and Wyatt (2017). To wit, the view can establish different escape clauses for the different ways in which a term's offensiveness varies depending on the context. Moreover, the argument doesn't explain how those escape clauses always interact with each other in the determination of an slur's offensiveness, as each clause can only explain one type of variation.

6.4 The proposal

6.4.1 The semantics

I contend that, at the semantic level, slurs are equivalent to their neutral counterparts. In other words, slurs are not semantically different from the expressions they modify (e.g., *Jap*, which derives from *Japanese*) or replace (e.g., *Kike*, which replaces *Jew*). In that sense, slurs are not conventionally

associated with negatively valenced (i.e., derogatory) attitudes neither at the truth-conditional nor the use-conditional level.

Yet, conventionality represents only one possible type of association between affective content and slurs. In the next section, I will argue that, at the pragmatic level, slurs are indexically associated with affective qualities derived from two affective dimensions, namely, pleasure (i.e., valence) and dominance. However, it is important to bear in mind that indexicality and conventionality should be understood as 'two phases of the same process, as opposed to a categorical difference between qualitatively separate kinds of content (Agha, 2003; Beltrama, 2020). As we will see in Section 6.4.3, this explains why some particularistic insults (e.g., *fatso*, *retard*) seem to have acquired a slurring function. For the purposes of the explanation, however, I will focus on those expressions which have acquired a certain degree of stability, thus becoming widely considered as slurring by a linguistic (or sub-linguistic) community.

6.4.2 The pragmatics

Slurs' indexical fields: introducing dominance

Slurs typically express the speaker's derogatory attitudes towards a target group. By uttering (1), the speaker is more likely to be interpreted as expressing a negative evaluation of South-American people:

(1) Juan es un Sudaca.
'Juan is a South-American.'

However, unlike other swear words (e.g., expletive adjectives like *damn* and particularistic insults like *bastard*), slurs are not typically associated with speaker's heightened emotions, that is, to a high degree of arousal. To wit, (1) doesn't come as odd or infelicitous in a situation where the speaker doesn't

feel excited or energetic about South-Americans at the utterance context. That is, slurs are part of the vocabulary of the racist or homophobic, not only when he is in a heightened state but in general.

Thus, it may be assumed that slurs are only associated with a particular type of evaluation. However, it is worth noting that (1) not only expresses that the speaker evaluates negatively South-Americans, but also that he feels superior with respect to them. That is, that South-Americans rank as low in worth with respect to the speaker and the social group to which he may belong. Indeed, (1) would come as odd in a situation where he speaker doesn't feel that South-Americans are *lesser* with respect to other social groups.

Now, it has been assumed that expressing that individuals are lesser with respect to others it itself a form of (negative) evaluation (Jeshion, 2016). However, even though these two aspects of slurring utterances often co-occur, they are nearly orthogonal: one can evaluate an individual negatively without feeling that he is lesser than others (e.g., when one qualifies someone as boring or lazy), and one could feel that someone is less than others without necessarily evaluating him negatively (e.g., racist ideologies about Chinese people are often built on positive evaluations, such as that they are better in math). Being evaluated as good in something doesn't preclude that one may be evaluated as inferior overall.

Therefore, I contend that slurs are, at the pragmatic level, associated with the pleasure (i.e., valence) and dominance dimensions. In Mehrabian and Russell (1974) multidimensional theory of emotions, these dimensions are defined as follows:

- Pleasure: this dimension serves to measure the pleasure experienced by the subject during an emotional episode. Thus, it corresponds to a scale including negative ([P-]), neutral ([P±]) and positive ([P+]) affective states. It is the evaluative component.

- DOMINANCE: this dimension serves to measure how 'in control' the subject feels in relation to a stimulus during an emotional episode. Thus, it corresponds to a scale including the sensation of being controlled ([D-]) to the sensation of being in control ([D+]). It is a relational component.

These dimensions distinguish slurs from other swear words (e.g., particularistic insults such as *bastard*) which are instead associated with the arousal rather than the dominance dimension. Now, how does pleasure and dominance are combined in order to determine specific types of affective states? Even though both dimensions haven't been studied independently from the arousal dimension, we can have a better grasp of how they interact in the following graphic (Tarasenko, 2010; Mehrabian and Russell, 1974). As we can observe, all possible combinations of low and high values in each dimension (i.e., pleasure, arousal, dominance) determine 8 'basic' affective states. For example, hostility corresponds to the [P-,A+,D+] state and anxiety to the [P-,A+,D-] state:

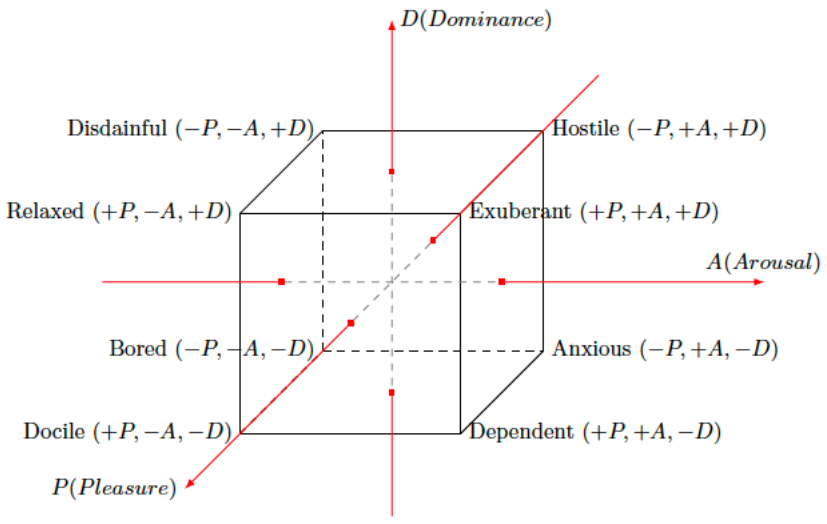


Figure 6.1: Mehrabian's 8 basic affective states

However, as we will ignore the arousal dimension to analyze slurs, we will only take combinations of values in the [P] and [D] dimensions, irrespective of whether qualify as [A+] or [A-]. Thus, the first step to build our model is to use the five qualities mentioned above ([P-], [D+], etc.) to formally define types of affective states. Inspired on Burnett (2017); **?**, we assume a structure $\langle \mathbb{Q}, > \rangle$, where $\mathbb{Q}$ is the set of relevant affective qualities and $>$ encodes relations of compatibility between them (e.g., that an individual cannot be in a [P-] and [P+] state simultaneously, etc.). Note that, this time, we haven't considered the [P±] nor the [D±] quality, as slurs seem to always involve some degree (or lack of) pleasure and dominance, respectively. Furthermore, the combination of [P±] with different degrees of dominance doesn't seem to correspond to any actual affective state:

(2) $\mathbb{Q} = \{$[P+], [P-], [D-], [D+]$\}$

a. [P+] > [P-]

b. [D-] > [D+]

Second, based on this structure, we derive 4 types of affective states α: the [P-, D+] affective state, which we label CONTEMPT, the [P+, D+] state, which we label AMUSEMENT, etc. Notice that we use these labels based on our analysis of slurs in Section 6.2.1: derogatory uses of slurs can express both [P-] states (e.g., contempt towards the group derogated) or [P+] states (i.e., amusement in derogating a social group). That is, what remains constant across weapon uses of slurs is that they express a high degree of dominance, that is, a feeling that the members of the group referred to are lesser than others:[3]

[3]As mentioned in previous chapters, we should keep in mind that these labels assemble different types of affective phenomena. For example, CONTEMPT represents [P-, D+] states in general (e.g., rage, hostility, etc.), and not only contempt.

(3) Possible affective states α:

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
α	[P+, D-]	[P+, D+]	[P-, D-]	[P-, D+]

Third, how can we characterize the association between a slur (e.g., *Sudaca*) and the affective states *alpha* it has the potential to express? As we have observed through this chapter, we cannot assign a single and stable affective meaning to a slur. To wit, the use of *Sudaca* may signal that the speaker feels dominant [D+] and negatively [P-] with respect to South-Americans (in derogatory uses), or positively [P+] and non-dominant [D-] with respect to them (in reclaimed or affiliative uses). For that reason, I will assume that PIs' link to affective states is 'indexical' (Silverstein, 1976; Podesva et al., 2015; Eckert, 2008) rather than conventional. That is, that such relation is grounded on the typical co-occurrence between the use of a slur and a range of affective states, anyone of which could become relevant in a particular context of interaction.

How can we characterize such indexical association? Even though we don't have statistical data about the affective states that typically motivate the use of slurs, it is certain that uses of slurs are more strongly associated with [D+] states (that is, with states like CONTEMPT) rather than [D-] states (that is, with states like AFFILIATION). Therefore, in order to characterize a PIs' indexical meaning, I use the notion of 'probabilistic indexical field' elaborated in Chapter 3. Namely, I associate a slur (e.g., *Sudaca*) with the probability distribution $\Pr(m|\alpha)$, read as 'the likelihood of uttering a slur m given an affective state α'. As can be observed in (4), such distribution captures the fact that slur like *Sudaca* typically express [D+] states across different contexts:

(4) Probabilistic field associated with *Sudaca*:

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
Pr(*sudaca*|α)	0.3	0.6	0.4	0.7

What is the alternative of *Sudaca*? As we observed in Section 6.4.1, *Sudaca* is truth-conditionally equivalent to *Sudamericano*, the term that it modifies, so we can assume that speaker's have the option to choose one or the other when describing an individual (Bolinger, 2015). As we observe in (19), Pr(*sudamericano*|α) = 1 - Pr(*sudaca*|α). That is, we assign a low value to *Sudaca* displaying AFFILIATION, and a high value to *Sudamericano* displaying CORDIALITY (due to its more 'polite' or 'formal' character):

(5) Indexical field associated with *Sudamericano*:

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
Pr(*sudamericano*|α)	0.7	0.4	0.6	0.3

Fourth, I assume that slurs are interpreted based on what it is believed about the speaker's affective disposition of the target of the insult. For example, in (1), *Sudaca* will be interpreted by the listener relative to what he assumes about the speaker's affective relation with South-Americans. Thus, inspired on Burnett (2017, 2019), I represent the listener's prior beliefs as the relativized probability distribution Pr(α), read as 'the probability distribution that the speaker S feels α with respect to social group G'. In a context where the listener doesn't know the speaker, and thus where he has no prior expectations about the speaker's affective relation with the the target social group (e.g., South-Americans), we represent Pr(α) as a uniform distribution over affective states:

(6) Listener's prior beliefs about the speaker's affective relation with the target group G:

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
$\Pr(\alpha)$	0.25	0.25	0.25	0.25

Which factors determine the listener's prior beliefs about the speaker's affective states/dispositions? In the case of slurs, I assume that listener's priors are not only determined by 'short-lived' affective cues (e.g., gestures, facial expressions) but also by more stable assumptions about the speaker's psychological relations with others and his perceived social identity:

A The speaker's psychological closeness with the target: this factors seems to be determined by their past experiences and interactions (e.g., the frequency with which speaker and target engage in affiliating behavior). This factor will be useful to explain why individuals that don't belong to the group derogated by a slur, but which have a certain 'insider' status among them, can use the slur non-offensively (Ritchie, 2017).

B The speaker's social identity: identities are labels that people use to group each other. When the speaker is identified with a label, then such identification is interpreted as giving reasons to the speaker to feel (and act) in certain ways (Appiah, 2010). If the speaker is Catholic, it will be assumed that he tends to feel positively about the Catholic church and to favour its teachings; if the speaker is South-American, it will be typically assumed that he doesn't feel that South-Americans are lesser than other groups or deserve to be discriminated. Even though these assumptions may be prove incorrect, speaker's recognizable social identities guide how listeners think about them.

In a context of utterance, the combination of these factors will influence the listener's expectations about the speaker's affective stance towards a certain social group.

Finally, once the speaker S utters a slur m directed at a social group G, the listener's L prior beliefs are updated by conditioning $\Pr(\alpha)$ on m's indexical

field, $\Pr(m|\alpha)$. In other terms, the interpretation proceeds by i) combining the likelihood of m's signalling an affective state α with the listener's prior beliefs about S's affective relation with G, and then ii) readjusting the resulting measure with a normalizing constant, i.e., the sum of these terms computed for all affective states α:

$$\Pr(\alpha|m) = \frac{\Pr(\alpha) \times \Pr(m|\alpha)}{\sum_{\alpha} \Pr(\alpha) \times \Pr(m|\alpha)} \tag{7}$$

The model set out in this section lead us to state the main prediction that our model makes about the affective information expressed by a slur. That is, that the affective information perceived by an audience member in relation to the use of a slur is constrained by the perceived affective relation -according to that audience member- of the affective relation between the speaker and the social group G targeted by the slur and by how the slur is typically interpreted in the linguistic community where the it is used. In the next section we will see four representative cases of how this model works.

Explaining slurs' affective variability

In this section, I model four types of situations using the proposal sketched above: first, a situation in which speaker uses a slur denoting a social group to which he doesn't belong; second, a situation which is similar to the first except that the speaker has a certain 'insider' status; third, a situation where the speaker uses of a slur denoting a group to which he belongs; and, fourth, a situation which is similar to the third one but where the speaker aims to convey that the target, member of his own social group, is lesser.

First, in a situation where Alex hears that John, an Asian person, utters (8), the slur *Sudaca* will receive its normal (negative and dominant) affective interpretation:

(8) Habrán varios Sudacas en la fiesta.
'There will be a lot of Sudacas in the party.'

How can we explain this? In this case, we assume that Alex doesn't have any prior expectation about Alex's affective relation with South-Americans. Thus, we plug the uniform distribution in (6) and the probabilistic indexical field associated with *Sudaca* in the formula in (7). As a result, we obtain that Alex's posterior beliefs indicate that John is more likely expressing CONTEMPT or AMUSEMENT towards South-Americans rather than AFFILIATION (cf. the fourth row):

(9) Alex's beliefs after hearing (8) (situation 1):

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
$\Pr(\alpha)$	0.25	0.25	0.25	0.25
$\Pr(sudaca\|\alpha)$	0.3	0.6	0.4	0.7
$\Pr(\alpha) \cdot \Pr(sudaca\|\alpha)$	0.075	0.150	0.100	0.175
$\Pr(\alpha\|sudaca)$	0.15	0.3	0.2	0.35

An important prediction of the model for situations like this one is that, independently of whether Alex perceives that John is feeling positively or negatively, the result of the update will still favour [D+] states. In other terms, if John is perceived as feeling positively, he will be interpreted as feeling amusement at the expenses of South-Americans; and if he is perceived as feeling negatively, he will be interpreted as feeling hostility towards South-Americans. In both cases, John expresses that South-Americans are lesser and/or worth of discriminatory attitudes.

Second, in a situation that is similar the first one but where Alex knows that John, despite being Asian, has a certain insider status within the South-American community (e.g., he migrated to South-American at a very young age, he recognizes himself as South-American, etc.), then he will assume

that John doesn't feel superior with respect to South-Americans, or doesn't consider them to be lesser than other social groups (even though he may evaluate them negatively nonetheless). In this case, we plug a distribution which favours [D-] states (e.g., AFFILIATION), and the probabilistic indexical field associated with *Sudaca* in the formula in (7). This, time, we obtain that Alex's come to believe that John is more likely expressing states such as *affiliation* or *anxiety*:

(10) Alex's beliefs after hearing (8) (situation 2):

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
$\Pr(\alpha)$	0.40	0.10	0.40	0.10
$\Pr(sudaca\|\alpha)$	0.3	0.6	0.4	0.7
$\Pr(\alpha) \cdot \Pr(sudaca\|\alpha)$	0.12	0.06	0.16	0.07
$\Pr(\alpha\|sudaca)$	0.292	0.146	0.390	0.170

Third, in a situation where Alex knows, or at least presumes, that John is South-American, we assume that Alex expects John to feel [P+] rather than [P-] states towards South-Americans. Additionally, we assume that Alex expects John not to feel [D+] states towards South-Americans (e.g., to consider them worth of discriminatory attitudes). In this case, we plug a distribution which favours [P+] and [D-] states (e.g., AFFILIATION), and the probabilistic indexical field associated with *Sudaca*, in the formula in (7). As a result, we obtain that, this time, Alex will interpret John as more likely expressing AFFILIATION (e.g., affection, friendship, etc.) rather than CONTEMPT towards South-Americans:

(11) Alex's beliefs after hearing (8) (situation 3):

Situation 2 and 3 illustrate how, despite their typical oppressive function, slurs can be uttered without harm when the speaker belongs or is perceived as belonging to the social group derogated by the slur. Over time, with a

AFF	AFFILIATION	AMUSEMENT	ANXIETY	CONTEMPT
$\Pr(\alpha)$	0.60	0.12	0.12	0.05
$\Pr(sudaca\|\alpha)$	0.3	0.6	0.4	0.7
$\Pr(\alpha) \cdot \Pr(sudaca\|\alpha)$	0.18	0.072	0.048	0.035
$\Pr(\alpha\|sudaca)$	0.537	0.214	0.143	0.104

high frequency of uses of slurs that express non-dominant affective states, the indexical field associated with a slur may be reinterpreted by the whole community and the harmful effects of slurs may come to be less salient (as, for example, in the case of *queer*).

Finally, what happens in situations where the speaker is perceived as belonging to the social group derogated by the slur, but where the speaker seems to use the slur to oppress? For example, in order to intimidate a South-American, John (a South-American who has obtained a non-South-American citizenship) can utter the following sentence:[4]

(12) Hay demasiados Sudacas en mi barrio!
'There are too many South-Americans-PEJ in my neighborhood!'

In such cases, we may consider that John's is trying to accommodate the idea that he is, in fact, not South-American, as such idea seems required in order to interpret his utterance as expressing that South-Americans are lesser. Thus, if Alex accepts the idea that John doesn't sees himself as South-American, he will interpret John's utterance as expressing CONTEMPT rather than AFFILIATION. However, if Alex doesn't accommodate such idea, he will see John as expressing ANXIETY instead (e.g., as expressing unease because of the mismatch between the culture he comes from and the culture he would

[4]For example, in a situation where an African-American boss refers to an African-American subordinate using the n-word, the speaker may be interpreted as expressing a dominant, contemptuous affective state towards that person (Popa-Wyatt and Wyatt, 2017).

like to belong to).

In this section, we have analysed different situations that can be accounted for with our model. In future work, the model should incorporate internal factors such as, e.g., interpreter's bias, that is, how the hearer's ideological perspective influences the interpretation of a slur, and speaker's utilities (see Chapter 3.3.5)

6.4.3 Hyper-projection

As briefly mentioned in Section 2.2, slurs are not only offensive when they are used in atomic sentences (e.g., 'John is a Chink', etc.) but also when they occur under the syntactic scope of entailment-cancelling operators (e.g., 'John is not a Chink'). Moreover, as Anderson (2014) point out, even placing a slur within quotation marks (e.g., "Sudaca' refers to South-Americans in Spanish') warrants offense. Thus, the offense arises by the mere pronouncement of the slur, i.e., by the presence of their tokens irrespective of the linguistic environment. In order to explain this phenomenon, the authors claim that 'it might be that groups prohibit names not explicitly adopted by them, for calling a group a name that its members have not chosen may be viewed as an attempt to usurp their authority to choose (p. 355). However, as various authors have pointed out, prohibitions are not enough to explain slur's complex offensive profile. To wit, if it were the case that slurs' offensiveness depended on their prohibited status, then all slurs would be equally offensive (Popa-Wyatt and Wyatt, 2017).

Yet, even though prohibitions are not sufficient to understand slur's complex offensive profile, Anderson and Lepore's explanation of *why* slurs are prohibited in the first place can help us understand their hyper-projective character. As we saw above, slurs allow the bigoted express their feeling of superiority (i.e., dominance) with respect to other individuals. But, what explains slur's indexical association with [D+] affective states? To wit, even

though many slurs are explicitly disparaging (e.g., *Jungle Bunny)*, slurs also include hypocoristic (e.g., *Jap*) and non-sensical expressions (e.g., *Kike*), so the source of their offensive character cannot be located on their etymology. Instead, a more promising explanation focus on the fact that slurs come as *impositions* to their bearers. In other terms, the indexical link between slurs and dominance derives from the fact that coining and using slurs violates the target's autonomy to determine how they want to be treated, perceived, etc. Such imposition constitutes a form of symbolic violence, which eventually feeds from (and reinforces) other forms of violence being endured by the target. As a result, slurs end up being expressing not only a negative evaluation of their targets, but also that they aren't worth of respect as individuals.

How does slurs being imposed labels explains their hyper-projective character? 'Being imposed' is not a semantic nor pragmatic property, but a feature of how a particular sign came to be coined and used within a linguistic community (e.g., in the same way that 'having four letters' is a property of the word *tree*, or 'being used in Spain' is a property of the word *Sudaca*). Now, pace Anderson and Lepore, independently of whether a sign was ever implicitly or explicitly prohibited within a linguistic community, its circulation within that community is itself offensive. Therefore, slur's offensiveness not only derives from what they come to express at a given utterance context (i.e., the speaker's dominant attitudes) but from their mere existence in the vocabulary of a linguistic community. Therefore, it is not slurs' 'prohibited', but 'imposed' character, which explains why they can, at the pragmatic level, be used to express dominant affective states and why, at the meta-pragmatic level, their mere existence (and thus pronouncement) is morally offensive for their bearers.

6.5 Conclusion

The complex affective meaning associated with slurs is multidimensional, that is, linked to different affective qualities derived from the pleasure and dominance dimensions. In this chapter, I have defended an indexical view of slurs, where different affective qualities can emerge during the process of interpretation depending on how the slur is typically used and how the listener's affective predispositions are perceived. Moreover, we have distinguished between the affective states expressed by slurs, from those they elicit: even though a slur might be used to express joy at the expenses of the derogated group, it still elicits offense. In our account, this is because slurs express a high degree of dominance.

Conclusion

Emotions are a complex phenomena. In this dissertation I have argued that, in order to study how some expressions (in particular, curse words) have the capacity to express emotions, we should decompose emotions using affective dimensions such as pleasure, arousal and dominance. This allow us incorporate emotions in a formal theory of communicative interaction. I have then proposed that the association between a given expression (e.g., *damn*) and such affective dimensions should be modelled as 'indexical' rather than 'conventional'. I have also argued that this perspective help us explain why curse words can have radically different interpretations depending on who uses them and how he uses them, in everyday communication. Thus, in this dissertation I mostly focused on how affective expressions are *interpreted*. Complementing these ideas with an analyses of how affective expressions are *used* is left for future work.

Bibliography

Abrusán, M. (2011). Predicting the presuppositions of soft triggers. *Linguistics and philosophy*, 34(6):491–535.

Agha, A. (2003). The social life of cultural value. *Language & communication*, 23(3-4):231–273.

Anderson, J. (2014). *Names*. Oxford University Press.

Appiah, K. A. (2010). The ethics of identity. In *The Ethics of Identity*. Princeton University Press.

Ballard, M. E., Green, S., and Granger, C. (2003). Affiliation, flirting, and fun: Mock aggressive behavior in college students. *Psychological Record*, 53(1):33–50.

Bänziger, T. and Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech communication*, 46(3-4):252–267.

Beller, C. (2013). Manufactured and inherent pejorativity. *Semantics and Linguistic Theory*, 23:136.

Beltrama, A. (2020). Social meaning in semantics and pragmatics. *Language and Linguistics Compass*, 14(9):e12398.

Blakemore, D. (2002). *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*, volume 99. Cambridge university press.

Blakemore, D. (2011). On the descriptive ineffability of expressive meaning. *Journal of Pragmatics*, 43(14):3537–3550.

Bolinger, D. (1972). *Degree words*. De Gruyter Mouton.

Bolinger, R. J. (2015). The pragmatics of slurs. *Nous*, 51(3):439–462.

Brandes, S. H. (1975). The structural and demographic implications of nicknames in navanogal, spain1. *American Ethnologist*, 2(1):139–148.

Bross, F. (2021). On the interpretation of expressive adjectives: pragmatics or syntax? *Glossa: a journal of general linguistics*, 6(1).

Burnett, H. (2017). Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics. *Journal of Sociolinguistics*, 21(2):238–271.

Burnett, H. (2019). Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy*, 42(5):419–450.

Burridge, K. and Mulder, J. (1998). English in australia and new zealand: An introduction to its history, structure, and use. *Oxford*.

Campbell-Kibler, K. (2005). *Listener perceptions of sociolinguistic variables: The case of (ING)*. Stanford University.

Campbell-Kibler, K. (2007). Accent,(ing), and the social logic of listener perceptions. *American speech*, 82(1):32–64.

Campbell-Kibler, K. (2008). I'll be the judge of that: Diversity in social perceptions of (ing). *Language in Society*, 37(5):637–659.

Cepollaro, B., Domaneschi, F., and Stojanovic, I. (2021a). When is it ok to call someone a jerk? an experimental investigation of expressives. *Synthese*, 198(10):9273–9292.

Cepollaro, B. et al. (2021b). The moral status of the reclamation of slurs. *Organon F*, 28(3):672–688.

Cepollaro, B. and Stojanovic, I. (2016). Hybrid evaluatives: In defense of a presuppositional account. *Grazer Philosophische Studien*, 93(3):458–488.

Cepollaro, B., Sulpizio, S., and Bianchi, C. (2019). How bad is it to report a slur? an empirical investigation. *Journal of Pragmatics*, 146:32–42.

Cole, P. M. (1986). Children's spontaneous control of facial expression. *Child development*, pages 1309–1321.

Cruse, A. (2006). *Glossary of semantics and pragmatics*. Edinburgh University Press.

Davis, C. and Gutzmann, D. (2015). Use-conditional meaning and the semantics of pragmaticalization. In *Proceedings of Sinn und Bedeutung*, volume 19, pages 197–213.

de Klerk, V. and Bosch, B. (1996). Nicknames as sex-role stereotypes. *Sex Roles*, 35(9-10):525–541.

de Melo, C. M. and Terada, K. (2020). The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma. *Scientific reports*, 10(1):1–8.

Dorian, N. C. (1970). A substitute name system in the scottish highlands. *American Anthropologist*, 72(2):303–319.

Driver, J. L. and Gottman, J. M. (2004). Daily marital interactions and positive affect during marital conflict among newlywed couples. *Family process*, 43(3):301–314.

Eckert, P. (2008). Variation and the indexical field 1. *Journal of sociolinguistics*, 12(4):453–476.

Eder, A.-M. A. (2019). Evidential probabilities and credences. *The British Journal for the Philosophy of Science*.

Eggins, S. and Slade, D. (2004). *Analysing casual conversation*. Equinox Publishing Ltd.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Franke, M. (2009). *Signal to act: Game theory in pragmatics*. Institute for Logic, Language and Computation Amsterdam.

Frazier, L., Dillon, B., and Clifton, C. (2014). A note on interpreting damn expressives: transferring the blame. *Language and Cognition*, 7(2):291–304.

Frye, M. (1983). *The politics of reality: Essays in feminist theory*. Crossing Press.

Furman, W. (1989). The development of children's. *Children's social networks and social supports*, 136:151.

Geurts, B. (2007). Really fucking brilliant. *Theoretical Linguistics*, 33(2).

Gilmore, D. D. (1982). Some notes on community nicknaming in spain. *Man*, 17(4):686.

Glanzberg, M. (2007). Context, content, and relativism. *Philosophical Studies*, 136(1):1–29.

Goodman, N. D. and Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Gutzmann, D. (2015). *Use-Conditional Meaning*. Oxford University Press.

Gutzmann, D. (2019). *The Grammar of Expressivity*. Oxford University Press, Oxford.

Gutzmann, D. and McCready, E. (2016). Quantification with pejoratives. *Pejoration*, (2016):75–102.

Hajek, A. (2002). Interpretations of probability.

Henderson, R. and McCready, E. (2019). Dogwhistles and the at-issue/non-at-issue distinction. In *Secondary content*, pages 222–245. Brill.

Hess, L. (2018). Perspectival expressives. *Journal of Pragmatics*, 129:13–33.

Hess, U. and Hareli, S. (2015). The role of social context for the interpretation of emotional facial expressions. In *Understanding facial expressions in communication*, pages 119–141. Springer.

Hess, U. and Hareli, S. (2017). The social signal value of emotions: The role of contextual factors in social inferences drawn from emotion displays.

Hom, C. (2008). The semantics of racial epithets. *Journal of Philosophy*, 105(8):416–440.

Hom, C. (2012). A puzzle about pejoratives. *Philosophical Studies*, 159(3):383–405.

Hom, C. and May, R. (2014). The inconsistency of the identity thesis. *ProtoSociology*, 31:113–120.

Hyatt, C., Maples-Keller, J. L., Sleep, C., Lynam, D., and Miller, J. (2017). The anatomy of an insult: Popular derogatory terms connote important individual differences in externalizing behavior.

Jackendoff, R. S. (1972). Semantic interpretation in generative grammar.

Jackson, J. H. (1958). *Selected writings of John Hughlings Jackson.* London: Staples.

Janschewitz, K. (2008). Taboo, emotionally valenced, and emotionally neutral word norms. *Behavior Research Methods*, 40(4):1065–1074.

Jay, T. (1992). *Cursing in America*, volume 10. Philadelphia: John Benjamins.

Jay, T. (2000). *Why we curse*, volume 160. Philadelphia: John Benjamins.

Jeshion, R. (2013). Expressivism and the offensiveness of slurs. *Philosophical Perspectives*, 27(1):231–259.

Jeshion, R. (2016). Slur creation, bigotry formation: The power of expressivism. *Phenomenology and Mind*, (11):130–139.

Jeshion, R. (2020). Varieties of Pejoratives. In Khoo, J. and Sterkin, R., editors, *Routledge Handbook of Social and Political Philosophy of Language*.

Kaplan, D. (1998). The meaning of ouch and oops. explorations in the theory of meaning as use.

Kayyal, M., Widen, S., and Russell, J. A. (2015). Context is more powerful than we think: contextual cues override facial cues even for valence. *Emotion*, 15(3):287.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, 30(1):1–45.

Koev, T. (2018). Notions of at-issueness. *Language and Linguistics Compass*, 12(12):e12306.

Kolbel, M. (2002). *Truth Without Objectivity*. Routledge.

Kratzer, A. (1999). Beyond ouch and oops: How descriptive and expressive

meaning interact. In *Cornell conference on theories of context dependency*, volume 26. Cornell University Ithaca, NY.

Labov, W. (1972). *Language in the inner city: Studies in the Black English vernacular*. Number 3. University of Pennsylvania Press.

Labov, W. (2012). *Dialect diversity in America: The politics of language change*. University of Virginia Press.

Lasersohn, P. (2005). Context dependence, disagreement, and predicates of personal taste∗. $Linguistics and Philosophy, 28(6) : 643 - -686$.

Lewis, D. (1979). Scorekeeping in a language game. In *Semantics from different points of view*, pages 172–187. Springer.

Lordat, J. (1843). Analyse de la parole pour servir à la théorie de divers cas d'alalie et de paralalie (de mutisme et d'imperfection du parler) que les nosologistes ont mal connus. *Journal de la Société de médecine pratique de Montpellier*, 7:333–353.

Lyons, W. (1980). *Emotion*. Cambridge University Press.

Martin, F. (2014). Restrictive vs. nonrestrictive modification and evaluative predicates. *Lingua*, 149:34–54.

McCready, E. (2010). Varieties of conventional implicature. *Semantics and Pragmatics*, 3:1–58.

McCready, E. (2012). Emotive equilibria. *Linguistics and Philosophy*, 35(3):243–283.

McCready, E. (2019). *The semantics and pragmatics of honorification: Register and social meaning*, volume 11. Oxford University Press, USA.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for

describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

Mehrabian, A. and Russell, J. A. (1974). *An approach to environmental psychology.* the MIT Press.

Morzycki, M. (2009). Degree modification of gradable nouns: size adjectives and adnominal degree morphemes. *Natural Language Semantics*, 17(2):175–203.

Morzycki, M. (2011). Expressive Modification and the Licensing of Measure Phrases. *Journal of Semantics*, 28(3):401–411.

Nunberg, G. (2018). The social life of slurs. *New work on speech acts*, pages 237–295.

Ong, D. C., Zaki, J., and Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, 143:141–162.

Ong, D. C., Zaki, J., and Goodman, N. D. (2019). Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science*, 11(2):338–357.

Padilla Cruz, M. (2018). Expressive aps and expletive nps revisited: Refining the extant relevance-theoretic procedural account. *Lingua*, 205:54–70.

Pinker, S. (1995). *The language instinct: The new science of language and mind*, volume 7529. Penguin UK.

Podesva, R. J., Reynolds, J., Callier, P., and Baptiste, J. (2015). Constraints on the social meaning of released/t: A production and perception study of us politicians. *Language Variation and Change*, 27(1):59–87.

Pollastri, A. R., Raftery-Helmer, J. N., Cardemil, E. V., and Addis, M. E.

(2018). Social context, emotional expressivity, and social adjustment in adolescent males. *Psychology of Men & Masculinity*, 19(1):69.

Popa-Wyatt, M. and Wyatt, J. L. (2017). Slurs, roles and power. *Philosophical Studies*, 175(11):2879–2906.

Potts, C. (2004). *The Logic of Conventional Implicatures*. Oxford University Press.

Potts, C. (2007a). The centrality of expressive indices.

Potts, C. (2007b). The expressive dimension.

Potts, C. and Kawahara, S. (2004). Japanese honorifics as emotive definite descriptions. In *Semantics and Linguistic Theory*, volume 14, pages 253–270.

Potts, C. and Roeper, T. (2006). The narrowing acquisition path. *The syntax of nonsententials: Multidisciplinary perspectives*, page 183.

Potts, C. and Schwarz, F. (2008). Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora. *Ms., UMass Amherst*, pages 1–29.

Qing, C. and Cohn-Gordon, R. (2019). Use-conditional meaning in rational speech act models. In *Proceedings of sinn und bedeutung*, volume 23, pages 253–266.

Radcliffe-Brown, A. R. (1940). On joking relationships. *Africa*, 13(3):195–210.

Recanati, F. (2019). Force cancellation. *Synthese*, 196(4):1403–1424.

Reisenzein, R. (1994). Pleasure-arousal theory and the intensity of emotions. *Journal of personality and social psychology*, 67(3):525.

Ritchie, K. (2017). Social identity, indexicality, and the appropriation of slurs. 17(50):155–180.

Ronderos, C. and Domaneschi, F. (2022). Predicting the f*** ing word: an eye-tracking study on negative expressive adjectives. *Experiments in Linguistic Meaning*.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Russell, J. A. (1989). Measures of emotion. In *The measurement of emotions*, pages 83–111. Elsevier.

Saka, P. (2007a). *How to think about meaning*, volume 109. Springer Science & Business Media.

Saka, P. (2007b). *How to Think About Meaning*. Springer Netherlands.

Saxe, R. and Houlihan, S. D. (2017). Formalizing emotion concepts within a bayesian model of theory of mind. *Current opinion in Psychology*, 17:15–21.

Schlenker, P. (2005). Minimize restrictors!(notes on definite descriptions, condition c and epithets). In *Proceedings of Sinn und Bedeutung*, volume 9, pages 385–416.

Schlenker, P. (2007). Expressive presuppositions. *Theoretical Linguistics*, 33(2).

Schlenker, P. (2017). Super monsters ii: Role shift, iconicity and quotation in sign language. *Semantics and Pragmatics*, 10(12):1–67.

Schlenker, P. (2018). Iconic pragmatics. *Natural Language & Linguistic Theory*, 36(3):877–936.

Seeman, M. V. (1976). The psychopathology of everyday names. *British Journal of Medical Psychology*, 49(1):89–95.

Sennet, A. and Copp, D. (2019). Pejorative verbs and the prospects for a unified theory of slurs. *Analytic Philosophy*, 61(2):130–151.

Silverstein, M. (1976). Shifters, linguistic categories, and cultural description. *Meaning in anthropology*, pages 11–55.

Silverstein, M. (1979). Language structure and linguistic ideology. *The elements: A parasession on linguistic units and levels*, 193:247.

Simons, M., Tonhauser, J., Beaver, D., and Roberts, C. (2010). What projects and why. In *Semantics and linguistic theory*, volume 20, pages 309–327.

Smith, P. K. and Boulton, M. (1990). Rough-and-tumble play, aggression and dominance: Perception and behaviour in children's encounters. *Human Development*, 33(4-5):271–282.

Stojanovic, I. (2021). Derogatory terms in free indirect discourse. In *The Language of Fiction*. Oxford University Press, Oxford, U.K.

Stojanovic, I. and Kaiser, E. (2022). Exploring valence in judgments of taste.

Sundell, T. (2016). The tasty, the bold, and the beautiful. *Inquiry*, 59(6):793–818.

Talbott, W. (2001). Bayesian epistemology.

Tarasenko, S. (2010). Emotionally colorful reflexive games. *arXiv preprint arXiv:1101.0820*.

Townsend, S. W., Koski, S. E., Byrne, R. W., Slocombe, K. E., Bickel, B., Boeckle, M., Braga Goncalves, I., Burkart, J. M., Flower, T., Gaunet, F., et al. (2017). Exorcising grice's ghost: An empirical approach to studying

intentional communication in animals. *Biological Reviews*, 92(3):1427–1433.

Umbach, C. (2001). (de)accenting definite descriptions.

Vayrynen, P. (2013). *The Lewd, the Rude and the Nasty*. Oxford University Press.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Williams, B. (2006). *Ethics and the Limits of Philosophy*. Routledge.

Williamson, T. (2002). *Knowledge and its Limits*. Oxford University Press on Demand.

Wundt, W. (1896). *Compendio de psicología*. La España Moderna.

Zaki, J. (2013). Cue integration: A common framework for social cognition and physical perception. *Perspectives on Psychological Science*, 8(3):296–312.