Samethinking

Romain Bourdoncle

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Ecole Normale Supérieure

SAMETHINKING

Soutenue par

Romain BOURDONCLE

Le 13 décembre 2022

Ecole doctorale n° ED 540

Lettres, arts, sciences humaines et sociales

Spécialité

Sciences cognitives

Composition du jury :

Ernesto PERINI-SANTOS Prof, U. fédérale du Minas Gerais	*Rapporteur & Président*
Sajed TAYEBI Prof assistant, Institute for Research in Fundamental Sciences (IPM), Teheran	*Rapporteur*
Eleonora ORLANDO Prof, U. de Buenos Aires	*Examinatrice*
Stacie FRIEND Reader, Birkbeck College, University of London	*Examinatrice*
Manuel GARCÍA-CARPINTERO Prof, U. de Barcelone	*Examinateur*
Michael MUREZ Maître de conférences, U. de Nantes	*Co-encadrant*
François RECANATI Prof, Collège de France	*Directeur de thèse*

ENS | PSL

ENS DEC DÉPARTEMENT D'ÉTUDES COGNITIVES

DEC
DÉPARTEMENT
D'ÉTUDES
COGNITIVES

PSL

This work was supported by a doctoral scholarship from Paris Sciences et Lettres ANR-10-IDEX-0001-02 PSL. The writing of the dissertation has been helped by the Département d'Études Cognitives of the École Normale Supérieure de Paris, by the Institut Jean Nicod, and received support from grants ANR-17-EURE-0017 FrontCog, and ANR-19-CE28-0019-01 AMBISENSE.

Samethinking

Romain Bourdoncle

What is it that makes for [interpersonal] coordination? I don't know. The fact that we each have two ways of having a particular *x* in mind doesn't suffice for communication. We need to be in synch. It is very hard to say what this interpersonal coordination amounts to, but it is not that we are on the same path, nor that we are on paths grounded in the same event. (...) In spite of all the work on reference, issues concerning the transmission and coordination of our knowledge of *things* have been understudied. I read Kripke's "A Puzzle about Belief" (1979) as a contribution to this study (though I don't know if he intended it that way).—David Kaplan, *An Idea of Donnellan* (2012: 154-156)

Abstract & Keywords

Abstract: This thesis investigates the nature of the relation between mental representations in successful verbal communication, thought attribution, agreement, and disagreement — a relation which I call “samethinking”. The nature of samethinking raises several foundational questions about the nature of (non-natural) meaning, and the cognitive underpinnings of the emergence of culture. It bears on long-lasting puzzles in the philosophy of mind and language (such as Frege’s puzzle and Kripke’s puzzle about belief). Samethinking does not amount to sharing a reference (with “sharing" I refer to two or more thinkers having something in common): it is more demanding. How can we explain and characterize this relation, more stringent than coreference, that is instantiated by a pair of thoughts when samethinking takes place? It is often assumed that this relation involves sharing a thought content more fine-grained than reference. In this thesis, I argue that the issue is more complex than what has been commonly assumed, and I suggest an alternative model in which sharing thought content is not necessary.

Keywords: Philosophy of mind & language · Communication · Content · Frege’s puzzle · Relationism

Résumé & Mots-clés

Résumé: Cette thèse étudie la nature de la relation entre les représentations mentales dans la communication verbale réussie, l'attribution des pensées, l'accord et le désaccord — relation que j'appelle "samethinking". La nature du samethinking soulève plusieurs questions fondamentales sur la nature de la signification (non naturelle) et les fondements cognitifs de l'émergence de la culture. Elle concerne également des énigmes de longue date en philosophie de l'esprit et du langage (telles que le problème de Frege et le problème de Kripke sur la croyance). Le samethinking ne se résume pas au partage de la référence (par "partage" je fais référence au fait pour deux ou plusieurs penseurs d'avoir quelque chose en commun) : il est plus exigeant. Comment pouvons-nous expliquer et caractériser cette relation, plus exigeante que la coréférence, qui est instanciée par une paire de pensées lorsque le samethinking a lieu ? On suppose souvent que cette relation implique le partage d'un contenu plus fin que la référence. Dans cette thèse, je soutiens que la question est plus complexe que ce qui a été communément supposé, et je propose un modèle alternatif dans lequel le partage du contenu plus fin que la référence n'est pas nécessaire.

Mots-clés: Philosophie du langage & de l'esprit · Communication · Contenu · Problème de Frege · Relationnisme

Acknowledgements

I would like to express my deepest gratitude to both of my supervisors, François Récanati and Michael Murez. I am very grateful for the confidence they put in my ability to complete this project, and I feel lucky to have benefited from their wisdom and mentorship.

I would like to express my deepest appreciation to Michael Murez (my mentor-turned-co-supervisor) for his unwavering support throughout the years. I cannot say how much this work owes to his help.

I would like to thank the members of my dissertation committee for having accepted to play this time-consuming role. I am very grateful to Stacie Friend, Manuel García-Carpintero, Eleonora Orlando, Ernesto Perini-Santos, and Sajed Tayebi.

Many thanks to Cathal O'Madagain for a number of fruitful exchanges and insights which convinced me that representations sharing was a fascinating topic in its own right.

Many thanks to François Récanati for giving me the opportunity to present materials on my dissertation topic in his vibrant reading group at the Collège de France. I thank all the participants, in particular Maryam Ebrahimi Dinani, Bruno Gnassounou, Philippe Lusson, Michael Murez, Jean-Baptiste Rauzy, Sajed Tayebi.

The Institut Jean Nicod is a great place to do philosophy. Meeting so many inspiring researchers and students over the years has enriched me deeply both professionally and personally. Thank you to all Nicodians.

Thanks should also go to Béatrice Longuenesse and Michael Strevens for welcoming me at New York University during the spring semester of 2019. My stay was made possible thanks to a ENS-NYU scholarship. I would like to extend my sincere thanks to the organizers and participants of the 2019 Norwegian Summer Institute on Language and Mind, in particular, Nicholas Allott, Robyn Anne Carston, Stephen Crain, Anna Drożdżowicz, Steven Gross, Terje Lohndal, Michael Rescorla, Georges Rey, Agustin Vicente. The classes and discussions I had in the summer institute helped me to have a more global viewpoint on many issues that I was interested in.

I would like to thank Sacha Altay, Gregory Bochner, Céline Boisserie-Lacroix, Denis Bühler, Géraldine Carranante, Roberto Casati, Paul Egré, Santiago Echeverri, Luca Gasparri, Anna Giustina, James Hampton, Dan Hoek, Benjamin Icard, Pierre Jacob, François Kammerer, Uriah Kriegel, Armando Lavalle, François Le Corre, Andrew Lee, Marie Lods, Guido Löhr, Robert Long, Salvador Mascarenhas, Olivier Morin, Jesús Navarro, Takuya Niikawa, David Nicolas, François Olivier, Louis Rouillé, Chris Scambler, Nura Sidarus, Merel Semeijn, Ori Simchen, Benjamin Spector, Tristan Thommen, Hugo Trad.

This dissertation was made possible thanks to a grant from the Ministère de l'Enseignement Supérieur et de la Recherche (MESR) of the French government. As a first-generation student, having the time and means to study philosophy means a lot to me. I am grateful to the French taxpayers who financed this work. I hope that at least some of this work will be of some use to other people. I have no doubt that the process of writing this dissertation will allow me to give back at least some of the benefit I gratefully enjoyed as a PhD student.

I am deeply indebted to the Institut Jean Nicod, the Département d'Etudes Cognitives, and François Récanati for the end-of-thesis grant they awarded me when my doctoral fellowship ended.

Je voudrais remercier Nathalie Boudard pour son accompagnement essentiel. Merci également à Edit Mac Clay. Merci à tou·te·s les bibliothécaires de la Médiathèque de Meaux dans laquelle j'ai découvert la philosophie.

Thank you Erika for helping me with the cover! Huge thanks go to the people who so kindly accepted to proofread a chapter or bit of this dissertation: Constant Bonard, Louis Pijaudier, Sam Ducourant, Charles Ehret (who was there for the last sprint!). Thanks also to Gautier Anselin, Jacopo Domenicucci and Sylvain Letoffe for their friendly support.

Une pensée pour mes parents Geneviève et Laurent, pour ma soeur Déborah. Et pour la merveilleuse et irremplaçable Caroline Andrieu.

Last but not least, je voudrais remercier Axel Baptista, Constant Bonard, Maryam Ebrahimi Dinani, Charles Ehret, Jean Tain, Philippe Yahchouchi. Merci pour leur amitié, leur soutien, et nos échanges très enrichissants tant sur le plan personnel que sur celui des idées.

Paris
September 2022

Contents

List of Figures

List of Tables

0 Introduction

This introduction has three parts. Section 1 presents the general problem of the dissertation. It also makes explicit several assumptions endorsed in this work, as well as the theoretical orientation taken. Section 2 introduces three principles which set the stage for what will follow. Section 3 offers an overview of the work.

1 The general problem

Mary says, while pointing to the École Normale Supérieure (ENS), "That is a beautiful place". John agrees, "Yes it is!". In this scenario, Mary and John's utterances refer to the same thing, and it's manifest to the speech participants that they do. It would be natural to say that a thought was communicated, or that Mary and John "share a thought". I will say that there is *samethinking* here, and that communication succeeds. In such cases, communication requires that there be coreference (where expressions corefer just in case they refer to the same thing), and that coreference be manifest to the participants in the discourse. Now, imagine that at some other time, Peter (who does not know Mary or John) says to a friend, while pointing to the ENS: "That is a beautiful place". It seems natural to say that Peter shares a thought with Mary and John, even though Peter does not interact with Mary or John. This is a case of samethinking between agents who do not interact. Contrast it with a further case: Mary says, pointing to the ENS, "That is a beautiful place". John, thinking to agree, says "Yes it is!"—but in fact he is looking at the École des Arts décoratifs in the same street. Here, Mary and John both assume they are thinking about the same thing, but they are not. Despite this mutual assumption of coreference, there is no samethinking here, and communication fails.

But even coreference and the mutual assumption thereof does not seem sufficient for samethinking. Imagine that Leslie and Chris, newcomers to Paris, just talked about the ENS. They do not realize it, but the place in front of them is the ENS. Leslie says to Chris (intending to refer to the institution they have just been talking, not to the place they see), "That is a beautiful place". Chris takes Leslie to refer to the building they both see and, thinking to agree, says

"Yes it is!". Here, Leslie and Chris have coreferential thoughts, and mutually assume that their thoughts are coreferential. Still, there is no samethinking here, and accordingly communication fails, given that Chris is confused about Leslie's communicative intention. Referring to the ENS *qua* the building they see is not what Leslie intended.

Samethinking seems also required for genuine agreement and disagreement between thinkers. If, in the aforementioned situation, Leslie and Chris were having an argument — Leslie saying "that is a beautiful place" and Chris replying "No, it's not" — that would not count as a genuine disagreement. This is because Chris thinks about the building in his visual scene, without recognizing it as the ENS which Leslie and he just talked about. Leslie, on the other hand, thinks about the ENS in an "offline" manner: she is not commenting on the building they both see, which she does not recognize as the institution she just talked about either. But two thoughts genuinely disagree when they could form the basis of a rational constructive dispute between their thinkers. And in order to form the basis of a rational dispute, it is not enough that the thoughts ascribe contradictory properties to the same object: it seems that the object must be thought of in the same way. After all, in the envisaged scenario, Chris might have (additionally) a belief he would expressed with "The ENS is a beautiful place". He would not thereby be guilty of internal inconsistency (even if, from an external point of view, his two thoughts would in fact ascribe contradictory properties to one and the same object).

Samethinking seems also *prima facie* involved in a certain class of attitude reports. Attitude reports are utterances of the form 'S ATTs that *p*', where 'S' stands for the subject whose attitude is reported ('the reportee'), 'ATTs' is a placeholder for any attitude verbs (such as 'believes', 'hopes', 'wonders', 'fears', 'wants', 'anticipates', 'suspects', 'recalls', 'feels', etc) and 'that-*p*' is the subordinate clause describing the content of the reported attitude. There is a useful distinction between attitude reports used *de re*, and attitude reports used *de dicto*. Roughly, a *de re* attitude report merely specifies the object about which an attitude is held, but not the way in which the object features in the attitude. By contrast, a *de dicto* attitude report aims to specify not only the object the attitude is about, but also the way in which the object features in the attitude. An initially tempting account of the contrast between *de dicto* and *de re* uses of attitude reports is the following. While a *de re* reading only requires coreference between the thought expressed by the subordinate clause and the cognitive state reported, a *de dicto* reading requires a more stringent relation, namely, that the thought expressed by the subordinate clause and the cognitive state thus reported, both feature the same object thought of *in the same way*.[1] Going back to the last example, if Leslie, recounting the argument to a friend, were to report Chris' thought with the words "Chris believes that the ENS is not a beautiful place", then Leslie's utterance would not accurately report the thought Chris expressed.

[1]This is *prima facie* not the case for indexical beliefs where the reporter is not in the same contextual relation to the object. Also, see Bach 2002 for the provocative claim that *belief reports do not report beliefs* (even on a *de dicto* use). Let me acknowledge in passing that footnotes in this introduction may only be clear to readers with a certain background knowledge in the relevant literature.

As a starting hypothesis, there thus seems to be some relation, more stringent than coreference, involved in successful communication, agreement, disagreement, and *de dicto* attitude reports. I call this putative relation *samethinking*.[2] The central problem of my dissertation is to specify the characteristics and explanation of such a relation—if there is indeed a unique relation involved in all these phenomena. (If it turns out that there is not one such samethinking relation but *several*, then the issue is to characterize and explain each of them).[3] We may distinguish the *characterization* question from the *explanatory* question. The characterization question asks about the formal properties of the relation underlying samethinking, namely: What is the relation R, more demanding than coreference, that a pair of thoughts of different agents must stand in for there to be samethinking? The explanation question, by contrast, asks about the conditions *in virtue of which* the relation R obtains whenever it does. Answering that question involves citing the factors that *ground* the relation of samethinking.[4]

The simplest and most elegant theory is that the relation R just is identity, and samethinking consists in the replication of thoughts. One version of this view is what I call *Samethinking as Content Identity* (SCI). "Content" in this theory refers to that which, being different between Leslie's and Chris' thoughts, explains the absence of samethinking between Chris and Leslie in the case described above.[5] (SCI) is in fact the combination of three claims:

1. Samethinking is a relation of identity;
2. What is identical is content;
3. Content is finer-grained than referential content.

(SCI) might capture the folk view on the matter; it is at any rate the classical philosophical conception of the phenomenon, see e.g. Frege 1892. Frege wanted content to be shareable

[2] At least some cases of intentional identity seem to exhibit samethinking as well—the classic example being Geach's statement of intentional identity (Geach 1967: 628):

> Hob thinks a witch blighted Bob's mare, and Nob wonders whether she killed Cob's sow.

Accordingly, intentional identity somehow belongs to the explanatory role for samethinking. I do not mention intentional identity in my formulation of the explanatory role for samethinking insofar as I am only marginally concerned with accounting for Geach-type statements in this dissertation. Clarifying the exact relation between samethinking and the typology of intentional identity cases is for another occasion.

[3] The qualification in parenthesis makes sure that this way of stating the explanatory role of samethinking is compatible with there being no unique *explanans* playing this role. But even this formulation might not be absolutely uncontroversial. We can imagine some *radical Referentialists* believing that coreference is necessary and sufficient for samethinking (tentative candidates include Salmon 1986, Soames 2002, Braun 1998). Such theorists will find that the proposed formulation is not a good definition. It's not clear that such a radical Referentialist view does not amount to the *rejection* of samethinking as a genuine phenomenon, however. On this view, successful communication is *merely* transmission of *referential content*: speakers can communicate using a term iff they attach the same referent to it. Relatedly, on this view, there is no robust notion of agreement or disagreement more fine-grained than coreference. Likewise, attitude reports admit only a *de re* reading. We may call *eliminativists*, proponents of such a view. Eliminativists should explain our intuitions about samethinking, perhaps as a pragmatic side effect. I will not engage with eliminativism as defined in this dissertation.

[4] On various sorts of explanation and the notion of grounding in metasemantics, see Burgess et al 2014.

[5] (SCI) is a particular version of the view I call *Samethinking as Identity* that defines the relation R as identity. Since thoughts may not be individuated by content, there are other versions of (SI) than (SCI).

— where "shareability" of content refers to the possibility for two or more thinkers to have thoughts with identical contents. Frege famously held that content "may well be common property of many and is therefore not a part or mode of the single person's mind: for it cannot well be denied that mankind possess a common treasure of thoughts which is transmitted from generation to generation." (Frege 1892: 188 in Martinich 1996). Transmission of thoughts passes for the most part through communication. For language to transmit thoughts, language must be shared. On Frege's conception, *content* is meant to play the role of the meaning of linguistic expressions. But for content to be a suitable candidate to play this role, content must be shared by the speakers of a language.[6]

For this reason, Frege thought one could not identify contents — he called them "senses", and identified them with abstract entities — with mental representations, which are essentially *private*, on his conception.[7] Thus, he writes:

> The referent and sense of a sign are to be distinguished from the associated idea. If the referent of a sign is an object perceivable by the sense, my idea of it is an internal image, arising from memories of sense impressions (. . .). Such an idea is often imbued with feeling; the clarity of its separate parts varies and oscillates. The same sense is not always connected, even in the same man, with the same idea. The idea is subjective: one man's idea is not that of another. There result, as a matter of course, a variety of differences in the ideas associated with the same sense. A painter, a horseman, and a zoologist will probably connect different ideas with the name "Bucephalus". This constitutes an essential distinction between the idea and the sign's sense, which may be the common property of many people, *and so is not a part or a mode of the individual mind*. (. . .) In the light of this, one need have no scruples in speaking simply of *the* sense, whereas in the case of an idea one must, strictly speaking, add whom it belongs to and at what time. It might perhaps be said: just as one man connects this idea, and another that idea, with the same word, so also one man can associate this sense and another that sense. But there still remains a difference in the mode of connection. They are not prevented from grasping the same sense; but they cannot have the same idea. *Si duo idem faciunt, non est idem*. If two persons picture the same thing, each still has his own idea. (Frege 1892: 154-155 Beaney 1997 ed., my emphasis)

We can reconstruct Frege's argument against the identification of contents to mental representations as follows:

(1) Contents (i.e. *senses*) are shareable;

[6] See Burge 1979b for discussion about this way of interpreting Frege. May 2006 is a response to Burge.

[7] For Frege, the fact that contents are abstract means that they belong to a reality distinct both from the physical world and from the internal realm of consciousness (which might not be physical). Since the advent of computers, it is much more natural to conceive that contents can belong to the physical world. In this dissertation, I endorse a naturalization constraint: samethinking must ultimately constitute a potential subject of study for the natural sciences.

(2) Mental representations (i.e. *ideas*) are unshareable;

(3) Therefore, it is not the case that contents are mental representations.

Frege is often said to have missed *the type/token distinction* (see e.g. Margolis & Laurence 1999: 7), which has been originally defined by Peirce (1931-58, sec. 4.537, cited in Wetzel, 2018). To illustrate that distinction, consider the following sentence:[8]

Agapanthe is a variety of violet, and Agastache is a variety of violet too.

In this sentence, there are two numerically distinct word *tokens* of the same word *type* "violet". Likewise, my thought token can be numerically distinct from your thought token, and yet our different thought *tokens* might be of the same *type*. Once we draw the type/token distinction, what Frege appears to be saying is that it is impossible for two persons to have the very same *token* mental representation. But this impossibility does not preclude the sharing of mental representations, since two people can have the same *type* of mental representation.[9] Once we make the type/token distinction, Frege's argument against the identification of contents with mental representations does not go through. In this dissertation, I will be assuming the type/token distinction.[10] The claim that content is identical across thinker's mental representations is thus to be construed as a claim of *type-identity*.

One assumption I make in this dissertation is that thoughts are *structured*. One thing I mean by this is that thoughts are composed of parts. These parts I call "concepts". For instance, the thought PARIS IS CHARMING has the singular concept PARIS as a constituent, where a singular concept is a mental representation that contributes a particular object to the thought in which it features — we may think of it as a *mental name*.[11] Why think thought is composed of concepts? One influential argument starts with the observation that a human mind has the capacity to think a potential infinity of thoughts (which is not to say, of course, that a human mind could *actually* entertain infinitely many thoughts). In virtue of this feature, thought is said to be *productive*. Now, the notion that thoughts are composed of concepts, together with the notion—called *compositionality of thought*—that the content of a thought depends solely on

[8] I mean the sentence *token*, namely, the very inscription you are reading on your examplar of this thesis. If the sentence is construed as a type, the word type "violet" has two *occurrences* in it. See Levy & Olson 1992 (mentioned in Perry 2012: 198) for the idea that electronic documents urge us to include an additional category of *templates*, namely structures that produce tokens on screen or paper.

[9] Frege might more charitably be understood as arguing that type-identity between mental representations would require comparison, and since that is impossible, so is type identity. At least that is what the passage immediately following the one I quoted suggests:

> It is indeed sometimes possible to establish differences in the ideas, or even in the sensations, of different persons; but an exact comparison is not possible, because we cannot have both ideas together in the same consciousness. (Frege 1892: 155 Beaney 1997 ed.)

[10] I also puts forward a different ontology in terms of *stages* and *continuants*, inspired by the metaphysics of biological species and personal identity, see e.g. Kaplan (1990). I need not enter into the details of this model at this stage.

[11] Following the standard usage, I use SMALL CAPS to represent concepts/thoughts.

the content of its parts and the way these parts are arranged, best explain the productivity of thought. Or so the argument goes (Frege 1923/1963, Fodor and Pylyshyn 1988).[12] Another influential argument relies on the fact that, if a human thinker can think the thought e.g. ANNA LOVES BOB, then she can think the thought BOB LOVES ANNA. In virtue of this feature, thought is said to be *systematic*. The notion that thoughts are structured, together with the notion that thought is compositional, best explain (according to this argument) the systematicity of thought.[13] "Thought" is ambiguous between thought-*content*, and thought-*vehicle*, i.e. whatever it is that instantiates thought-contents in the brain.[14] [15] The same ambiguity applies to concepts, which can be construed as content-constituents, or as vehicule-constituents. Accordingly, the claim that thoughts are structured and admit concepts as constituents has two readings: one is about thought-contents, another is about thought-vehicles. Hence, what I am assuming is the following: either thought-contents are structured, or thought-vehicles are structured, or both. This assumption is manifest as I rely on the description of the mental representations of agents when dealing with communication, (dis)agreement and reporting.[16]

The principle according to which agents entertain type-identical thoughts whenever they communicate successfully, genuinely agree/disagree with each other, or successfully ascribe a thought, I call *Shareability* (**SHAR**). Assuming that thought is structured, asking whether thoughts are shareable is *ipso facto* asking whether concepts are shareable: thoughts are shareable if and only if concepts are shareable. I am assuming that genuine disagreement between two thinkers requires the sharing of a thought if Shareability is true. For example, let us say that Anna believes that Paris is charming, and Bob believes that Paris is not charming. Accord-

[12] See e.g. Stalnaker 1984 and the cited literature in section 5 of Rescorla 2019 for discussion.

[13] Let me mention Frege's version of the argument from productivity, which is remarkable in that it involves linguistic communication:

> It is astonishing what language can do. With a few syllables it can express an incalculable number of thoughts, so that even a thought grasped by a terrestrial being for the very first time can be put into a form of words which will be understood by somebody to whom the thought is entirely new. This would be impossible, were we not able to distinguish parts in the thought corresponding to the parts of a sentence, so that the structure of the sentence serves as an image of the structure of the thought. (Frege 1923/1963: 1)

Language is trivially structured (it is not *trivially* compositional). The conclusion of Frege's argument is that both language and thought must be compositional. The compositionality of thought entails that thought is structured. So Frege's argument may be read as an argument for the claim that thought is structured. A human interlocutor has the capacity to understand a never-heard-before sentence uttered by a speaker by thinking a content that is the same as, or suitably related to, the content which the speaker wants to communicate. Pagin (2003) and Westerstahl (2011) propose to read Frege as claiming that the task of the hearer is possible only if there is a structure-preserving mapping between the construction of a sentence, the thought the hearer associates with it, and the thought the speaker associates with it.

[14] I don't mean that vehicles-types are identical to neurological types by definition.

[15] "Thought" is arguably also ambiguous between the *act* of thinking a particular thought, and the output *state* of that act. But this need not concern us here. See Hanks (2015) and Soames (2013) for the view that thoughts (i.e. *contents*) are act-types.

[16] This assumption is not universally shared in philosophy of mind and language. For example, Stalnaker (1976, 1978, 1984, 1999) and Lewis (1973, 1977, 1980) conceive of thought-contents as unstructured propositions (sets of possible worlds without constituent structure). This family of views is purposefully agnostic about the nature of thought-vehicles. While I suspect translations from the representationalist idiom to the sets-of-worlds idiom are possible, I won't provide such a translation manual in this thesis.

ing to Shareability, Anna and Bob genuinely disagree *only if* Bob believes the negation of the very thought that Anna believes.[17] Likewise, I am assuming that the successful ascription of a thought involves that the reporter shares a thought with the reportee if Shareability is true. If Anna sincerely and accurately asserts that Bob believes that Paris is not charming, then (as I will understand the Shareability principle) Anna tokens the thought she ascribes to Bob as part of the thought about Bob she expresses, and shares it with Bob, if Shareability is true. Shareability thus relates activities such as communication, reporting, and rational relations such as agreement and disagreement, with the individuation of thoughts. Individuating thoughts means producing a criterion that decides, for an arbitrary pair of thought tokens, whether these thought tokens are type-identical. The question whether communication and agreement require entertaining identical thoughts may then be formulated in terms of whether Shareability governs the individuation of thoughts. (I will explain what this means in section 2 of this introduction, when I introduce other principles for the individuation of thoughts). If Shareability is true, then sameness of thought is the benchmark of successful communication/understanding, agreement/disagreement, and (at least in central cases) of *de dicto* reporting. In other words, if Shareability is true, then it is part of the *explanatory role* of thoughts that their sharing by agents explain why samethinking obtains whenever it does between them in communication, agreement, disagreement, and thought ascription.

Shareability is a very intuitive idea. It says that if different thinkers communicate successfully with each other, or genuinely agree with each other, then there is a thought that they all entertain. This is an initially plausible hypothesis to explain communication and agreement across thinkers. More generally, (**SHAR**) is an initially plausible hypothesis to account for collective knowledge, and for the fact that collective knowledge is transmitted from generation to generation. Relatedly, on the face of it, we need (**SHAR**) to explain that a speaker can *learn*, be *wrong* about, or can grasp only *part of* the meaning of a word. For assume that meanings are not shared. How are we to explain that a speaker can be wrong about the meaning of a given word? Obviously, we cannot say that the speaker is wrong about her own meaning. But what is it she is wrong about, if not an objective meaning capable of being grasped by several speakers? If people can get concepts wrong, then concepts must be shareable (or so the argument goes). (**SHAR**) is sometimes also thought to be required to account for the existence of psychological laws: if psychological laws subsume psychological states by reference to the concepts they contain, and psychological laws generalize across individuals, then presumably concepts must be shared.[18]

Recently, alternative approaches have emerged, that reconceptualize samethinking in terms of a relation weaker than identity. This characterization is misleading, and I hasten to correct it,

[17] As Frege says, "There can be no negation without something negated, and this is a thought" (Frege 1923: 2).

[18] I will say little about this latter motivation for (**SHAR**). Althought I won't argue for this claim, it is also the weakest motivation to accept (**SHAR**), in my opinion. See Fodor 1994, Schneider 2009, Gray & Almotahari 2020 for discussion.

because it seems to include theories that samethinking is a question of *resemblance* or *similarity* between thoughts.[19] But the family of views I have in mind does not include resemblance theories of samethinking. A better characterization of the family of views I have in mind is to say that samethinking is an *external* relation between thought tokens, on such views. "External", in this context, means that the relevant relation is not determined by the *intrinsic* properties of thought tokens. It takes some work to define this notion precisely, but for present purposes we can be satisfied with the following rough but orienting illustration, due to Gray 2017:

> Consider the difference between being *soul-mates* and being *married*. Being soulmates is a matter of a *match* between the properties of two people. If the personalities, experiences, and so forth, of X and Y match, they are soulmates. This is consistent with X and Y never having met or interacted. Contrast this with *being married*. No facts about X and Y's personalities, tastes, and so forth, determines whether they are married. One has to consider how X and Y are related—in particular, whether they have entered into certain social or legal relations. (Gray 2017: 4)

So for instance, *is-similar-to* (like *is-soulmate-of*, or *is-identical-to*) is not an external relation, because whether two representations are similar is determined by intrinsic properties of those representations. The family of views I have in mind says that *samethinking* is a relation like marriage in that it is *external*. Let us call the family of views on which samethinking is an external relation, *Relationism*.[20] (I will sometimes use, following Prosser (forth), the useful label *Intrinsicalist* to refer to the family of views on which samethinking between representations is—in contrast with the relationist views— determined by intrinsic properties of these representations. *(SI)* is a particular version of the intrinsicalist view. *Samethinking as similarity* is another version of the intrinsicalist view.)[21]

Why might one want to reject *Samethinking as Identity* (SI)? One motivation comes from the communication of so-called "indexical thoughts".[22] Indexical thoughts are thoughts that include mental counterparts of words such as "I", "my", "you", "he", "his", "she", "it", or the demonstrative pronouns "that", "this", adverbs such as "here", "now", "tomorrow", "yesterday", the adjectives "past", "present", etc. Indexical thoughts can been said *essentially context-bound*

[19] Fodor and Lepore proposed an argument against this class of theory that proved to be very influential. Here is a relevant passage: "It seems sort of plausible that you can't have a robust notion of *similar* such and suches unless you have a correspondingly robust notion of *identical* such and suches. The problem isn't, notice, that if holism is true, then the conditions for belief identity are hard to meet; it's that, if holism is true, then the notion of "tokens of the same belief type" is defined *only* for the case in which *every* belief is shared. Holism provides no notion of belief-type identity that is defined for any other case and no hint of how to construct one." (Fodor & Lepore 1992: 19)

[20] The label "Relationism" has been popularized by Fine 2007. Cumming 2013a and Schroeter 2012 use the label "relational".

[21] The reader will find in the general conclusion of the dissertation, a tree diagram that delinates all the competing views on the samethinking problem.

[22] In the thesis, I deploy arguments against (SI) that are independent from the communication of indexical thoughts.

(Burge 1979), because they can be entertained only when thinkers are in a certain contextual relation to the object the thought is about. For example, it seems that the thought I AM MAKING A MESS thought by me (RB) in a given shop at a certain time cannot be thought *in the very same way* by someone other than myself, because nobody except me is in the relation of identity to RB.[23] As Frege says, "everyone is presented to himself in a special and primitive way, in which he is presented to no one else". I-thoughts seem to involve this primitive way in which everyone is presented to himself and to no one else. So it is *prima facie* the case that I-thoughts cannot be shared (whereas their referential content or their role can). Similar remarks seem to apply to other mental indexicals. Here is an example to convey the correpsonding intuition. Imagine that Chris and Leslie are jointly perceiving another person, call her Leïla. Chris sees Leïla from the side, but Leslie sees Leïla from the front. Leslie says to Chris (attending and pointing to Leïla) "This person is beautiful". Imagine that Leïla is indeed beautiful, but that she is not at her profile advantage. Chris and Leslie have different visual perspectives on Leïla. (They may also have different character traits, different standards for what count as beautiful, etc, which make their perspectives different also in a non-visual sense). Assumes that Chris understands what Leslie says. Intuitively, this is compatible with Chris having a demonstrative thought about Leïla different from Leslie's. There is a clear sense in which the demonstrative thoughts they each deploy cannot be matched by another thinker at the same time.[24] As an upshot, to explain how some indexical thoughts *can* be communicated, it seems that we need to reject (SI). As Martin Davies writes,

> The doctrine that in successful communication the hearer (audience) comes to have a thought with the same content as the thought expressed by the speaker obviously needs to be complicated in the case of communication using demonstratives. (Davis 1982: 293 cited in Recanati 2016: 112).

Schiffer makes a similar point when he recognizes that some propositions have what he calls *the relativity feature*:

> A proposition has [the relativity feature] provided it's an x-dependent proposition [i.e. a singular proposition, on which more shortly] the entertainment of which requires different people, or the same person at different times or places, to think of x in *different ways*. (Schiffer 2005: 141)

[23] Of course, there is a sense in which different subjects who think I AM MAKING A MESS are sharing an indexical thought. But that sense of "sharing an indexical thought" comes apart from agreement/disagreement, and seems to require a distinction between narrow and wide content, or character and content, or the introduction of centered worlds. I refer the reader to Garcia-Carpintero & Torre 2016, an edited collection on *de se* thought and communication with many interesting discussions on this issue. See also Torre & Weber 2021 for a review of the state of the art on *de se* attitudes.

[24] We can follow Kaplan 1989 to model perspectives. On Kaplan's proposal, to think of an object under a perspective is to think of it as "the individual that has appearance A from here now—where an appearance is something like a picture with a little arrow pointing to the relevant subject" (Kaplan 1989: 526 cited in Recanati (forth)). Kaplan's model arguably leaves out many factors that are involved in having a perspective on an object, but it is a place to start.

To recap: it seems that some indexical thoughts, being perspectival, cannot be shared. If this is true, then the communication of some indexical thoughts does not involve the replication of thought from speaker to hearer. This suggests a reconceptualization of communication, and more generally samethinking, as a form of *coordination* of thoughts—where "coordination" refers to an external relation. I call this conception, *Relationism*.[25] In this dissertation, I aim to contribute to this alternative picture of communication, and develop this alternative picture with respect to samethinking more generally. Accordingly, my investigation of the explanation and characteristics of the *samethinking* relation — in particular, whether *samethinking* can be construed as type-identity between thoughts — is largely internal to this new way of conceiving cognitive and linguistic sameness across agents. The dissertation can be seen as a cumulative argument against (SI) combined with a proof of concept for an alternative model.[26]

This dissertation investigates samethinking between a certain kind of thoughts, which philosophers call *singular*. When a thought is not singular, it is general (and there might be indeterminate cases in between). Let me illustrate this distinction. Marc forms, on purely general grounds, the belief that the most talented illustrator is interesting. This is an example of general thought. By contrast, if the most talented illustrator is Marc's girlfriend called Caroline, and Marc thinks that Caroline is interesting, Marc's thought is singular. Even assuming that Marc's general thought in fact refers to Caroline, the two thoughts are importantly different in the way they refer to their object. For example, the object of the singular thought is determined via a relation its thinker has to the referent; not so for the general thought. Or, Marc's *Caroline* thought could not exist without Caroline. But Marc's general thought *could*. Etc. Defining what makes thoughts singular, as opposed to general, is not a trivial task.[27] Fortunately, we may hope to say something helpful and explanatory on samethinking between singular thoughts, without precisely defining what it is for a thought to be singular.[28] My dissertation focuses more specifically on samethinking between singular thoughts involving non-indexical *referential* concepts such as the ones thinkers associate with *names*. (I am sorry to say that I will not be concerned with fictional thoughts in this dissertation).[29] You can think of the dialectic of this dissertation as follows. I mentioned that indexical thoughts provided the best or most obvious case for a reconceptualization of communication as a form of coordination of thoughts, as opposed to

[25]Note that some relationists (Dickie & Rattan 2010, Cumming 2013a,b, Heck 1995) want to validate (**SHAR**) by construing the coordination relation as an equivalence relation—i.e. a relation that is reflexive, symmetric and transitive. Hence (SI) entails (**SHAR**), but (**SHAR**) does not require (SI) because (**SHAR**) can be obtained through a relation other than identity, provided that the relation is an equivalence relation. Following the useful terminology of Dickie & Rattan 2010, I call this version of relationism, *Equivalence Class Fregeanism*. They are my main interlocutors in this dissertation.

[26]The dissertation argues not only against (SI) but also against (**SHAR**). Specifially, I take issue with Equivalence Class Fregeans. I refer the reader to the preceding note on the relation between (SI) and (**SHAR**).

[27]For a recent edited collection on the issue see Jeshion 2010.

[28]This methodology is defended in Kaplan 2012: 145 and passim.

[29]On samethinking with respect to fictional concepts and thoughts—sometimes referred to as 'co-identification' in the literature (after Friend 2014) see e.g. the aforementioned article; discussed in Garcia-Carpintero 2018; Garcia-Carpintero & Martí 2014; Sainsbury 2005; Everett 2000 & 2013; Perry 2012; Orlando 2017; Maier 2017; Recanati 2018; Kamp 2021; Maier & Stokke 2021; Semeijn 2021.

Samethinking as Identity (SI). So, if one can argue against (SI) *even in cases not involving indexical thoughts*, that makes the case against (SI) and for Relationism even stronger. That being said, indexical thoughts, and in particular (as I will explain) joint attention, is a paradigm, indeed the main inspiration for the relationist model of samethinking I develop in the dissertation.[30]

This was the presentation of the problem, some of the assumptions made, and the orientation taken in this thesis. I will now discuss three notions that will be used recurrently in this work (one of which I have already mentioned).

2 Three principles for the individuation of thoughts

Another way to frame some of the main questions of this work is in terms of whether communication transmits content, or whether reporting attitudes demands a match in content between what the reporter says and what the reportee thinks. A theory of samethinking in linguistic communication and reporting is thus a theory of the relation between mental and linguistic content. But what is content? This section presents a preliminary discussion of three relevant principles for the individuation of content as I approach this issue in this work.

There are many different views of mental content. Moreover, as already noted, thoughts may not be individuated by content on some views. But it is uncontroversial that thoughts and concepts are posited to explain/predict behaviour. In order to fulfill this explanatory role, thoughts and concepts must be (as is well-known) individuated more finely than reference.

Consider an example. Anna believes that Ajar is a French writer and she believes that Gary is a different, American, writer. In fact, Ajar is Gary. This is said to be an example of a "Frege case", that is, a case in which a rational agent has two distinct representations that refer to the same object but which the agent does not take to be coreferential. In other words, going back to Anna, she is in a Frege case with respect to Gary/Ajar because there is a property F such that Anna has conflicting attitudes towards the content referentially individuated that Ajar is F, without being thereby irrational. For example, the action of Anna buying a book that mentions "Gary" on the cover is rationalized by (and can be explained in part by citing) the desire of hers that would be expressed with "I want to read Gary" (together with the belief that the book is authored by Gary, etc.), but not rationalized by (and not explainable by citing) the desire of hers that would be expressed with "I want to read Ajar". Things are different for someone, say Bob, who is unconfused about Gary/Ajar. In Bob's case, the action of buying a book tant mentions "Gary" on the cover *is* rationalized by his desire that would be expressed with "I want to read Ajar", because his attitudes about Gary/Ajar, unlike Anna, are all in rational contiguity with one another. Intuitively, Anna's desire that would be expressed with the

[30] A *prima facie* reason for modelling samethinking uniformly for both indexical and non-indexical thoughts is that in many thinkers some non-indexical thoughts are in the samethinking relation to some indexical thoughts.

name "Ajar" and the desire that would be expressed with the name "Gary" should be distinguished, even though they are coreferential, because they have different *cognitive roles* for Anna.

As an upshot, to account for thoughts' cognitive roles, we need to individuate them in such a way that any two thoughts with respect to which a rational agent can have conflicting attitudes at the same time—should be counted as different. This is Frege's constraint (**FC**), on which I do not dwell here, because it will be articulated in the chapters to come. A remark on the force of this constraint. Without this constraint, one could not even recognize Frege cases as an empirical phenomenon. How we should account for the Frege cases is a question that is subsequent to recognizing their existence as an empirical phenomenon in one's theory. Hence in this thesis I take (**FC**) to be a non-negotiable principle for any theory of mental representation and samethinking, no matter how the theory deals with the Frege cases.[31]

I have already mentioned the principle, which I will now discuss again, according to which thinkers routinely have thoughts that are type-identical (for example, whenever they communicate, or agree with each other). This is the principle of Shareability (**SHAR**)—a classical constraint on the individuation of thoughts. This principle rules out the variety of *same-thought* relations or criteria that would make thinkers' thoughts almost never shared in practice.[32] For example, a criterion on which two concept tokens from different thinkers are type-identical if and only if they have the same *total* computational roles would make (**SHAR**) false, because it is an extremely demanding criterion that is (we may suppose) virtually never met, hence ruled out by (**SHAR**). (To see why, consider that a subtle difference in affective valence (or "micro-valence") between, say, your APPLE concept and my APPLE concept, everything else being shared, would arguably suffice to make their respective computational roles different).[33] [34] In short, Shareability imposes the need for coarse-graining the individuation of thoughts.[35]

Another principle, which does not involve intersubjectivity or communication, also pushes the granularity of thought individuation towards the coarse-grained. I will call it "Campbell's constraint", after Murez (2016 :170). *Frege's Constraint* is useful to individuate thoughts that a

[31] But see Gray 2022 and Speaks 2013 for discussion.

[32] See Fodor 1998 p. 28 for another formulation of this principle.

[33] The claim that micro-valence is part of the total computational role is not uncontroversial, and strongly depends on how 'computational' is defined. I use it only as an illustration.

[34] If the aforementioned argument by Fodor and Lepore against content similarity is correct, then it is not clear what this notion of "everything else being shared' means, if the identity of a concept involves all the other concepts possessed by the cognitive system. I ignore this complication here, see e.g. Pollock 2020 for a recent response.

[35] What do we mean by 'coarse-graining' the individuation of thoughts? Abstractly, we may think of any *same-thought* relation as an equivalence relation for grouping thought tokens into subsets in terms of their content. Given a set of thought tokens, the *same-thought* relation thus determines a partition of that set, that is, a disjoint union of non-empty subsets—the "parts" of the partition—in such a way that every thought token is included in one and only one subset. For example, that a given *same-thought* relation $\sim_C$ is finer-grained than the *coreference* relation means that $\sim_C$ splits the parts of the referential partition into smaller parts: the more demanding the criterion, the finer the partition, the more parts the partition will have by the criterion, and conversely. One clear consequence of (SHAR) with respect to the partition of the set of thought tokens is that it must *be properly coarser* than the partition of singletons (i.e. such that every part of the partition is a singleton).

single agent has at a particular time. But, being a synchronic criterion, it does not go beyond that. However, it seems essential to the explanation of behavior that concepts of a single agent be able to recur across time or across attitudes. Consider the following sequence in Anna's life. Anna believes that Gary's books are interesting. She reasons that she can find one of Gary's books in the Gibert Joseph Paris bookstore. Later in the day, she forms the intention to go to the bookstore and buy a book by Gary. When Anna forms her intention to go to the bookstore, it is obvious to her that the thoughts she had in the sequence concern the same individual, namely Gary. Likewise, it is obvious to her that each time she thinks BOOK in the sequence, she thinks about the same thing. Now, the way that sameness of reference in such cases is manifest to her does not rely on an explicit representation of coreference. Rather, Anna simply "trades on identity" (Campbell 1989) or—as we might also say—she simply uses e.g. the concepts BOOK or GARY as a "middle term" in her train of thoughts (Millikan 2000: 141-143). Trading on identity/using a concept as a middle term can be illustrated with respect to the formal validity of certain arguments. Consider the two following arguments:

Argument 1:	**Argument 2:**
(P1) Superman is F	(P1) Superman is F
(P2) Superman is G	(P2) Clark Kent is G
(C) Therefore, $\exists x$ (Fx $\wedge$ Gx)	(C) Therefore, $\exists x$ (Fx $\wedge$ Gx)

Argument **1** is formally valid. By contrast, even if the occurrences 'Superman' and 'Clark Kent' corefer, the argument **2** is not formally valid. For it to be valid, we need to add the premiss that Superman is identical to Clark Kent. Trading on identity does not seem essentially tied to the inferential pattern of existential generalization but may occur in any pattern of reasoning whose validity turns on whether expressions allows trading on identity, for example:

Argument 3:
(P1) Superman is F
(P2) If Superman is F, then Superman is G
(C) Therefore, Superman is G.

It is controversial how to analyze the relation in virtue of which trading on identity is allowed whenever it is.[36] According to one view, trading on identity is a matter of different token expressions being of the same type, i.e. the relevant relation is identity (of content, or of expression). According to another view, trading on identity is a matter of an external relation, in particular, weaker than identity (Fine 2007, Gray 1017). This dissertation is concerned with this debate with respect to the *interpersonal* domain. In this domain, the issue can be formulated as follows: what is it that allows different agents to trade on the identity of each other's thoughts? One family of views (SI) claims that this is a matter of type-identity between thoughts. Another

[36]This relation is often called "coreference *de jure*" in the literature.

family of views (Relationism) claims that this is an external relation (weaker than identity). Ultimately, I will propose to distinguish the relation—whatever it is— that allows trading on identity across different agents, and the relation—whatever it is— of samethinking proper. I will not go over the details of this proposal here, because the material required to make the distinction will be introduced gradually in the relevant chapters. I will now provide the reader with a bird's eye view of the work.

3 Overview of the dissertation

This dissertation is divided into two parts. The first part deals with samethinking in communication. The second part deals with samethinking outside of communication, namely, in thought ascription, and in agreement and disagreement. Note that this section will use technical expressions that I will define in the relevant chapters.

Part 1 — "Samethinking in communication" — is constituted of two chapters. The first one sets a problem that needs to be solved, and argues against two proposed solutions. The second one defends an alternative solution.

In chapter 1 — "Communication, content, and the (Super-)Loar cases" — I raise the following problem: under which conditions do people communicate successfully, if not simply as a result of recovering the right referential content? I clarify what communicative success amounts to. In particular, I deploy an argument to the effect that successful communication cannot be lucky. Then I explain why the conception according to which communicative success is a matter of thinking identical content on the part of speaker and hearer is not satisfactory. Since the recovering of referential content is not sufficient, as shown in the *Loar case* presented in the chapter (a communicational variant on Frege Cases), the content that must putatively be grasped for communicative success is finer-grained than reference.

I examine two important instances of samethinking in communication as content identity. The first one I call the Standard Fregean view. It says that speech participants successfully communicate about an object *o* just in case they share descriptive modes of presentation for *o*. Drawing on Buchanan 2013 and Tayebi 2013, I use intuitions about cases to show that this condition is not sufficient: it is always possible that participants share the same referential content under the same descriptive mode of presentation, but do so by luck. The second view I call the Sophisticated Fregean view. It says that modes of presentation (MOPs) are non-descriptive, in particular that their reference is determined via causal relations to the environment. It has a *relational* conception of shared contents, on which a shared content is an equivalence class of non-descriptive MOPs suitably related to each other in a situation. Importantly, the relevant relation is external. By this I mean that participants can fail to notice when the relevant relation does not obtain. I will say that the postulated shareable contents are not *transparent*. In part

because such postulated contents are not transparent, I argue that they create more problems than they solve. The *pars destruens* may be summarized in terms of the following dilemma: either identity in content is gettierizable i.e. may be arrived at by mere luck (on the Standard Fregean conception), or else difference in content is not transparent (on the Sophisticated Fregean conception). The dilemma gives us reason to think that we should not understand successful communication in terms of fine-grained shared content. The upshot of the chapter is that the condition of non-luckiness of coreference in communication must be understood as a *causal* condition and not as a *semantic* condition.

In chapter 2 — entitled "On what might prevent communicative luck" and which is the central chapter of part 1 — I examine another important candidate solution to the problem, and explain why it too is inadequate. Then, building on it, I offer my own solution. The chapter shifts the focus of the discussion to the idea that communication is a matter of intention recognition. A central theme is the idea that the referential plan of a speaker (namely, her plan to make her audience think of a certain object) typically includes the intention that certain features of the utterance or of the context be utilized in how the hearer retrieves what the speaker intends to communicate about. Drawing on Buchanan 2013, I incorporate this idea into the following anti-luck condition: the hearer must interpret the speaker's utterance in virtue of attending to the information the speaker intends the audience to use in order to retrieve the referent (I call this bit of information *ib-feature*, after Schiffer (forthcoming/b)).

Relying on the literature, I present two cases (Tayebi 2013, Peet 2016) showing that this condition is not a general solution to the problem. I then introduce joint attention as a communicative safety mechanism. I distinguish two kinds of communication : *deictic* where the object talked about is present and observable in the discourse situation; and *non-deictic*, where the object is not present or observable in the discourse situation. I subsequently explain how joint-attention can be used to analyze communicative success in both kinds of communication. The criterion I arrive at is (roughly) the following: the hearer's interpretation of the speaker's utterance is wholly governed by the intended ib-feature the speaker and hearer jointly attend, and as a nondeviant result of this, the hearer retrieves the right referent.

The idea behind the proposal is this. Joint attention provides coreferential safety because it is a factive state—one the discourse participants can only be in if they are actually focussing on the same object with the common awareness that they are. When this happens, speaker and hearer are referring *together*, as it were. Joint attention on ib-features thus brings it about that every element of contextual information used in interpreting the utterance is not only mutually known, but (roughly) commonly known.[37] As a result, speaker and hearer have common knowledge that the hearer is recovering the correct interpretation, and luck is eliminated. I

[37] Where x is *mutually known* among a set of agents if each agent knows x; whereas x is *commonly known* among a set of agents if x is mutually known among that set of agent, and each agent knows that each agent knows x, and each agent knows that each agent knows that each agent knows x, and so forth *ad infinitum*.

call this criterion "IB joint attentional criterion for communicative success". I explain why this criterion is a step towards an approach to the common ground (roughly, the set of propositions and references assumed to be already shared between the speech participants) which is less intellectualistic than mainstream views thereof.[38] In closing, I offer some reflections on the following outstanding question: if the joint attentional approach is on the right track, how is the common ground established in non face-to-face communication? Lastly, I compare the proposed solution to the Sophisticated Fregean view examined in chapter 1.

Part 2 — "Samethinking outside of communication" — deals with the following problem: what is it for different thinkers who do not interact to samethink? This part, which consists of three chapters, proceeds in a similar fashion as the first part: it considers different variants of the conception of samethinking outside of communication as content identity, and explains why they are not satisfactory (chapter 3); then it gradually defends an alternative model (chapter 4 & 5).

Chapter 3 — "From *alignment* to *pragmalignment*" — examines communication involving proper names as a touchstone for competing views of samethinking outside of communication. It thus makes the transition between the two parts of the thesis, and is a centerpiece of it. How can name-involving communication lead us to samethinking *outside* of communication? To illustrate, if you know the name "Napoleon", it is because it was transmitted to you through an utterance. The transmission path leads back to an initiating use of the name that establishes the name-using practice. All users of the name "Napoleon" are connected to each other by such a transmission path.[39] I observe that, on the face of it, membership in the network seems to guarantee that a concept is shared. If a speaker is competent with the name "Napoleon", she can be said to have common knowledge of Napoleon—and to share the concept NAPOLEON—with all the users of "Napoleon" (or so goes the initially tempting thought). In other words: when it comes to name-involving thoughts, the *same-concept-as* relation seems to reduce to membership in a same transmission path.

The chapter starts by considering a theory due to Onofri 2018 that precisifies this idea (an idea common to causal-historical models of samethinking). I show that membership in a same

[38] The common ground (CG) is classically defined as common belief (Stalnaker 2002):

> (. . .) The common ground of a conversation is just what is common belief among the participants in a conversation. (Stalnaker 2002: 706)

In the same article, Stalnaker subsequently proposes another characterization of the common ground as common *acceptance*, where to accept a proposition is "to treat it as true for some reason. One ignores, at least temporarily, and perhaps in a limited context, the possibility that it is false." On this widely accepted characterization of the CG:

> It is common ground that ϕ in a group if all members accept (for the purpose of the conversation) that ϕ, and all believe that all accept that ϕ, and all believe that all believe that all accept that ϕ, etc. (Stalnaker 2002: 716)

[39] I ignore the different types of transmission path in this chapter synopsis insofar as they have no role at this level of abstraction.

communicative path, when construed as a relational individuation criterion for thoughts, is too coarse-grained to account for thoughs' cognitive significance and transparency: it identifies thoughts that are different for their thinkers. I subsequently propose a modification of this idea which technically solves the problem. I explain that the resulting relational individuation criterion for thoughts is stipulative: it seems to arbitrarily excludes agents from communicative chains just for the sake of restoring a compatibility with Frege's constraint. However, evidence is needed that the stipulated clause excluding relevant Frege cases from the communicative chains is necessary for explaining communicative success.

This leads to the following question: can a speaker in a Frege case with respect to an object *o* communicate successfully about *o* with a conversational partner who is not in the relevant Frege case? A *negative* answer to this question imposes a condition of *alignment* on successful referential communication. Alignment obtains between agents' conceptual repertoires just in case (very roughly) the agents' communicative dispositions relate their concepts in a one-to-one manner. In the second part of the chapter, following Cumming 2013a,b, I show that alignment is (modulo some assumptions) necessary for any relational individuation criterion of concepts to satisfy both Frege's constraint and Shareability. Hence the aforementioned question has a crucial status, indeed decides whether thoughts are shareable. This is another important dimension why I consider communication as a touchstone for competing views of samethinking *simpliciter* in this chapter. Rejecting alignment as a background condition for successful communication is *ipso facto* rejecting Shareability.

The rest of the chapter puts forward a set of arguments for the claim that misaligned agents *can* successfully communicate. I argue that if the standards for communicative success are context-sensitive, then alignment is not a necessary condition on successful communication. I then go on to argue that the standards for communicative success *are* context-sensitive. Assuming that knowing what is said involves being able to rule out all relevant alternatives, *which* alternatives are relevant depend on the conversational context. I suggest two different specific views of this context-sensitivity.

The first view I suggest is the *pragmatic encroachment* view, according to which (roughly) the standards for knowing what is said may depend on the practical costs of misunderstanding. My discussion culminates in an attempt to provide a pragmatic twist to the constraint of alignment, which constitutes the second view of the context-sensitivity of the standards of communicative success I suggest, namely, the *psychological status* view. Drawing on the linguistic theory of the *Givenness Hierarchy* (GH), I observe that the cognitive status of a concept (i.e. roughly, its degree of accessibility in memory and attentional states)—as assumed by the speaker—plays an important role in linguistic communication. According to (GH), whenever speakers use pronouns and determiners (such as *it*, *this/that/this NP*; *the NP*; *a NP*, etc), they make implicit assumptions about the cognitive statuses the object under discussion has in the minds of their

conversational partners (e.g. Hedberg 2013, Féry & Krifka 2008).[40]

These cognitives statuses help define a notion of *relevance* applied to concepts: which concepts are relevant are, I suggest, thoses with a certain degree of accessibility (namely the activated ones). In contrast, the standard notion of alignment is insensitive to any notion of relevance, which yields (I argue) some counterintuitive predictions about cases. With this notion of relevance applied to concepts, I define a pragmatic version of the constraint of alignment restricted to the domain of activated concepts. I call the resulting constraint *pragmalignment*, and illustrates how it works. Pragmalignment seems to make more intuitive predictions about cases than the standard notion.

Having argued that the domain of the standard notion of alignment was too broad, I suggest that it is also, in an important sense, too narrow. *Representing the perspective* of a misaligned agent is, I suggest, a way to successfully coordinate in misaligned communication. I provide a definition of pragmalignment that incorporates this idea, relying on the mental file theory (Recanati 2012, 2016). This is the last pragmatic twist on the constraint of alignment I explore in the chapter. In closing, I point out a strong limitation of pragmalignment: it is a synchronic and context-bound notion. Hence, as is, it is unable to account for samethinking in thought attribution (as when I report beliefs Aristotle held long ago), or in agreement and disagreement between agents who do not interact.

Chapter 4 — "Pragmalignment in action: attitude and speech reports" — generalizes the relation of pragmalignment to diachronic and context-spanning samethinking, thus filling the gap pointed out in the end of chapter 3. An account of samethinking in attitude/speech reports, and agreement/disagreement without shared content is presented and defended. To do this, I examine networks of mental files associated with the use of names in causal-historical chains—more specifically, Perry's description of them (Perry 2012). Perry calls them *intersubjective file-networks*. The chapter articulates how Perry defines a *same-saying* relation without alignment in terms of the file-networks. Perry's solution involves a further partitioning of the network—which he calls *thread*—tracking which file of an agent is involved in a particular discourse or thought context, and how that file is used or updated in that context. I emphasize the significant convergence of Perry's notion of a thread with the notion of pragmalignment—and the psychological status view—introduced in the previous chapter. Getting hold of Perry's notion of a thread, I use it to generalize pragmalignment to diachronic and counterfactual attitude reports. When reporting the attitudes of a thinker in a Frege case with respect to an object, reporters have in mind particular ways the thinker has of thinking about the object.

[40] As I explain in the chapter, the idea that referring expressions encode features that indicate whether a referent is present in the common ground and its degree of accessibility in the memory/attentional states in the hearer's mind—as assumed by the speaker— fits well with the IB-joint attentional criterion of communicative success defended in chapter 2. If (GH) is on the right track, then the use of pronouns and determiners involve joint attention both on these ib-features and the intended referent of the linguistic expression. Hedberg 2013 proposes that such cognitives status determine necessary and sufficient conditions on the use of pronouns and determiners.

In doing so, they implicitly distinguish a thread in the file-network, which accounts for the sensitivity of speech and attitude reports to the status of particular mental files.

I point out an important aspect of the view that is not explicit in Perry's characterization (who often uses an intrinsicalist/(SCI) terminology):[41] in misaligned configurations, content sharing relative to a thread does not amount to content identity other than coarse-grained content, because (as shown in chapter 4) agents that are in a Frege case with respect to a referent introduce additional spurious information not matched by misaligned agents. Perry is, I suggest, best construed as a relationist. Capitalizing on the previous chapter, I provide a particular view on how speakers implicitly target threads when ascribing thoughts: they do so, I suggest, by indexing files to the perspective of the ascribee. I define this idea and illustrate how it works on Kripke's puzzle about belief. The last part of the chapter deals with agreement and disagreement without interaction. I contrast the moderate contextualist, according to whom (roughly) issues of agreement and disagreement are fully decided by agents' communicative dispositions, with the radical contextualist, according to whom such issues irreducibly involve an interpreter. I claim that Perry is committed to the latter position, and offer some thoughts on the costs and benefits of each position.

Chapter 5 — "Participating in representational traditions" — addresses the following problem. If concepts are not shared, how is it that a speaker can learn, be wrong about, or have merely partial knowledge of—what a word means?[42] The chapter begins by offering a typology of the distribution of concepts (i.e. the extent and manner in which they are spread in a population, in a theory-neutral sense of "spread"), and locates word meanings within this typology. The rest of the chapter provides a metasemantic story to account for *learnability*, *being wrong* and *partial grasp* without shared meanings other than extensions (i.e. referent, class, property, etc). The proposed metasemantic story relies on two claims. The first claim is that the use of the same words trigger appearances of semantic sameness in language users.[43] The second claim—drawing from Schroeter 2012— is that these semantic appearances make it the case that things happen as if meanings were shared, and give rise to representational traditions. Speakers intend to conform their uses of words to these representational traditions.

I propose a particular view of what these representational traditions are. Following Recanati 2016 and again Schroeter 2012, I propose that what underlies semantic deference at the metasemantic level are peer-to-peer distributed files managed at the community level. How are they managed? I suggest that the way *Wikipedia* encyclopedic entries are managed is a good reflection of the social mechanisms by which the community manages a distributed file, and outline

[41] SCI= Samethinking as Content Identity.

[42] I am ignoring many difficult issues on how concepts and lexical meanings relate for present purposes. See e.g. Glanzberg 2018 for discussion.

[43] As I make it clear in the chapter, I don't mean we cannot have a functional characterization of such mutual semantic appearances.

some of them. Why do speakers commit to corefer with both their past uses and the (assumed) community use? Drawing inspiration from O'Madagain 2018, I explore the notion that people defer in the semantic sense in order to defer in the epistemic sense. Seeking knowledge about encyclopedic entries of shared interest is an important reason for using words in a deferential way.

In the general conclusion — "What is samethinking?" — I offer a comprehensive map of the competing views. Taking advantage of the theoretical distance provided by the work carried out in this thesis, this chapter is concerned with defining general classes into which views of samethinking might fall, identifying their main respective consequences. Next, I locate the model I have suggested in this dissertation within the delineated solution space, and draw some notable implications of this model. I then indicate two lines of research, which I think are worth pursuing in order to further develop the model suggested in this thesis. Finally, I conclude this work by distinguishing between two important notions, which I believe have not been clearly distinguished in the literature.

Part I

Samethinking in communication

1 Communication, content, and the (Super-)Loar cases

Abstract

Successful communication requires participants not only to think about the same thing, but also to think about it in the same way — as illustrated by an example due to Loar 1976. What is thinking about the same thing in the same way (which I abbreviate with the term *samethinking* in this thesis)? To account for this, many authors have felt the need to posit a level of shared content finer-grained than referential content. One may call this view *samethinking as content identity* (SCI). In this chapter, I argue that (SCI) is not a good theory of successful communication.

Following Buchanan 2013, I argue that successful communication is not the same as sharing content, whatever the notion of content one chooses. My main argument takes the form of a challenge: for whatever notion of fine-grained content one chooses, one can always *gettierize* the recovering of fine-grained content in communication, that is, make it so that the interpretation process is *lucky*. I call such cases, Super-Loar cases (using a recipe found in Tayebi 2013). The possibility of constructing a Loar case for any account of fine-grained shared content one chooses thus saps the motivation for preferring it to the referential view.

However, when modes of presentation (MOPs) – the mental representations that one needs to postulate in order to account for the possibility of *Frege cases* – are individuated by *external* relations to the environment, they are such that they cannot be shared accidentally across thinkers (i.e. their type-identity in communication is not gettierizable). MOPs externalistically individuated thus seem to escape the threat of the Super-Loar cases.

But it is also the case that externalistically individuated MOPs are not transparent: that is, whether two MOPs are the same or different may not be subjectively distinguishable, on this picture. A similar consideration applies to *shared senses* individuated as equivalence classes of inter-coordinated MOPs relative to a discourse situation: senses so individuated are such that a difference in sense across participants' respective MOPs need not be transparent to them. Hence my resulting argument against (SCI) as a theory of communication takes the form of a dilemma: either identity in content is gettierizable, or else difference in content is not transparent. The true moral of the Loar case is that communication is causal, not that communicated content is fine-grained.

1 Introduction

1.1 Chapter plan

I start by presenting what I call the naïve conception of communication, namely, the pre-theoretical idea that successful communication involves a match in content between the thought expressed by the speaker and the interpretation of the hearer. Then I present an hypothetical scenario put forward by Loar (1976) suggesting that recovering the right referential content is not enough for successful communication; discuss the relation between communication, luck, and knowledge on the basis of Loar's example; lastly I precisify the notion of communicated content that is the focus of my investigation.

In section 2, I present a theoretical flesh out of the *Naïve Conception* in terms of shared content, where the relevant notion of communicated content is (based on Loar's example) more fine-grained than referential content. I call this family of views, the (standard) Fregean views. Then in section 3, I argue that one can always gettierize the recovering of content in communication according to a recipe I call 'Super-Loar case'. Demanding that a finer-grained level of content be shared cannot do anything against this: it is at best a necessary, but not a sufficient condition for successful communication.

In section 4, I present a more sophisticated Fregean view, which acknowledges the causal aspect of communication, but still want to account for it in terms of fine-grained shared content. As a first approximation, Sophisticated Fregeans (as I call them) may be construed as reversing the order of explanation put forward by the standard Fregean view: roughly, identity in content is explained by the relation underwritting communicative success, rather than the other way around (Dickie & Rattan 2010; Campbell 1987). This variety of Fregean view is *relationist* because the relevant notion of shared content is not reducible to intrinsic features of representations. Instead, it depends on external relations between them. By individuating shared content in terms of an external relation, Sophisticated Fregeans seem to escape the threat of the Super-Loar cases, which are simply construed as cases where the relevant relation does not hold between the MOPs of the participants. While there are many features I like in the Sophisticated Fregean approach, I argue (in section 5) that it is ultimately misleading to construe the safety condition required for successful communication in terms of content sharing: in doing so, sophisticated Fregeans are still confusing content with a condition on the causal transmission chain.

My main argument against the Fregean views of both sorts take the form of the following dilemma: either identity of content is gettierizable (on the standard Fregean conception) or else difference in content is not transparent (on the sophisticated Fregean conception).

1.2 The naïve conception of communication

Humans communicate with each other using language. Linguistic communication is a very complex phenomenon, which involves many different aspects of the mind. At a certain level of abstraction however, this phenomenon may be, and has often been characterized as *the transmission of content* from speaker to hearer. According to this characterization, when a speaker produces an utterance, she expresses a certain thought to the audience; communication is successful when, on the basis of the utterance, the hearer comes to entertain the very same thought that the speaker thereby expressed. Call this, the *Naïve Conception of Communication* (see Figure 1.1) .[1]

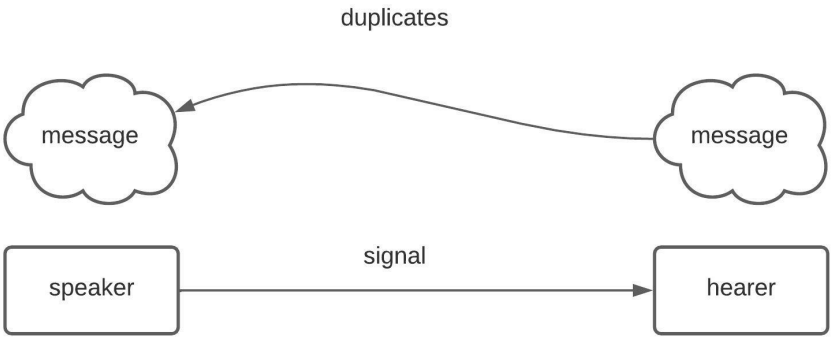


Figure 1.1 – The naïve conception of communication

This abstract philosophical model of communication is intuitive. It seems to derive from our pre-theoretical understanding of the phenomenon. While it may be that there is no such thing as a well-defined folk concept of successful communication that could be the target of conceptual analysis, it is uncontroversial that people generally have intuitions of success or failure about communicative events. Such intuitions lead people to make statements about cases. For example, people say things like "They misunderstood who we were talking about", "Gotcha!". Speakers generally have intuitions about the correctness of such utterances against various scenarios (I do not mean to deny that some cases might appear inderterminate with respect to communicative success or failure).

These folk intuitions about communicative success are data when theorizing about communication. I believe such data should constrain any philosophical/scientific analysis of the phenomenon. Philosophers ought to provide a systematic account of the folk standards by which communicative success and failure are routinely judged by speakers. Accordingly, any philosophical criterion of successful communication should line up with the folk demarcation

[1] This label is from Heck 2002, used also in Recanati 2016. The same conception has received other names in the literature: the 'conduit metaphor' (Reddy 1979), the 'belief transfer model' (Egan 2007), 'the package delivery model' (Moss 2012) or 'the FedEx Model' (Weber 2013), see note 16 in Recanati 2016.

of communicative success. More precisely, any philosophical definition of successful communication should roughly deliver the same verdicts about cases than the folk standards underlying the pre-theoretical practice of assessing communicative success.

One might object to this idea that it wrongly assumes the folk standards for assessing communication are good as they stand. But these standards might be imperfect and in need to be revised. I think there is something odd about considering this as a possibility, for the following reason. Human verbal communication, unlike (say) physics or biology, *does involve* our common understanding of semantic/pragmatic facts, including our notion of when communication is successful, and when it fails. For instance, the belief that a given semantic hypothesis is shared by one's interlocutor plays an essential role in the interpretation and production of utterances. Common sense semantic/pragmatic hypotheses, including the folk standards of successful communication, are just part of the phenomenon. In other words, the concept of human verbal communication *is* essentially a folk concept. Hence I will assume that the goal of analyzing and systematizing the folk standards about successful communication is fruitful.[2]

Implicit in utterances such as "gotcha!' is, arguably, the idea that communicative success demands a match in content between the thought expressed by the speaker and the thought the hearer comes to entertain as a result of understanding the utterance. The notion of 'content' is a technical term in linguistics. Plausibly, this notion is involved as a mere *placeholder* in the naïve conception: as whatever it is the sharing of which helps explain communicative success. Communicative success thus might allow us to better understand the notion of content: we may define content as that which must be grasped for communicative success, and, studying communication, discover what it is. [3]

As a first pass towards making the *Naïve Conception* more precise, let us ask what is the relevant notion of content, the sharing of which *putatively explains* communicative success. Linguistic communication enables speakers to reveal the propositional contents of their thoughts to their audience. One might expect that what is asserted in singular communication just are singular propositions. For instance, imagine that it is common ground that we are talking about my little sister Déborah, and I tell you "she is an osteopath". What I communicate thereby is the singular thought[4] composed of my sister Déborah and the property of being an osteopath. You understand me, we might be tempted to say, if and only if you grasp the singular proposition

[2] I don't mean to deny that there might be more than one notion of communicative success worth studying. I will say more about revisionary attitudes about the folk concept of communication in due course.

[3] See the' availability based approach' in Recanati 2002 (1.6) where a similar methodology is discussed at length. Recanati 2002 defines what is said as:

> The primary truth-evaluable representation made available to the subject (at the personal level) as a result of processing the sentence.

[4] In the space of this chapter I use 'proposition' and 'thought' interchangeably. For a working definition of singularity of thought, see the main introduction of the dissertation.

that I intended to express.

1.3 Loar cases

But in fact, things are not so easy. As Brian Loar (1976) showed, assigning the right referent to a singular term is not always sufficient for referential communication to be successful. One can fail to understand a singular term while correctly identifying the referent associated with that singular term. Here is a variation on Loar's example inspired by Buchanan's (2013) (depicted in Figure 1.2):

> KASPAROV: Anna and Bob are sitting on a bench in Central Park, as Anna is reading the *New York Times*. The front page features a large photo of Garry Kasparov to which Anna gestures and says 'that man is smart'. Bob takes Anna to be intending to refer to a certain man sitting in front of them in the park, reading a book, who happens to be Garry Kasparov. Intuitively, even thought Bob has produced an interpretation that correctly identifies to whom Anna is referring with 'that man', Bob has misunderstood Anna's utterance.

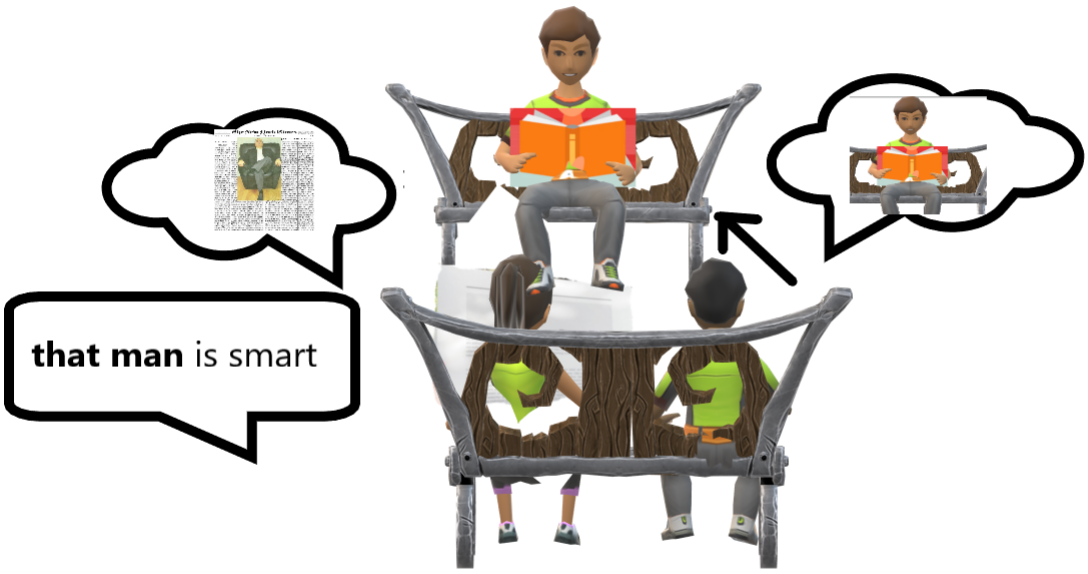


Figure 1.2 – A Loar case

In this example, Anna intends to communicate of a certain man that he is smart, and Bob

entertains the intended proposition on the basis of Anna's utterance. However, the example elicits the intuition that communication has failed. Why? One element which I think triggers the intuition of communication failure in this case is that it is merely a matter of *luck* that Anna and Bob are entertaining the same singular proposition. For it is fortuitous that the man sitting in front of Anna and Bob in the park is the man Anna is talking about.

In what follows, I will take it that Loar cases are characterized by the element of *epistemic luck* that they contain. In a Loar case, the interpretation of the hearer ascribes the same property to the same object as the thought expressed by the speaker, but the fact that both thoughts corefer is a matter of luck. In this respect, the Loar cases are with respect to communication what *Gettier cases* are with respect to knowledge. Gettier cases are cases in which justified true beliefs fail to constitute knowledge because of an element of luck in how the beliefs manage to be true and justified.[5] Just as Gettier cases are cases in which an element of luck prevents the episode from being one of knowledge, Loar cases are cases in which an element of luck prevents the interpretation of the hearer to constitute genuine understanding of the speaker's utterance.

1.4 Communication and knowledge

I have just presented the Loar cases as if they were to communication what the Gettier cases are to knowledge, that is, as if there was a mere relation of *analogy* between communication and knowledge.[6] However, on reflection, the relation between Gettier cases and Loar cases might be more intimate than a mere analogy. To see why, consider the role of linguistic communication in testimony. Testimony is the procedure by which one acquires knowledge through endorsing the content that one recovers through a communicative exchange. For example, I ask my friend what the capital of Ecuador is. She tells me "Quito is the capital city of Ecuador". I understands what she says, I trust her, and I thereby acquire the knowledge that Quito is the capital city of Ecuador. My reason for believing that Quito is the capital of Ecuador is that my friend said this to me. So it seems that successful communication can transmit knowledge from speaker to hearer.

Here are two questions to get started:

(Q1) What is the epistemic relation that a hearer must bear to the content expressed by the speaker's utterance in order to be able to get knowledge from that utterance?

(Q2) Can successful communication be lucky, given that successful communication can transmit knowledge?

The answer I will provide to the first question also answers the second: if the relation that

[5] Gettier 1963 is the *locus classicus*. Pritchard (2005) is a monography on epistemic luck. Pritchard (2007) is an overview of the anti-luck approach to the definition of knowledge. Williamson 2000 argues that knowledge is a primitive notion.

[6] See Bach (2006: 524) for the claim that Loar cases are analogous to Gettier cases.

a hearer must bear to what is said in order to get testimonial knowledge is *knowledge*, then recovering what is said cannot be lucky if knowledge cannot be lucky. I shall explain.

The fact that knowledge of what is said is required to get testimonial knowledge would explain why communication cannot be lucky, given that knowledge cannot be lucky. Seen in this way, the Loar cases would simply be Gettier cases. Here is an argument to bring this point home:

Argument from testimony to the effect that successful communication cannot be lucky (ALA)

(1) Successful communication transmits knowledge in the following sense: if S knows that p at t and S says that p to H at t and communication is successful between S and H at t and H trusts S at t, then H can know that p from S at t. (Testimony is a source of knowledge)

(2) In order for H to come to know p on the basis of understanding p as what S said at t, H must know that S said that p at t.

(3) In order for H to understand what S said at t, H must know what S said at t. (Corollary of (2))

(4) Knowledge cannot be lucky.

(5) In particular, knowledge of what is said cannot be lucky. (By 4)

(6) Therefore, successful communication/understanding cannot be lucky. (By (5) and (3))

A few comments are in order. I take it that premiss 1 is obvious, for most of our knowledge is obtained on the basis of linguistic interactions with others (including the interpretation of written utterances). Someone who denies it would be committed to some form of radical scepticism.

Premiss 2 is the crucial one in this argument. By "to come to know p on the basis of understanding p as what was said", I mean to rule out propositions known in a way that depends on a speaker's testimony but that are intuitively *not* testimonially acquired. For example, I may come to know *that you speak French* (call this proposition q) if you say somethink in French (no matter what it is) that I understand. However, assuming you said something other than q, q will not be understood as what you said. Likewise, I may come to know *that you have a radiophonic voice* (call this proposition r) if you say something to me, but I will not have learnt r on the basis of understanding it as what you said. Even if what you said is *that you have a radiophonic voice*, I need not come to know it on the basis of understanding r as what you said (Peet 2019 following Goldberg 2007).

Moreover, in "understanding p as what was said", I am assuming that p is not an implicature, but instead graspable somehow *directly*. By this I don't mean to exclude that *understanding*

p as what was said may be the result of an inferential process. First of all, there is a liberal sense of "inference" that counts unconscious and sub-personal transitions between thoughts as inferences. Obviously, grasp of what is said could be inferential in this sense. For instance, your grasp of *p* as what the speaker said may result from an inference from the proposition that the speaker has uttered sentence *s*. Grasp of what is said may even result from some inferential process in a more narrow sense, on which inferences may be conscious albeit spontaneous and automatic (Recanati 2002). This issue, though interesting, need not concern me for present purposes. What I do want to exclude from the etiology of "understanding a proposition as what was said" are the *explicit* kind of inferences, from which implicatures result. (See also 1.6 below).

Lastly, when using "coming to know *p* on the basis of understanding *p* as what *S* said", I don't mean to preclude that the thought of the hearer standing for *p* might be (for at most some notion of content finer-grained than referential content[7]) in a relation *weaker than identity* to the thought expressed by the speaker. Let me clarify what I mean. The issue whether successful understanding/communication involves content *identity* (other than referential content, that is) is precisely the issue I am investigating in the chapter. So I want to leave it open that an audience may come to know that *p* on the basis of understanding *p* as what the speaker said, although the thought of the hearer and the thought of the speaker do not match in their fine-grained content. I now turn to my justification in support of the crucial premiss 2.

For *reductio*, assume that understanding/successful linguistic communication does *not* require knowledge of what is said. For example, assume that understanding an utterance to the effect that *p* requires no more than the mere justified (and presumably tacit) belief that *p* is what was said.[8]

First of all, note that this view need not be committed to the idea that people should have explicit thoughts about what is said. Instead, people may have tacit beliefs about what is said just by interpreting the utterance they hear, provided they experience the thought they entertain as the interpretation of the utterance they hear (as I have tried to elucidate this notion in previous paragraphs). Understanding does not have to be joined with a second-order attitude about what was said. (Of course, a similar consideration applies to knowing what is said).

So a person *H* could be justified in believing that *p* is what *S* said at *t*, and in forming the belief that *p* on that basis (assuming the speaker is trustworthy and reliable), despite the fact that *p* is not in fact what was said. This is because it is possible for a person to be justified in believing a proposition that is in fact false. But it does not even matter whether the speaker is trustworthy or reliable, if the content recovered is not what was said!

[7] I am dealing with referential communication.

[8] I assume that epistemic justification is non-factive, as most epistemologists do.

Now, the line of thought just deployed does not yet establish that knowledge of what is said is required for understanding. Rather, it establishes at most that a *true* justified belief about what is said is required for understanding. But, as we know since Gettier, a justified true belief does not amount to knowledge. So we need a further argument for the claim that knowledge of what is said is necessary for understanding/successful communication.

Such an argument is not hard to find. Remember the Loar case: despite the fact that the hearer had a true justified belief regarding the proposition expressed by the speaker, the hearer could not be said to know what was said. (Importantly, in section 3, I argue that we can construct a Loar case for any notion of content one chooses as the proposition expressed by the speaker). This is because an element of luck was in play in how the interpretation of the hearer managed to be true and justified. Therefore, the best explanation of this intuition is, I submit, that *knowledge* of what is said is required for understanding/successful communication. To recap this line of thought:

Argument for the claim that knowledge of what is said is necessary for understanding (P3 of ALA)

> **(1)** In the Loar case, the fact that the interpretation process is lucky is deemed to make communication fail.
>
> **(2)** The best explanation of this intuition is that knowledge of what is said is required for understanding/successful communication.
>
> **(3)** Therefore, (we have reasons to believe that) knowledge of what is said is required for understanding/successful communication

The upshot of this argument is that for referential communication to the effect that p to succeed between persons S and H at t, it is not enough that H be justified in believing that p was said by S at t: H must *know* that p is what was said.

Premiss 1 may be read as a psychological premise (i.e. about what people find intuitive when exposed to the Loar cases). Read in this way, it is an empirical claim. Machery and al. (2015) provide empirical evidence that intuitions on Gettier cases do not vary across cultures. One prediction attached to the claim that genuine understanding requires knowing what was said is the following: intuitions to the effect that communication does not succeed in the Loar case will likewise not be prone to cultural variations.[9]

[9] I mean the original Loar case, where participants have false identity belief about the object under discussion. In the next subsection, I present cases having the structure of the Loar case where communication seems nevertheless fine. I argue that we accommodate a certain degree of communicative failure when the context (linguistic context, purpose of the conversation, etc) is favorable. This is orthogonal to the cross-cultural prediction I am mentionning.

But this by itself does not imply that there are *no* case of correct but lucky interpretation that could be deemed knowledge transmitting. Let name **LC** (*x*,*p*) the relation according to which a hearer *x* stands to the proposition *p* as in the original Loar case described in Loar 1976.[10] (For example, my own variant is a Loar case to the extent that it instantiates the relation **LC** (*x*,*p*)). In fact, the relation **LC** (*x*,*p*) may be instantiated in many different ways. Some of those possible instantiations, as we shall see in a minute, are such that they render *false* the claim that communication is deemed to fail or that the hearer failed to understand whenever **LC** is instantiated (I present some of them in the next subsection).

Relatedly, we can anticipate such a possibility by considering an epistemic contextualist take on knowledge of what is said.[11] Epistemic contextualism is, roughly, the view that what is expressed by a knowledge attribution to the effect that S "knows" that *p*, depends partly on factors in the context of the attributor[12]. Which factors of the context of the attributor (including the speech participants assessing *on-line* whether the conversation is going well) might *lower* or *increase* the standards for testimonial knowledge/communicative success attribution? I present some of them in the nextsubsection.

1.5 Purpose of the conversation, linguistic context, common ground mentalizing & communicative luck

Here I will deal with apparent counterexamples to the claim that successful communication cannot be lucky, and bring further factors relevant to communicative success (and communicative success attributions) on the table, namely, the purposes of the conversation, the intra-discourse linguistic context, and a specific class of (meta-)representational states of the speech participants. This subsection brings to bear these factors on the issue of the relation of communication and luck.

Unnsteisson 2018 makes the observation that Loar cases in which the participants do not have false identity beliefs about the intended referent seem *not* to exhibit the same degree of communicative failure as Loar cases in which they are in a contextually relevant Frege case with respect to the object under discussion (i.e. like in the original Loar case). For instance, consider a case which is in every respect like KASPAROV (the Loar case I have presented in the introduction), except that it is common ground between Anna and Bob that the man in front of them is Kasparov. Call it KASPAROV(b):

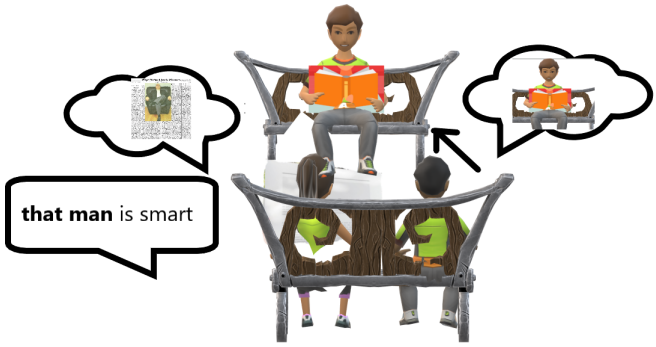
> KASPAROV(b): in every respects like KASPAROV except that it is common ground between Anna and Bob that the man in front of them is Kasparov.

What would happen if Anna were to realize that Bob has interpreted her utterance by looking

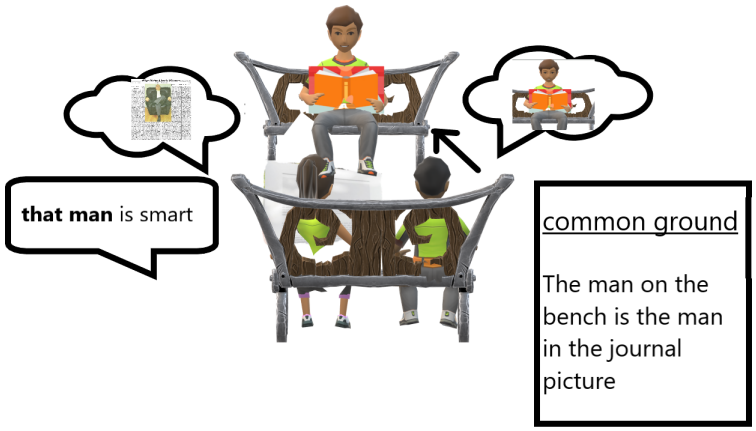
[10] "LC" stands for "Loar Case".
[11] I defend such a view in the chapter 3.
[12] See Rysiew (2021) for an overview.

at the guy in front of them and not - as she intended him to do - by looking at the newspaper she is gesturing towards? Presumably, if Anna knows that Kasparov is the man in front of them, and knows Bob knows it as well, and knows he knows she knows etc., then Anna can be more relaxed about Bob's mistake, were she to realize it. Indeed, we can sensibly imagine she would not even bother correcting Bob ('who cares, that's the same guy!').

By contrast if, as in the original KASPAROV case, Anna does not believe the man in front of them is Kasparov, then it is clear that she *ought* to be more preoccupied by Bob's mistake, were she to realize it. The contrast between the two cases is depicted in Figure 1.3.



(a) BAD



(b) GOOD

Figure 1.3 – Loar Case with (a) **vs.** without (b) false identity belief

Importantly, in KASPAROV (b), there still seems to be an element of luck in the interpretation of the hearer: Anna intended Bob to look at the *New York Times* picture, not the man in front

of them, as Bob did. However, the fact that it is common ground that the man in the photo is the man in front of them seems to somehow make up for the element of luck occurring in the episode. Indeed we do not have the intuition that communication fail in KASPAROV (b).

One might be tempted to conclude from KASPAROV(b) that communication may succeed while being lucky to some degree. Unnsteisson (2018) extrapolates from the contemplated contrast that the absence of false identity beliefs about the object under discussion may be a critical factor in communicative success. On this construal, in a Loar case without false identity belief, communication would be successful while still being lucky to some degree. The contemplated contrast in Unnsteinsson should thus shift the attention of the communication theorist from the *luck* factor to the criterion of *absence of false identity beliefs*.

First, observe that having specific *metarepresentational* false beliefs about identity can be just as damaging for communicative success as the presence of false first-order identity beliefs. Contrast KASPAROV (b) with the following scenario (call it KASPAROV(c)) :

> KASPAROV(c): in every respects like KASPAROV(b) except that Anna believes Bob does not know the man in front of them is Kasparov.

To see why communication is not fine in KASPAROV (c), imagine that Anna realizes Bob's mistake, namely, that he interprets her utterance by looking at the guy in front of them. Since Anna believes that Bob does not know the relevant identity, Bob's interpretation is *lucky* from the perspective of Anna, even though Bob in fact knows that his interpretation corefers with the thought Anna wants to express. We might say that Anna has a false *metarepresentational* identity belief concerning Bob, and that is where the breakdown in communication lies. From Anna's perspective, there is a competing object *she worries Bob might construe as a 'competing referent'*. Given Anna's take on Bob's perspective on the target, Anna ought to tell Bob about the relevant identity to prevent a possible misunderstanding.[13] The contrast between KASPAROV(b) and KASPAROV(c) shows that having specific metarepresentational false beliefs about a relevant identity as it is represented in one's interlocutor perspective can make communication fail just as well as the presence of false first-order identity beliefs.[14]

Now, is it true that the absence of false (first-order or metarepresentational) beliefs about relevant identity allows that successful communication may be compatible with some degree of luck, as my (possibly faulty but not aimed exegetically) construal of Unnsteisson's diagnosis

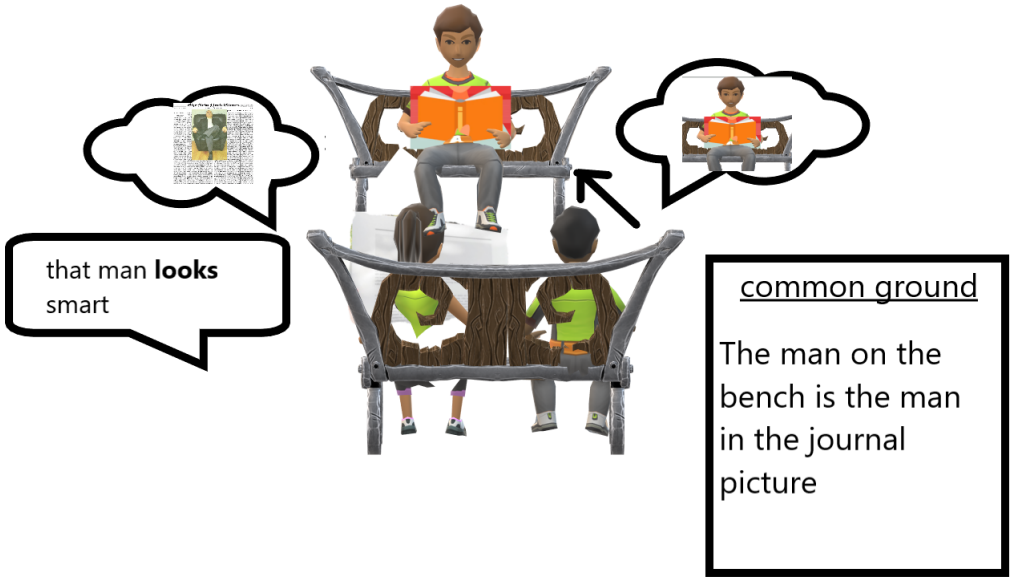
[13] In passing, I note that if my interpretation of KASPAROV(c) is correct, it constitutes a counterexample to Onofri 2018's criterion (examined in chapter 3): Anna knows that the thought she expresses corefers with the thought Bob entertains when interpreting her utterance; and Bob knows that his interpretation corefers with the thought expressed by Anna. However, because Anna does not know that Bob knows that her thought and his interpretation corefer, communication seems faulty here: it seems that Anna ought to warn Bob about the identity.

[14] The notion of common ground is useful here. See Schiffer (1972), Stalnaker (2002; 2014), Bach and Harnish (1979), and Clark (1992; 1996). Two standard definitions of the common ground as common belief and common acceptance are provided in section 3 of the main introduction of the dissertation, based on Stalnaker 2002.

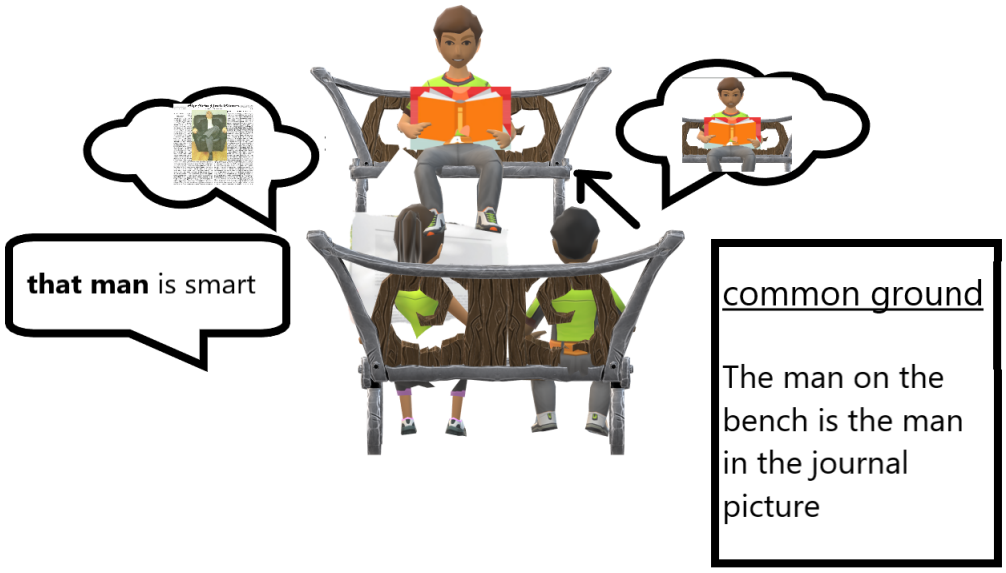
of the contrast at issue suggests? I believe that Unnsteisson 2018 is making an interesting observation, however I do not share the generalization he (at any rate, as I interpret him) draws from it. I think a better way to construe Unnsteisson's observation is to say that, given certain relevant factors, we are quite happy to tolerate some degree of communicative failure in some cases. The true upshot of the contrast at issue is that certain pivotal features such as the *purposes of the conversation* may impact (i.e. lower or increase) the standards for communicative success attribution. In other words, depending on the purposes of the conversation, we may *tolerate* a certain amount of communicative *failure*. For example, this is (I submit) what happens in KASPAROV(b).

Here is another class of factors which can trump that of the absence of false identity belief as to communicative success attribution. Unnsteisson's generalization regarding the absence of false identity belief as criterial for successful communication will prove too quick for yet another reason: communication can go wrong *Loar case style* even though the speech participants do not harbor false identity belief about the referent at issue and correctly represent the absence of relevant false identity belief in the perspective of the other. Indeed, in some conversational contexts, *the linguistic context* requires (hyper)intensional discrimination even when it's common knowledge that no participant is in a relevant Frege case about the targeted referent. In other words, the conversational context may be (hyper)intensional (i.e. such that necessarily co-intensional (co-extensional, resp.) representations are not substitutable *salva veritate* in the discourse context) even if the speech participants are *not* relevantly confused with respect to the object under discussion. This detail seems to be overlooked by Unnsteinsson 2018. Let us minimally change my previous example KASPAROV(b) to make the point vivid. Short version: SMARTLOOKING is in every respects like KASPAROV(b) except that Anna says "That man looks smart". (see Figure 1.4):

> SMART-LOOKING: Anna and Bob are sitting on a bench in Central Park, as Anna is reading the *New York Times*. It is common ground between them that the man in front of them is Garry Kasparov (perhaps they saw together a crowd of journalists around him). The front page of Anna's newspaper features a large photo of Garry Kasparov to which Anna gestures and says 'that man **looks** smart'. Bob takes Anna to be intending to refer to the man sitting in front of them in the park, reading a book. Intuitively, even though Bob has produced an interpretation that correctly identifies to whom Anna is referring with 'that man', and both Anna and Bob know that the man in front of them is the man on the front page of the newspaper, know that the other know it, etc., Bob has misunderstood Anna's utterance.



(a) BAD



(b) GOOD

Figure 1.4 – SMART-LOOKING (a) vs. KASPAROV (b)

Here Bob misunderstands Anna's utterance even though Bob and Anna do not have contextually relevant false identity beliefs or false metarepresentational identity belief about the targeted referent. This is because the conversational context seems hyperintensional: a man

may look smart under *some* mode of presentation but not under others. A speaker using "x looks smart" may want to draw the attention of their audience to a particular smart-looking feature of x, associated with a particular mode of presentation of the object. Bob may have falsely thought Anna said what she said because of the book, for instance. But Anna did not have the book in mind when she said what she said.

To conclude: it might look as if communication can be lucky and yet succeed in some cases, in particular, in cases that are just like Loar cases except that the participants to not have false identity beliefs. However, I think a better way to look at such cases is to say that *the purpose of a conversation* and *the linguistic context*, given the participant metarepresentational perspective on the perspective of the other on the intended referent, are pivotal contextual features in light of which we may tolerate a certain degree of luck. Good enough communication is very generally imperfect.[15]

1.6 Which notion of content is at stake?

Before we begin, let me clarify the notion of content at issue. There are different things one might mean by 'content' in linguistics and the philosophy of language. One notion of content corresponds to the purely semantic compositionally determined meaning associated with an utterance *qua* sentence-type. In the literature, such a notion is often called 'character' or 'standing meaning'. Intuitively, this amounts to the linguistic meaning of an utterance out of context. To illustrate, let us take my toy example of sect.1.2 again:

(1) She is an osteopath.

If I am a competent English speaker and I overhear "she is an osteopath" while lacking the relevant contextual details, what I grasp is the content that some salient female individual is an osteopath. Accordingly, communication may fail at that very level. For instance, that will be so if I believe that 'osteopath' means, say, *psychopath*.

Another notion of content corresponds to the standing meaning attached to an utterance *relative to context*. Intuitively, this notion amounts to *what is said*. In my example, the utterance context is such that "she" in fact refers to my little sister Déborah. Similarly, communication may thus fail at that very level too. For instance, that will be so if my interlocutor believes that I intend to refer to the women in front of us (and not Déborah) with "she".[16]

[15] In chapter 3, I reconsider the matters, and say that the standards for *knowing* what is said *are themselves* contextual.

[16] I am presenting a very standard Kaplanian picture. This picture is a rough approximation, which could be refined in many ways. In particular, as already noted, it is plausible that pragmatics (e.g. modulation/enrichment) operates not merely at the level of implicatures, but already at the level of what is said. That is, grasp of what is said may involve some inferential process, albeit not *explicit* inferences. This issue, while interesting, need not concern me in this chapter. I refer the reader to Recanati 2002 for a discussion of the standard picture. See also Carston 2008.

Finally, there is a broader notion of content, one that encompasses so-called conversational implicatures. Conversational implicatures are things the speaker conveys albeit indirectly, in a way that is not systematically constrained by the conventional meaning of the sentence uttered, but that depends on features of the conversational context. For instance, consider the following dialogue:

(2) **You:** What does Déborah think of MOP-involving propositions?

Me: She is an osteopath.

By saying "She is an osteopath", I am conversationally implicating that my little sister has no opinion whatsoever on MOP-involving propositions. Accordingly, communication may fail at this level, for instance if you fail to grasp what I am conversationally implicating, even though you have successfully grasped the standing meaning and what was said.

In this chapter, I will be concerned with the level of content of *what is said*. Hence, by *communicative success*, what I mean has to do with the successful grasp of what is said. *Is what is said somehow finer-grained than referential content? How does what is said relate to the content of the mental states of the speech participants? Can we decide whether a hearer has grasped what is said, solely on the basis of the individual psychological states of the participants?* These are guiding questions for the chapter.

2 The Standard Fregean conception

I have presented the Loar case as showing that grasping the correct referential content of an utterance is not enough for understanding/successful communication to occur. One reaction to the Loar case – which I call *Fregean*, because Frege (1892) is well known to have observed that representations sharing referential content might still differ 'content-wise' – is to claim that the contents which must be grasped for communicative success are *finer-grained* than referential content. Assuming that the relevant notion of content is *that which must be grasped for communicative success*, the idea is thus to eliminate luck by making the relation of *same-content* more demanding.[17]

We may think of these finer-grained contents as *ways of thinking*, or *modes of presentation* of the referents (MOPs hereafter). On the Standard Fregean view, MOPs are taken to be *descriptions*.

[17] What does it mean to say that a type of content is "finer-grained" than another type of content? Abstractly, we may think of any *same-content* relation as an equivalence relation for grouping representation tokens into subsets in terms of their content. Given a set of representation tokens, the *same-content* relation thus determines a partition of that set, that is, a disjoint union of non-empty subsets—the "parts" of the partition—in such a way that every representation token is included in one and only one subset. The parts of such a partition are the different contents as individuated by the relevant equivalence relation. That a given *same-content* relation $\sim_C$ is finer-grained than the *same-referential-content* relation means that $\sim_C$ splits the parts of the referential partition into smaller parts. See also section 2 of the general introduction.

To illustrate, remember KASPAROV – the Loar case I have presented in the introduction (see also Figure 1.2):

> KASPAROV: Anna and Bob are sitting on a bench in Central Park, as Anna is reading the *New York Times*. The front page features a large photo of Garry Kasparov to which Anna gestures and says 'that man is smart '. Bob takes Anna to be intending to refer to a certain man sitting in front of them in the park, reading a book, who happens to be Garry Kasparov. Intuitively, even thought Bob has produced an interpretation that correctly identifies to whom Anna is referring with 'that man', Bob has misunderstood Anna's utterance.

The Fregean diagnosis goes as follows. Although Bob and Anna both think to the same man, communication fails because, whereas Anna thinks of the referent as *the man depicted on the front page of the journal*, Bob thinks of the referent as *the man on the bench* (or something like this). But the purpose of the conversation, given the epistemic perspective of the participants on the target, require that Bob think of the referent in the same way as Anna, namely (on a Standard Fregean way to construe MOPs), according to a description such as *the man depicted on the front page of the journal* (or something in the vicinity).

The Fregean picture of communication is very intutive, and the Loar case seems to cry out for it. Loar's own diagnosis is a good preliminary characterization of the Fregean reaction:

> "It would seem that, as Frege held, some 'manner of presentation' of the referent is, even on referential uses, essential to *what* is being communicated" (Loar 1976 p. 357, my italics.)

What the Loar case shows, through the Fregean lens, is that what is said is richer than mere singular propositions. The Fregean will claim that, in the aforementioned case, Anna has communicated a proposition about Kasparov with an instruction to think of him as the guy depicted on the journal (or what have you). The proposition communicated by Anna might thus be represented like this:

Descriptive MOP-enriched proposition (*) [The x: x=Kasparov & *Man* (x) & $F(x)$] (*Smart* (x))

where 'F' expresses the property of *being depicted on the front page of the journal*, or something like this.[18]

In general, the Standard Fregean will say that *what is said* includes some contextually relevant property such as F, namely the descriptive MOP under which the audience is meant to think of the referent in order for communication to succeed. My representation of the enriched

[18] For the sake of simplicity, I leave aside the context-sensitivity of some consitutents in the description.

proposition putatively communicated by Anna as (*) is controversial. It is merely meant as an illustration of the Fregean conception. Even leaving aside the question whether the nominal (*Man*) to which the demonstrative "that" is conjoined contribute to the content of the complex demonstrative,[19] the question arises as to whether the putatively communicated descriptive MOP (here expressed with the predicate *F*) should appear in the truth-conditions of the enriched singular proposition.[20] One salient option among Fregeans is to say that the putatively communicated MOP merely place a constraint on how the audience must think of the referent, but is otherwise truth-conditionally irrelevant (Recanati 1993, Carpintero 2000).[21]

Relatedly, Standard Fregeans might disagree amongst each other with respect to how we should best think of these MOPs. For instance, once they have agreed that the content that must be grasped for communicative success is richer than a singular proposition, they might disagree as to whether the hearer should think of the referent under a MOP exactly *identical* to the one that the speaker 'had in mind' in producing the utterance; alternatively, they might allow that the MOP of the hearer could be merely suitably *similar*. I will not discuss such issues at this stage. Instead, I would like to argue against the Standard Fregean conception of singular communication. In the next section, I want to show that the proposal according to which the content that must be grasped for communicative success includes descriptive MOPs is still compatible with a Loar case, and consequently, does not guarantee communicative success.

3 A problem for the Standard Fregean conception

3.1 Identity of MOPs is gettierizable: the threat of Super-Loar cases

My main argument against the Standard Fregean view of successful communication is that Loar cases can be provided in which the speech participants think the same content under the same descriptive MOPs.[22] These iterations of Loar cases for finer-grained conceptions of shared content, I call *Super-Loar cases*. Consider the following case (see Figure 1.5):

[19] Minimal theories assign no semantic role to common noun phrases in complex demonstratives. Maximal theories say that the nominal helps determine the referent and its content appears as a constituent of the content, in a context (See Braun 2017 for an overview). This issue need not concern me here.

[20] See Soames 2002, 2008, for a not purely pragmatic strategy in terms of descriptively enriched singular propositions.

[21] I discuss this view in section 4 as one of the tenets of the Sophisticated Fregean view.

[22] The discussion in this section was inspired by Tayebi 2013.

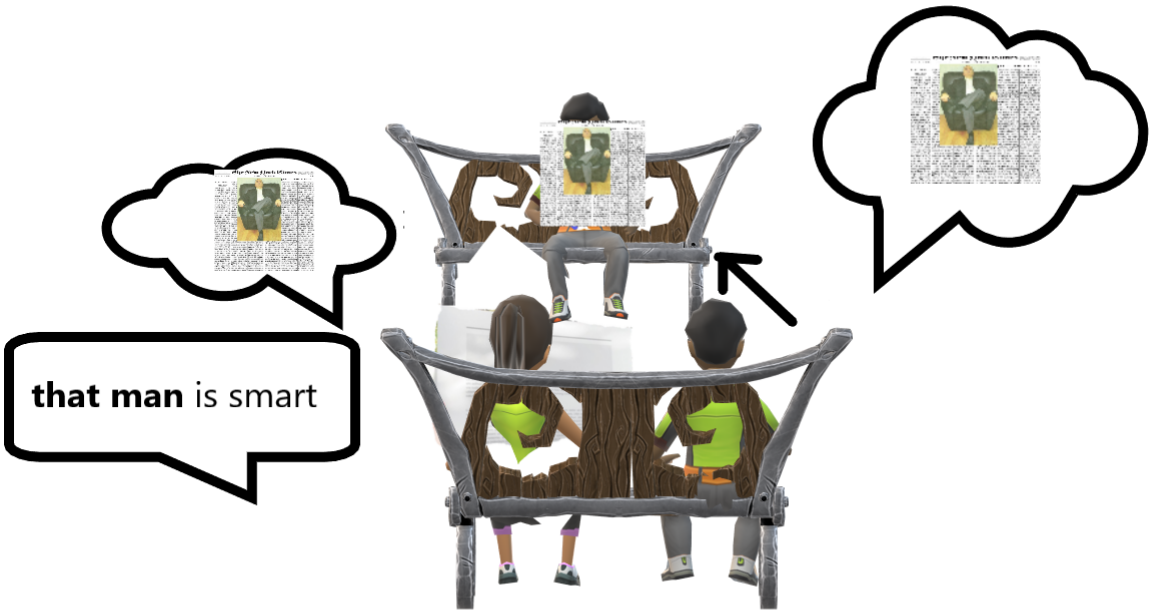


Figure 1.5 – Super-Loar case 1

> SUPER-LOAR CASE 1: Anna and Bob are sitting on a bench in Central Park, as Anna is reading the *New York Times*. The front page features a large photo of Garry Kasparov to which Anna gestures and says 'that man is smart '. Bob takes Anna to be intending to refer to the man depicted on the front page of the newspaper the man in front of them is reading. As it happens, this is also the *New York Times*.

In this scenario, Bob and Anna share the same content under the same descriptive MOP (Kasparov is thought as *the man depicted in the front page of the* New York Times), yet communication does not seem successful. The newspaper involved in Bob's interpretation is the newspaper the man in front of them is reading, which could have been other than the *New York Time*, e.g. featuring another object on the front page. So it is a matter of luck that Bob's interpretation is coreferential with the thought Anna expressed with her utterance.

A Fregean may object that Bob and Anna are not really sharing the same content under the same descriptive mode of presentation in SUPER-LOAR CASE 1. For Anna wanted Bob to look at *her specimen* of the New York Times, or something in the vicinity. The objection insists that the relevant MOP here should be expressed with *The man depicted in the front page of the New York Time Anna is gesturing towards*, or something like this.

As a rejoinder, here is a scenario (see Figure 1.6) in which Anna and Bob both share the relevant MOP, yet communication fails. I borrow the structure of this case from Tayebi 2013, who is more generally a strong influence here:[23]

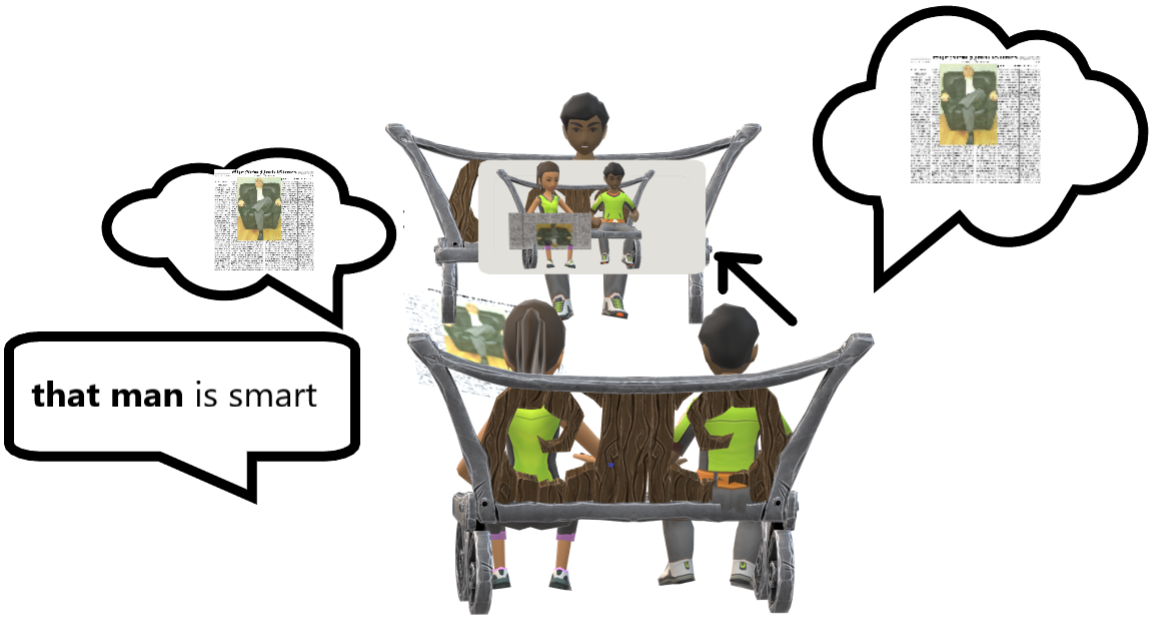


Figure 1.6 – Super-Loar case 2

> SUPER-LOAR CASE 2: Anna and Bob are sitting on a bench in Central Park, as Anna is reading the *New York Times*. The front page features a large photo of Garry Kasparov to which Anna gestures and says 'that man is smart '. Bob does not realize Anna is sitting next to him on the park bench. Unbeknown to Bob, the person sitting in front of him holds a mirror in the direction of Anna and Bob. Bob looks at Anna in the mirror but does not realize he is looking at a mirror. In the mirror, he sees what he thinks is another person than himself, sitting next to Anna, and to whom Anna is speaking. He thinks he overhears what Anna is saying to "that guy". He interprets Anna as talking about the man depicted on the front page of the newspaper Anna is gesturing towards.

In this scenario, both Anna and Bob think of the same referent under the same descriptive MOP, namely something like *the man in the picture of the New York Times Anna is gesturing towards*.

[23] I discuss Tayebi (2013)'s case in the next chapter.

However, miscommunication occurs. Where is the breakdown in communication between Anna and Bob? By hypothesis, the audience is required to think of Kasparov as *the man in the New York Times picture Anna is gesturing towards*. But that is what Bob is doing! However, while Bob entertains the right thought on the basis of Anna's utterance, Bob does not realize that Anna is in fact talking to *him*. So communication fails.[24]

3.2 Diagnosis

The conjecture I would like to put forward is that, for any possible communicative exchange, it is possible to conceive a Super-Loar case for any descriptive MOP that is the most plausible candidate as to how the audience must think of the referent in that communicative exchange. Let me explain the rationale for this conjecture.

Let us call **SLC**(*x*,*p*) the relation that obtains in a communicative scenario between an audience *x* and a proposition *p* just in case the audience *x* share with the speaker the same content under the same descriptive mode of presentation, yet communication fails in that scenario, as in the SUPER-LOAR CASE 1 or 2. Super Loar cases arise for the Standard Fregean conception because it is always possible that the audience think of the right referent under the right descriptive mode of presentation, albeit in a deviant or lucky way.

If this diagnosis is correct, then making the relation of *understanding* more stringent by requiring coordination on finer-grained contents such as descriptive-MOPs-enriched singular propositions is not an antidote against the Loar cases. One thing that the Super-Loar cases teach us is therefore that the true upshot of the Loar case is not that communicated contents are fine-grained, but that there are *non-semantic conditions* for successful communication having to do with the fact that the causal transmission chain should be non-lucky and non-deviant.

3.3 A fix for the Fregean conception: the "two-factor" Fregean theory

I have argued that the Standard Fregean conception was not immune to the Loar cases because it is always possible that the audience share with the speaker the same content under the right mode of presentation but in a lucky manner or as a result of a deviant causal chain. Consequently, I am suggesting that whatever the notion of communicated content one chooses, recovering the right content is never sufficient for communication. A causal, non-semantic condition on the interpretation process must also obtain, namely the interpretation process must be non-lucky and non-deviant.

In fact, this suggestion has its place in a Fregean theory of communication. Fregeans could grant that, because sameness of descriptive MOPs is gettierizable (that is, may arrived at by mere

[24] Replicability matters, even for thought experiments, because by replicating structurally similar counterexamples, we can be more certain that the counterexample is robust. I leave it to my reader to find other Super-Loar cases as an (optional) exercise.

luck or as a result of a deviant causal chain), sameness of descriptive MOPs is not *sufficient* for communication to be successful. But Fregeans could insist that sameness of descriptive MOPs is still a *necessary* condition on successful communication. *In addition*, they should conceded that the interpretation must result from *a non-deviant causal chain* – a non-semantic condition on successful communication. For the record:

The Two-factor Fregean view: A speaker S successfully communicates that p to a hearer H iff:

> **(a)** S utters s to communicate some proposition and H understands S to be communicating some proposition with s.
>
> **(b)** S and H think the same content under the same descriptive MOP.
>
> **(c)** (a) and (b) are not satisfied by luck or in virtue of a deviant causal chain.

Condition (a) is meant to be a necessary condition for successful communciation everyone should accept. (b) and (c) constitue the two factors (semantic and causal, respectively) necessary for successful communication according to the view. The "Two-factor" Fregean theory seems better than the standard Fregean view. It can account for intuitive judgements about cases that seem unexplainaible on the Standard Fregean view. For example, the "Two-factor" Fregean view is able to explain the intuitive difference between cases in which sameness of MOPs is arrived at by luck, and cases in which there is a mismatch in MOPs between the speaker and the hearer. Let me illustrate this difference with two cases.

> MOPs-MISMATCH: I say "Hesperus is visible in the evening". You mishear me and think I have made the claim that *Phosphorus* is visible in the evening.

Here, the speaker thinks of Venus as the entity named "Hesperus" whereas the hearer thinks of Venus as the entity named "Phosphorus". The Standard Fregean view and the "Two-factor" Fregean view can make the same verdict: the mismatch in MOPs is (all other things being equal) a sufficient reason why communication failed. Compare with this other case:

> DEVIANT: I say "Hesperus is visible in the evening". You do not hear what I say. Instead, you are listening to your inner voice subvocally uttering the sentence "Hesperus is visible in the evening". Furthermore, you are mistaking what your inner voice is saying subvocally for my utterance.

Here, the speaker and the hearer think the same content under the same MOP, but the causal chain is deviant. That is, DEVIANT instantiates the relation **SLC** between the hearer and the proposition that Hesperus is visible in the evening. Crucially, unlike the Standard view, the "Two-factor" Fregean theory can make the verdict that DEVIANT is a case in which there is a *match in content*, but the causal chain is *deviant*. Because the theory features a twofold criterion,

it can account for the difference between these two kinds of cases.

To anticipate, the "Two-factor" Fregean theory will have to compete with a theory on which content just is referential content, and an anti-luck condition makes sure that the referential contents of the interpretation of the hearer and of the thought expressed by the speaker are connected in the right way. What is important to note is that *the condition that the hearer coordinate with the speaker on a fine-grained content* is not able to play the role of the anti-luck condition. We need a non-semantic condition to play that role.

We have improved the Standard Fregean View of communication by making it more externalist. Once we remarked that coordination on fine-grained content is always compatible with a Loar case, we arrived at the Two-factor Fregean view of communication by adding the condition **(c)** in the definition, a causal and purely external condition on successful communication. In what follows, I want to suggest that the Two-factor theory can be further improved by making it even more causal and externalist.[25] I will do so by contrasting two cases. The first case is introduced in Byrne and Thau 1996:

> WINSTON: A patient checks into the hospital and is assigned room 101. Tony dubs him "Winston" and the cognitive value she attaches to the name is: *the amnesiac in room 101*. Alex is thoroughly unaware that Tony has seen the patient, but by sheer chance she also dubs him "Winston" and attaches the same cognitive value to the name. Alex utters "Winston will never recover" in Tony's presence, and Tony forms the belief that she would express by saying "Winston will never recover". (Byrne & Thau 1996: 147 cited in Peet 2019)

Seen through the Standard Fregean view, WINSTON is a Super-Loar case. That is, Tony and the speaker Alex think of the same content under the same descriptive MOP, albeit by pure luck. The Two-factor Fregean view can accept the diagnosis that descriptive MOPs are shared, and explains that communication fails because the condition that the causal transmission chain be non-deviant and non-lucky is not satisfied in WINSTON.

But are we sure that Alex and Tony think of the referent under the same mode of presentation? After all, it seems that Alex and Tony are not justified in taking their use of the name "Winston" to be the same. Alex's use of "Winston", and Tony's use of "Winston" do not belong to the same chain of deference, or so it seems. In making this remark, the causal dimension of name use is emphasized over its purely qualitative and descriptive dimension. Here is a case to motivate this methodological orientation:

> CARLA: Tony and Alex have a colleague in common whose name is Carla. Carla

[25] I am drawing on Peet 2019 in what follows, who advocates a radical anti-luck approach. I won't discuss Peet's proposal here.

just had a child. Tony intends to communicate this fact to Alex, and says "Carla just had a child!". As a result Alex forms the belief that Carla just had a child.

In CARLA, it is not by luck that Tony and Alex use the name "Carla" to think and talk about the same individual. Their use of the name belong to the same network of deference and use (e.g. Kripke 1980, Devitt 2015). CARLA suggests that a more externalist version of the Fregean theory is possible, which individuate MOPs not by descriptions, but by external relations to the social, cultural, natural environment. We can use this insight from CARLA to produce another diagnosis on WINSTON: in WINSTON, Tony and Alex do not in fact share a MOP, hence communication fails. I examine this externalist trend of Fregean theory on our problem at hand in the next section.

4 The Sophisticated Fregean conception

4.1 Non-descriptive MOPs

What I call "Sophisticated Fregean view" is the conjunction of four innovations with respect to the Standard Fregean view of communication:[26]

1. For communication to be successful, the audience must think of the right referent under *a* MOP suitably related to the one of the speaker.

2. The equivalence class of all the MOPs suitably related to each other relative to a discourse situation constitutes a *shared sense*.

3. The shareable sense attached to a singular term as used by a speaker places a constraint on how the audience must think of the referent, but it does not appear in the truth-conditions of the communicated proposition.

4. MOPs are non-descriptive.

Some clarificatory comments are in order. The *Standard* Fregean view construed any putatively communicated MOP as a contextually relevant property that enters into the truth-conditions of the communicated proposition. By contrast, the Sophisticated Fregean allows that *how the audience must think of the referent* is typically truth-conditionally irrelevant. Moreover, the Sophisticated Fregean allows that the MOP of the speaker and the MOP of the hearer *may differ*, provided that these MOPs are *suitably related*. "Suitably related" does not mean "suitably similar". One main innovation of the Sophisticated Fregean view is that the *same-sense-as*

[26] I propose that we consider all these innovations together for the following reason. On the one hand, innovations (1) and (2) go hand in hand: they are implied by the notion that the level of content relevant for (**SHAR**)—senses—is *distinct* from narrow psychological content i.e. MOPs, as per (1) and (2). On the other hand, while (3) is indeed an independent innovation from (1) and (2), it is in the spirit of a widely endorsed conception among sophisticated Fregeans (e.g. Recanati 1993, Carpintero 2000), and forms an interesting package with the others. The same is true of innovation (4): since senses are equivalence classes of MOPs (as per 2), and are grounded in an external relation (as per 1), it is natural (in light of these) to claim that MOPs involve an external relation as well (as per 4).

relation is not reducible to intrinsic features of representations (such as *descriptive contents* or *intensions*, whose similarity would determine sense sharing). Instead, the relation of sameness of sense is partly external. By this I mean the following: that a MOP *a* share a sense with MOP *b* is not determined by a combination of intrinsic features of MOP *a* (that does not mention MOP *b*) and of intrinsic features of MOP *b* (that does not mention MOP *a*).[27] No intrinsic features of MOP *a* and of MOP *b* are sufficient to entail that they share a sense. As Dickie & Rattan 2010 write,

> Our version of [the *same-sense* relation] collects together modes of presentation differently in different situations: cases in which a speaker tracks an object, or interlocutors jointly attend to an object, collect together modes of presentation differently from cases in which tracking or joint attention is absent, *despite the fact that the modes of presentation involved may be the same in both kinds* (p. 149; my italics)

The "suitably-related-to" relation between MOPs necessary for successful communication/ understanding is construed as an equivalence relation individuating *senses* (the equivalence classes induced by the relation), such that two MOPs suitably related with each other *share a sense* relative to a discourse situation. Lastly, on the Sophisticated Fregean picture, MOPs do not refer via descriptions. Instead, they refer through causal relations, such as what Recanati (2012) calls 'epistemically rewarding relations' (ERs).

MOPs are needed independently of the problem of successful communication. They are introduced to explain how it is that a rational subject can believe of a given object *o* that it is both *F* and not *F*. I will cite a (non-modal) version based on Schiffer 1990:[28]

Frege's constraint (FC) A minimally rational subject *S* cannot simultaneously believe and disbelieve of a certain object *o* to be *F* under the same MOP. In other words, if a rational

[27] I am relying on a formulation by Goodman & Gray (2022: 10).

[28] One limitation of Schiffer's version of (**FC**) is that it is silent about the difference of MOPs in all cases where thinkers do not *actually* ascribe contradictory predicates to a same object. To overcome this limitation, we must formulate (**FC**) in *modal* terms. Modal versions of (**FC**) can be more or less strong, depending on which notion of "rational doubtability" they feature. A standard version is due to Evans 1982:

> **Frege's constraint: Evans' version**
> The expressions '*a*' and '*b*' are associated with distinct modes of presentation (for a given subject in a given context) if the subject could rationally assent to '*a* is F' and simultaneously withhold assent from, or reject, '*b* is F'.

This formulation is stronger than Schiffer's (entails it), because here the *mere* possibility of conflicting attitudes entails an actual difference in MOPs. Recanati (2016) proposes a middle ground between Schiffer's characterization and Evans'. We may formulate his proposal as follows:

> **Frege's constraint: Recanati's version**
> The expressions '*a*' and '*b*' are associated with distinct modes of presentation (for a given subject in a given context) if the subject can take (given her actual dispositions in the context) different attitudes vis à vis '*a* is F' and '*b* is F'.

Whereas in Evans' criterion the *mere* possibility of conflicting attitudes (independently of the subject's actual disposition in the context) entails an actual difference in MOPs, the criterion due to Recanati 2016 is sensitive to thinkers' actual cognitive dispositions. See Recanati (forth) for a discussion of these various non-equivalent formulations of (**FC**).

> subject simultaneously believes of a given object *o* both that it is *F* and that it is not *F*, then there are two distinct MOPs such that *S* believes *o* to be *F* under one, and *S* believes *o* not to be *F* under the other.

Frege's constraint partially defines MOPs in terms of their role in psychological explanation. In virtue of this role, MOPs so defined satisfy the Transparency constraint. There are several ways to characterize transparency (Wikforss 2015, Murez 2022). One way involves the notion of knowledge and is thus concerned with *epistemic access*. Here is a relevant formulation:

Transparency constraint With respect to any two of her thoughts or beliefs an individual can know apriori via introspection whether or not they exercise the same MOPs.[29]

Frege's constraint only provides us with a sufficient condition for a *difference* in MOP at a time and intrapersonally. So it is not very useful for theorizing about diachronic and interpersonal MOPs sharing. That is where the notion of *shareable sense* enters the picture.

4.2 Sharing a sense

We have the intuition that speech participants may deploy distinct MOPs but still share a perspective on the object nevertheless, at a more abstract level. This is the idea behind the notion of a shareable sense. Intuitively, the relation *same-sense-as* thus collects the MOPs through which thinkers share a perspective relative to a discourse situation, notwithstanding the difference in the MOPs involved. (Senses are thus distinguished from MOPs: they are equivalence classes of MOPs). We want to be able to draw a constrast between cases where the participants each deploy distinct MOPs but do not share a perspective on the object, and cases in which they do share a perspective with the distinct MOPs deployed. This contrast is well presented by a series of examples from Dickie & Rattan 2010 inspired from Heck 2002 (see Figure 1.7 & 1.9 vs. 1.8):

> **Example (A)** I say (while attending to a bottle at time t_1 from perspective π_1 and intending to refer to it) "That[*said at* t_1 *from perspective* π_1] is half-full". I then walk around the bottle, attending to it all the while. When I get to the other side I say, still intending to refer to the bottle I am attending to "That[*said at* t_2 *from perspective* π_2] is not half-full". Given that I have been keeping track of the bottle all the while, my mistake leaves me in a situation of rational incoherence. So my uses of 'that' must share a sense.

[29] Falvey and Owens 1994: 109 -110 cited in Murez 2016. Another formulation of access transparency (as Wikforss 2015 proposes to call it) is the following:

> For any two modes of presentation m and m' under which x thinks of y and z, respectively (where y may identical to z), x can know a priori, via introspection alone, that m and m' are the same, if they are the same, and that m and m' are different, if they are different. (Boghossian 1994: 39-40 cited in Murez 2016.)

Example (B) I say (while attending to a bottle at time t_1 from perspective π_1 and intending to refer to it) "That[*said at* t_1 *from perspective* π_1] is half-full". I can also see a bottle reflected in a mirror on the wall and (attending to that bottle, which I am seeing from perspective π_2, and intending to refer to it) I say "That[*said at* t_2 *from perspective* π_2] is not half-full". In fact my 'that$_{t_1}$' and 'that$_{t_2}$' refer to the same object. But my mistake does not leave me in a position of rational incoherence. So my uses of 'that' must differ in sense. (See Figure 1.7)

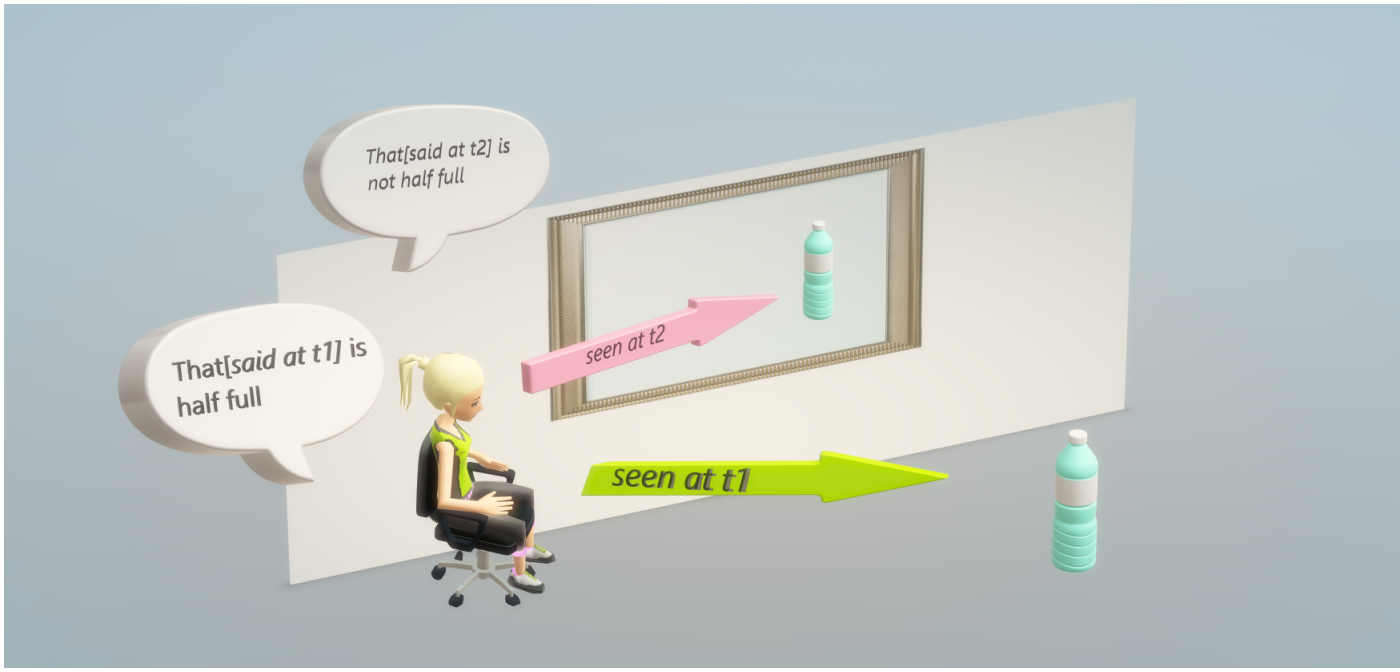


Figure 1.7 – Example (B) – Difference in sense

Example (C) You say 'That[*said at* t_1 *from perspective* π_1] is half-full'. I say 'That[*said at* t_1 *from perspective* π_2] is not half-full'. We understand one another's uses of the term in virtue of the fact that each of us is using it to refer to the object of our joint attention. Because we understand one another's uses of the term, our disagreement puts us into rational conflict with one another. So my use of 'that' and your use of 'that' must share a sense. (See Figure 1.8)

Figure 1.8 – Example (C) – Trading upon coreference is warranted

Example (D) You and I are sitting on opposite sides of a screen. Each of us is looking at a bottle in the unscreened part of the room. I am seeing it from perspective π_1. You are seeing it from perspective π_2. We do not realize that we are looking at the same bottle. I say 'That[*said at* t_1 *from perspective* π_1] is half-full'. You say 'That[*said at* t_1 *from perspective* π_2] is not half-full'. We are not in rational conflict with one another. So our uses of 'that' differ in sense. (See Figure 1.9)

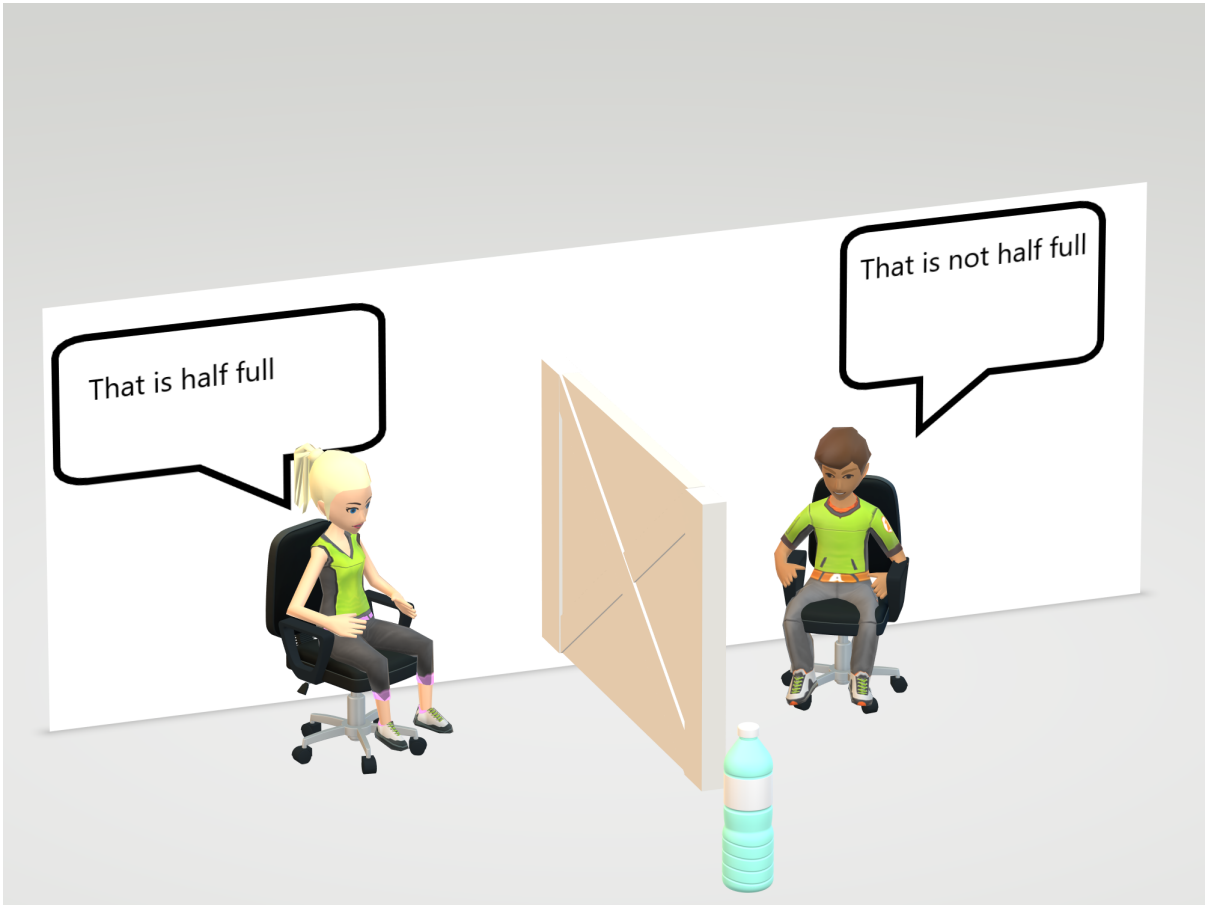


Figure 1.9 – Example (D) – Trading upon coreference is not warranted

Note the *externalists factors* in play in whether two distinct MOPs share a sense. In EXAMPLE (A), the thinker has kept track of the bottle while walking around it, as a result, a sense is shared by the various MOPs deployed at distinct times. In EXAMPLE (B), a mirror in the room misleads the thinker into believing that there are two distinct bottles, consequently, the non-synchronous MOPs at issue do not share a sense (see Figure 1.7). In EXAMPLE (C), two thinkers are jointly attending to some object and they are both aware that they each use the singular term "that" to refer to the object of their joint attention (see Figure 1.8). As a result, the MOPs of the different thinkers share a sense. In EXAMPLE (D) a screen in the room prevents the thinkers to jointly attend to the object, because of this the thinkers do not share a sense with the MOPs they each deploy, even though they are in fact both looking at the same object (see Figure 1.9).

Given these externalist factors at play in the relation of *sense-sharing*, we can already anticipate that the Super-Loar recipe will has less force against this conception of fine-grained content identity. Before I come to this, however, I need to provide a more theoretical way to individuate shareable senses.

I will rely again on Dickie & Rattan 2010's useful characterizations. When two thinkers respectively deploy MOPs which share a sense, Dickie & Rattan 2010 say these thinkers are *rationally engaged* with one another. What is it for two thinkers to be rationally engaged with one another? To answer this question, D&R (following Campbell 1987) consider a pattern of inference they call 'immediate extension of knowledge'. It is the relation between two premisses '*a* is F' and '*b* is G' (where '*a*' and'*b*' are token singular terms which may or may not be of the same type) such that it allows a minimally rational subject who accepts both of them to *trade on coreference* (Campbell 1987) and to directly derive the existential generalization to the effect that there is an *x* which is both F and G. Here is an example in the interpersonal domain:

> Consider a situation in which Ann and Bob are using a demonstrative to refer to an object to which they are jointly attending. Ann says:
>
> That is $F_{(\text{at } t_1)}$
>
> Bob says:
>
> That is $G_{(\text{at } t_1)}$
>
> Each of Ann and Bob hear and understand what the other says. So either of them would be warranted in moving directly to the conclusion:
>
> Something is both $F_{(\text{at } t_1)}$ and $G_{(\text{at } t_1)}$. (Dickie and Rattan 2010: 147; slightly modified by me)

Ann's use of "that" and Bob's use of "that" share a sense, because both Ann and Bob trade upon coreference of their respective uses of "that", and are *warranted* in doing so, given their relation to each other and to the target in the discourse context.[30] The relation of *warrant* here means something more than mere rational permissibility, because the notion of warrant at play is partly externalistic. The situation of use must be such that the environment cooperate: e.g. that the subject has effectively keep track, or that the subjects have effectively jointly attended, etc. But whether one has effectively kept track, or whether one has effectively jointly attended, cannot be decided on the basis of introspection: because of the external factors involved, these matters are opaque to introspection. If the situation of discourse is such that the speech participants are *warranted* in trading on the coreference of their respective uses of singular terms,

[30] See section 3 of the general introduction for an informal presentation of Trading on Identity (ToI) in the intrapersonal domain and Campbell's constraint, a constraint on the individuation for thoughts that incorporates trading on identity. Here is a relevant formulation, drawing on Recanati (forth):

> **Campbell's constraint**
> If two singular terms allow trading on identity (ToI), then they have the same sense.

According to Dickie & Rattan, Campbell's constraint applies to the interpersonal domain. Note that the constraint is vague as long as the invoked notion of "allowability" has not been defined. I show that there is an ambiguity attached to that notion in section 5.1.

then these singular terms share a sense relative to that situation of discourse.

In other words, two token singular terms in a given discourse situation have the same sense if, relative to that discourse situation, they warrant the presupposition that they have the same reference. There is nothing more to the notion of sense associated with a singular term than the warranted presupposition of coreference, according to the Sophisticated Fregean view. Again, let me insist on the *externalist* element of warrant here: whether the mutual presupposition of coreference is warranted is not wholly determined by the individual psychological states of the two participants; external factors enter into the determinants, such as whether joint attention is taking place; or whether keeping track is taking place, and so on. I discuss implications of this externalist element in the notion of shared sense in section 5. Here is D&R's criterion for the individuation of senses relative to a situation of use:

> **Individuation Principle for Senses** – The sense of ν (used by S_ν at t_ν with MOP M_ν) = the sense of μ (as used by S_μ at t_μ with MOP M_μ) iff the engagement-relevant factors in the situation of use generate the possibility of the immediate extension of knowledge upon full understanding. (Dickie & Rattan 2010: 150)

What D&R mean by "engagement-relevant factors" are, again, partly external factors such as whether the protagonists are jointly attending to some object and are both aware that they both refer to the jointly attended object; whether the subject has kept track of an object over time, etc. We may now provide the Sophisticated Fregean theory of successful communication, as follows:

The Sophisticated Fregean view: A hearer H understands a singular term ν as used by a speaker S iff

> **(a)** S utters ν to refer to some object and H understands that S is referring to some object with ν;
>
> **(b)** S and H attach the same sense to ν, that is:
>
> > **(b1)** S presupposes that the object S refers to with ν is the object H is thinking about when interpreting ν;
> >
> > **(b2)** H presupposes that the object S refers to with ν is the object H is thinking about when interpreting ν;
> >
> > **(b3)** the engagement-relevant factors in the situation of use warrant trading upon corefence as it occurs in **(b1)** and **(b2)**.

NB: Although the Sophisticated Fregeans are not always clear about this, as my formulation shows, an anti-luck condition (here, (b3)) is built into their notion of *sameness of sense*.

Having presented the central concepts of the Sophisticated Fregean view, and introduced a working formulation of its criterion for successful communication, I now present how the Sophisticated Fregean view intends to solve the Loar cases.

4.3 The Sophisticated Fregean solution to the (Super-)Loar cases

Remember KASPAROV, the Loar case I have presented earlier in the chapter (depicted in Figure 1.2). In KASPAROV, Anna and Bob both understand Anna's use of "that" as referring to the object of their putative joint attention. However, Anna and Bob are not jointly attending to anything: Anna is looking at her newspaper whereas Bob is looking at the guy in front of them on the bench. In virtue of this feature of the case, the Sophisticated Fregean will notice that the engagement-relevant factors in KASPAROV *do not* warrant trading upon corefence. In other words, condition (b3) in the Sophisticated Fregean definition of communicative success is not satisfied, so communication fails. We get the right prediction. Moreover, it seems that this way of construing KASPAROV can be generalized to *any* Loar case: In a Loar case, the speech participants are disposed to trade on coreference (i.e. they all think that the object the other is thinking about is the object one is thinking about), however the situation of use is such that trading on coreference is not warranted in that situation.

What is more, as already mentioned, it seems that the Sophisticated Fregean can iterate this style of explanation to my Super-Loar cases as well. For example, in the SUPER-LOAR CASE 1 (see again Figure 1.5), Bob and Anna are not jointly attending to the same token of the *New York Times* picture. Although they are in fact both looking at a (type- but not numerically-) identical photo, they are not jointly attending. Because of this factor in the situation of use, their disposition to trade on coreference is not warranted. Similarly, in the SUPER-LOAR CASE 2 (see again Figure 1.6), because Bob is looking at a mirror without realizing it, the disposition of the protagonists to trade on coreference is not warranted. More generally, for any situation of use instantiating a Super-Loar case, we can expect that *at least one factor* of the situation of use *defeat the warrant* for trading on coreference.

Having presented how the (Super-)Loar cases should be solved according to the Sophisticated Fregean view, I will now criticize this way of explaining the (Super-)Loar cases. My critique will have a foundational aspect. In a nutshell, I will argue that Sophisticated Fregeans are still confusing content with a condition on the causal transmission chain. But we get a better picture of communication and samethinking when we do not confuse these two dimensions.

5 Problems for the Sophisticated Fregean conception

The "two-factor" Fregean view (section 3.3) had the merit of distinguishing between what belongs to the sharing of content, and what belongs to the causal transmission chain. The Sophisticated Fregean view does not have this merit. For instance, the Sophisticated Fregean view cannot distinguish a MOPs-MISTMATCH type of case from a DEVIANT type of case (see sect. 3.3). This is an indication that something is wrong with the Sophisticated Fregean notion of sense: one would have hoped to be able to distinguish these two types of case with a notion of fine-grained shared content. But the doubts are confirmed when we examine the notion of

shared content that it puts forward.

5.1 Senses are half opaque

There is an ambiguity with the notion of 'rational engagement' the Sophisticated Fregeans are using. On one construal, a thinker A and a thinker B are rationally engaged with one another if and only if they are both diposed to trade on coreference of their respective thoughts, or singular term uses. Call *rational engagement* construed in this way, COORDINATION. COORDINATION obtains just in case both the conditions **(b1)+(b2)** defined above obtain. This is roughly what Recanati (2016) calls *weak coreference de jure*. However, this notion is *not* what explains communicative success and samethinking. For it is too weak. To see this, observe that in the Loar cases, speech participants *are* disposed to trade upon coreference of their thoughts or token uses. On another construal, a thinker A and a thinker B are rationally engaged with one another if and only (i) if they are both disposed to trade upon coreference AND (ii) the situation of use warrants the disposition to trade on coreference as in (i). Call *rational engagement* construed in this way, SUCCESSFUL COORDINATION. SUCCESSFUL COORDINATION obtains whenever the conditions **(b1)+(b2)+(b3)** as defined above obtain. The notion of *sense* the Sophisticated Fregeans are using involves SUCCESSFUL COORDINATION, not COORDINATION.

I am now in a position to formulate my criticism against the Sophisticated Fregean notion of sense:

> **(1)** SUCCESSFUL COORDINATION is not transparent to thinkers.
>
> **(2)** The point of introducing a notion of content finer-grained than referential content consists in its TRANSPARENCY (otherwise, why not just reference?)
>
> **(3)** Therefore, SUCCESSFUL COORDINATION should not be understood as a matter of fine-grained content-sharing (i.e. in terms of shared sense).

Here are some clarificatory comments. Premiss 1 says that SUCCESSFUL COORDINATION is not transparent to thinkers. This is imprecise. In what follows, I will specify this notion using an epistemic characterization of transparency.[31] *Knowing whether coordination is successful* (that is, knowing whether what I am thinking about is what my interlocutor is thinking about) has two parts. I can see two versions for each part, as follows.

What the transparency of successful coordination would be like

STRONG VERSION

[31] I offer another characterisation in terms of functional role, inspired by Murez 2022, in the general conclusion of the dissertation.

For any communicative event between A and B:

(i) If coordination is successful, then A and B are able to have common knowledge that it is;

(ii) If coordination is not successful, both A and B are able to have common knowledge that it is not.

MODERATE VERSION

For any communicative event between A and B:

(i') If coordination is successful, then both A and B are able to know that it is;

(ii') If coordination is not successful, then both A and B are able to know that it is not.

Let me put aside the strong version for the moment.[32] I cannot imagine a scenario in which coordination is successful, but the speech participants fail to know this. The theoretical support in favor of this intuited impossibility is that SUCCESSFUL COORDINATION may in part consist in the common belief that coordination is successful, or as the Sophisticated Fregeans like to say, in rational engagement.[33] If this lack of hypothetical scenario is not the result of a limit of my imagination, then it seems that at least one of (i) or (i') is true (note that (i') is entailed by (i)). Using the terminology of senses, we may say that *sameness of sense* IS transparent. However, it is clear that (ii) and (ii') are both false. Since (ii') is weaker than (ii), it is enough to show that (ii') is false to show that (ii) is false as well. But Loar cases show that (ii') is false. I shall say that *senses are half opaque* to mean that difference in sense is not transparent to thinkers:

Senses are half opaque
Difference in sense is not transparent to thinkers.

Another remark should be made about premiss 1 and the notion of TRANSPARENCY it features. The characterization of the TRANSPARENCY constraint for MOPs requires that knowledge of sameness and difference in MOPs be attained via introspection alone.[34] But even if successful coordination were transparent in the sense of (i)+(ii), it would surely not be on the basis of introspection alone. So I find that it is almost a category mistake to speak of transparency in the intersubjective domain. It is only in a limited sense that successful coordination/sameness of sense is made transparent to us. For example, knowledge of sameness of sense typically involve perception and mind reading (e.g. in cases involving demonstratives).

[32] See section 3 of the general introduction for the distinction between common knowledge and mutual knowledge, which underlies the distinction between what I call the strong and the moderate version.

[33] See KASPAROV(c) of sect. 1.5, in which the participants both know that their thoughts corefer, but *because* one of the participants believe the other does not know their thoughts corefer, coordination is not successful.

[34] I am dealing with the standard characterization of the notion of access transparency for present purposes (Wikforss 2015).

This brings me to premise 2. Premise 2 says that it is not useful to introduce a level of content finer-grained than reference in one's theory of content, if that notion is not transparent. Why not just reference? It seems that one does not need an intermediate level of content between psychological (narrow) content and referential content in order to explain the disposition thinkers may have to trade on coreference of one another thoughts, or uses of singular terms (see Recanati (forthcoming) where this idea is defended). Nor do we need this intermediate level of content to explain successful communication, or samethinking more generally. Here is a more general line of thought, 'by Ockham's razor', to bring this point home:

> **(P1)** For any explanandum X and any two equally explanatory/predictive theories T_1 and T_2 each explaining X, if T_1 is more parsimonious than T_2 then, everything being equal, we should prefer T_1 over T_2.
>
> **(P2)** It is more parsimonious to explain communicative success without *senses*.
>
> **(C)** Therefore, a *senses*-free theory of communicative success should be preferred over a Fregean theory.

In the absence of shared content, it might be tempting to say that thinkers who are warranted in being disposed to trade upon coreference of their MOPs, *share a distributed file* on the object. Clearly, such distributed files exist. However, if my argument against the sophisticated Fregeans is correct, such distributed files are not transparent.[35]

5.2 We don't have to construe non-lucky coreference in terms of sameness of sense

What would the *form* of a *senses*-free theory of communicative success be like? I already alluded to it when I introduced the Two-factor Fregean theory, let me requote the relevant passage:

> The "Two-factor" Fregean theory will have to compete with a theory on which content just is referential content, and an anti-luck condition makes sure that the referential contents of the interpretation of the hearer and of the thought expressed by the speaker are connected in the right way. What is important to note is that *the contition that the hearer coordinate with the speaker on a fine-grained content* is not able to play the role of the anti-luck condition. We need a non-semantic condition to play that role.

[35] See the next chapter, where I continue the discussion with the sophisticated Fregeans.

So the theory at issue would have the following form:

***Senses*-free view:** A speaker S successfully communicates that p to a hearer H iff:

> **(a)** S utters s to communicate some proposition and H understands S to be communicating some proposition with s.
>
> **(b)** S and H think the same referential content.
>
> **(c)** (a) and (b) are not satisfied by luck or in virtue of a deviant causal chain.

In the next chapter, I go in search of a theory that instantiates this promising scheme. Ideally, one would like to be able to provide a necessary condition that *explains* the elimination of luck, rather than citing a brute anti-luck clause such as (c).

2 On what might prevent communicative luck

Abstract

Studying the Loar cases in the previous chapter has taught us a few things about communication. I have argued that successful communication requires knowledge of what is said, and that a satisfactory theory of communication must distinguish conditions on successful communication that have to do with content, and conditions that have to to with the causal transmission chain. Furthermore, I have argued against theories that introduce a level of shared content finer-grained than reference (whether descriptive or non-descriptive) in order to eliminate luck. In the end, all this leads us to search for the following kind of theory of communication:

***Senses*-free view:** A speaker S successfully communicates that p to a hearer H iff:

(a) S utters s to communicate some proposition and H understands S to be communicating some proposition with s.

(b) S and H think the same referential content.

(c) (a) and (b) are not satisfied by luck or in virtue of a deviant causal chain.

In this chapter, my goal is to provide a theory that instantiates this schema. What needs to be defined is the condition (c). Leaving (c) as is i.e. mentioning a pure anti-luck clause is a last resort, when one has exhausted the candidate analysans of (c).

I first consider the following anti-luck condition: the hearer must interpret the speaker's utterance in virtue of attending to the intended *inference-based feature*, namely, the information the speaker intends the audience to use in order to retrieve the referent (Buchanan 2013). I show that this condition, although it explains how luck is eliminated in some cases, is not a general solution to the problem of the Loar cases.

I then consider the following anti-luck condition: the hearer must interpret the speaker's utterance in virtue of jointly attending with the speaker to the intended ib-feature. The idea behind this approach is that joint attention provides coreferential safety, because it is a factive state, one the participants can only be in if they are actually focussing on the same object with the common awareness that they are. Joint attention on ib-features thus brings it about that every element of contextual information used in interpreting the utterance is not only mutually known, but commonly known. As a result, the speech participants have common knowledge that the hearer is recovering the correct interpretation, and luck is eliminated.

1 Introduction

1.1 Coreference by coincidence vs referring together

In the previous chapter, I have characterized the Loar cases as cases in which it is a matter of luck that the interpretation of the hearer corefers with the thought the speaker wanted to express. Accordingly, if we can formulate a condition that eliminates communicative luck, then we can solve the Loar cases and thereby provide a criterion for communicative success. This chapter is the second part about communicative luck, and conditions on the interpretation process that might eliminate it.

Although we do have an intuitive grasp of the notion of luck, it is not easy to say what makes an event lucky in general terms (in the sense at issue). One idea is that typical instances of luck (e.g. lottery winning events) could easily have failed to occur. Applying this characterization to communication, to say that the output of an interpretation process turned out to be *luckily* coreferential with the thought the speaker intended to express is to say that it could easily have been wrong i.e. not coreferential at all. On this characterisation of luck, an anti-luck condition is therefore a safety condition: one that makes the interpretation process robust in a modal sense i.e. such that the interpretation would still be coreferential with the message in close possible worlds (see e.g. Pritchard 2005).

A notion close to that of communicative luck relevant to characterizing the Loar cases is the notion of coincidence. In typical Loar cases, calling '*o*' the referent at issue, the fact that the thought of the hearer refers to *o* and the fact that the thought of the speaker refers to *o* are produced by independent causal factors. In other words, to explain the coreference of the respective thoughts of the protagonists, we need to cite independent explanations for each thought and conjoin them[1]. The coreference between the two thoughts is *coincidental*.

Contrast a typical Loar case with a case of successful demonstrative communication where the speech participants jointly attend to the referent. In such a case, it is not a coincidence that the thought of the hearer and the thought of the speaker corefer: for the hearer is successfully monitoring the direction of the speaker's attention, and the hearer's own attention is controlled by the direction of the speaker's attention. Calling *o* the referent at issue, the fact that the hearer's thought refers to *o* is explained *not independently* of the fact that the speaker's thought refers to *o*. Informally we may say that the speech participants manage to refer *together* rather than by coincidence.

The structure whereby the reference of a thought is explained not independently of the reference of another thought is not the prerogative of demonstrative communication. We find a similar structure in communication involving proper names. When I use the name 'Garry

[1]See Aristotle, chapters 4–6 of Physics II (Charlton 1970) for a similar definition of 'coincidence'.

Kasparov' and you use the name 'Garry Kasparov', it is no coincidence that we are referring to the same person. Rather, both of our uses belong to the same causal chain of deference and use. Safety is causal, but it can span conversational contexts. We may think of each successful conversational context as connected components of MOPs suitably related, and, zooming out, each conversational contexts as nodes in turn, connected to each other by deferential relations.[2]

I have presented *joint attention* and *deference* as possible safety mechanisms, participation in which guarantees a non-coincidental coreference with the thought of one's interlocutor. Another, perhaps more generic way to theorize about coreferential safety in communication is in terms of *intention recognition*. In this framework, understanding an utterance just *consists in* recognizing the speaker's referential intention. Roughly, when a hearer successfully recognizes which object the speaker is intending to refer to and communicate about (together with other aspects of the referential plan, perhaps) then it's not coincidental that their thoughts corefer. Intention recognition should not be here conceived as a safety mechanism in competition with joint attention or deference, but rather as possibly relying on them.

I have informally defined luck or coincidence as the concurrence of independent causal factors, or what could have easily not happened. The idea of this chapter is to explain whether and how joint attention or intention recognition may explain how communicative luck is eliminated. Namely, how joint attention or intention recognition may bring it about that the thoughts of speech participants refer *together*, in a suitably related way, rather than by coincidence.

1.2 Chapter plan

In the second section, I introduce the idea that communication is a matter of intention recognition. Looking at communication in this way, a natural line of thought is that intention recognition might typically be what eliminates communicative luck. A central theme of this section is the idea that the referential plan of a speaker – that is, her plan to make her audience think of a certain object she intends to communicate about – typically includes the intention that *certain features of the utterance* be utilized in how the hearer infers what the speaker intends to communicate about. We may refer to such features, following Schiffer (forthcoming a, b), as an utterance's *inference-based* features ("ib-features" in short). The main goal of this section will be to introduce ib-features, and to show the work their recognition can do to eliminate communicative luck.

In the third section, I put to the test the robustness of this intentionalist conception of communicative safety. In particular, I wonder whether ib-features recognition may be lucky. I argue that it can by presenting two cases from the literature on communicative luck.

[2] I will introduce chains of deference and use in the next chapter.

In the fourth section, I introduce *joint attention* as a communicative safety mechanism. I distinguish two kinds of referential communication: *deictic*, where the object is present in the discourse situation; and *non-deictic*, where the object is not present or not observable in the discourse situation. I explain how joint attention might be used to analyze communicative success in both kinds of communication.

I present a characterization of joint attention that is less intellectualistic than the notion of *common knowledge* usually invoked to theorize about communicative success. Its virtue is that it allows us to understand how common knowledge is reached by having joint attention as its sole source. Doing this, I attempt to contribute to a newly evolving alternative approach to shared knowledge in communication (Campbell 2005, 2017, Peacocke 2005, Wilby 2010, Seeman 2019, Schroeter 2012).

In the fifth section, I point out that joint attention to ib-features is not necessary in non-face-to-face communication. I argue that there is a natural distinction between oral face-to-face communication and non-face-to-face communication which answers this worry. Corresponding to this distinction is a distinction among two types of understanding, one of which requires joint attention, but not the other. Then I compare the proposed approach to the sophisticated Fregean view which I have discussed in the previous chapter. Despite strong surface similarities, I emphasize the differences between the two conceptions, and I defend the criterion proposed here.

2 Ib-features recognition as an anti-luck condition

This section presents the idea that intention recognition might explain how communicative luck is eliminated. I start with a simple definition of communicative intention, then I add more structure to the referential plan in order to account for the Loar cases.

2.1 Intention recognition

The idea that intention recognition plays a key role in verbal communication is intuitive, and has been enormously influential in linguistics and the philosophy of language.[3] As Searle (1965) says, paraphrasing Grice (1957) :

> In speaking a language I attempt to communicate things to my hearer by means of getting him to recognize my intention to communicate just those things.

The central idea is that what a speaker means in any given event of verbal communication is a matter of what the speaker intends to communicate. Furthermore, a speaker intends do

[3]For critical discussions, see e.g. Millikan 1984 and Gauker 2019. More generally, there is an alternative tradition to the Gricean program which is broady Austinian, which appeals to the perfomance/recognition of rules or conventions. The *locus classicus* is Austin 1975; see e.g. Witek 2022 for an introduction. Neale 1992 is a good survey of the Gricean program. I rely heavily on Buchanan 2013 for this part of the chapter.

communicate something only if she wants to impact their hearer's mind *by way of* making transparent to them just that intention. Here is a first rough definition of communicative intention, adapted from Grice (1957):

Intention to communicate
S intends to communicate that *p* with *u* to *H* if and only if *S* produced *u* intending:

(i) *H* to entertain that *p*;

(ii) *H* to recognize that *S* intends (i) on the basis of their recognition that *S* produced *u*. (Buchanan 2013 slightly modified)

One may then define successful communication in terms of successful recognition of the communicative intention, as follows:

Successful communication
A speaker *S* successfully communicates that *p* to a hearer *H* iff:

(a) *S* intends to communicate that *p* with *u*;

(b) *H* recognizes that *S* produced *u* to bring it about that *H* entertains that *p*.

(c) H entertains that *p* as a result of (b).

Abstracting from the complexities, the definition says that communication succeeds if and only if the speaker's communicative intention is recognized by the hearer. One problem with this definition is that the conditions it gives for a speaker's intending to communicate that *p* are not *sufficient*.[4] Buchanan provides the following case to illustrate why:

> FLAT-TOP MOUNTAIN: In observance of a religious holiday, Smith is forbidden to read, write, or speak for the day. Because Smith is looking so bored, his friend, Jones, tells Smith he will take him to a movie, but they need to decide what to see. It is mutual knowledge between them that a cowboy movie entitled 'Flat-top Mountain' is one of the many movies playing at their local Cineplex. Smith grabs his notebook and draws a mountain (in clear view of Jones), intending to communicate thereby that he would like to go to see *Flat-top Mountain*. Jones, however, mistakes the drawing for one of a cowboy hat, and infers thereby that Smith would like to go to see *Flat-top Mountain*. (Buchanan 2013: 62)

In FLAT-TOP MOUNTAIN, Smith intended Jones to recognize his intention to communicate that he would like them to go to see *Flat-top Mountain* by recognizing that he drew *a mountain*. But the pictorial content Jones experiences and utilizes to recover what is said is not the pictorial

[4] As Grice himself already anticipated, see his discussion (cited in Buchanan 2013) of Searle's putative counterexample involving the American Soldier captured by Italian troops in World War II in Grice [1969: 161–5].

content Smith intended Jones to recognize, for Jones sees the picture as a *hat-picture*. As a result, communication is lucky to some degree. Note that, given the purposes of the conversation, it does not really matter that Jones failed to recognize the drawing in the way intended, because the communicative exchange still enables them to successfully coordinate on the action plan.[5]

FLAT-TOP MOUNTAIN suggests that what is missing from conditions (i) and (ii) in our previous definition of communicative intention is the requirement that the speaker intends the audience to recover his communicative intention *according to a certain inferential path*. As Bach (2006: 524; mentioned in Buchanan 2013) says: "If your audience identifies the [object] in some other way [than the one intended], that's a matter of luck, not of successful communication". In other words, the referential plan of a speaker is typically more fine-grained than what the condition (i) and (ii) above describe. It includes what we may call an intended inferential path for identifying the referent from the recognition of certain *features* of the utterance. (An utterance, in the sense at issue here, may be any overt behavior that serves as evidence of an agent's intentions, such as – like in FLAT-TOP MOUNTAIN – drawing something). Following Schiffer, let us call *inference-based* features (*ib-features* in short) the features whose recognition must serve as premisses in the intended inferential plan of a speaker. I now turn to them.

2.2 ib-features recognition

Let me repeat the upshot of the last paragraphs. The speaker not only intends the hearer to think of the correct referent. In addition, the speaker intends the hearer to use certain informations provided with her utterance in order to retrieve the intended referent. So a speaker's referential intention not merely involves an instruction to think of a certain object *o* as the object at issue. In addition, it involves an instruction to identify *o* through a certain inferential path, based on the recognition that the utterance has certain features.

For example, in FLAT-TOP MOUNTAIN, Smith intended Jones to think of the object at issue by recognizing the drawing of a mountain. So the intended ib-feature in this exemple includes: *being the drawing of a mountain*. In *linguistic* communication, the relevant ib-features typically include the fact that the speaker uttered such and such words with such and such standing-meanings in the relevant shared language. However, ib-features are not limited to features that are evidentially relevant irrespective of facts about the circumstances of utterance. On the contrary, ib-features can be just about anything: any feature of the circumstances of utterance may in principle be recruited as an ingredient of the ib-feature. Hence the sense in which ib-features are features of an utterance is very inclusive, for it may include virtually any feature of the extra-linguistic context.

Consider this example from Schiffer 1981 (mentioned in Buchanan 2013: 64). A 'prelinguistic

[5] See 1.5 of the previous chapter for a theoretical support for this kind of diagnosis.

speaker' loudly utters 'GRRRR' in order to communicate thereby that he is angry. Here, to recognize the relevant communicative intention involves the recognition that the sound he produces resembles the sound that nearby dogs make when they are angry. Because ib-features can be about anything, we should be reluctant to put them in *what is said*. Instead, ib-features should be seen as pertaining to the pragmatic overall background success conditions for carrying out communicative intentions. Having acknowledged ib-features recognition as being integral to communicative intentions, we are now in a position to formulate the sophisticated Gricean view on successful communication:

Buchanan's criterion

Intention to communicate*
S intends to communicate that *p* with *u* to *H* if and only if, for some ib-feature Ψ of *u*, *S* produced *u* intending:

> **(i)** *H* to entertain that *p*;
>
> **(ii)** *H* to recognize that *S* intends (i) at least partly on the basis of their recognition that *u* has ib-feature Ψ. (Buchanan 2013 slightly modified)

We may then define successful communication in terms of the successful recognition of the communicative intention, as follows:

Successful communication*
A speaker *S* successfully communicates that *p* to a hearer *H* iff:

> **(a)** *S* intends to communicate that *p* with *u* and ib-feature Ψ of *u*;
>
> **(b)** *H* recognizes that *u* has Ψ and that *S* intends *u – cum –* Ψ to bring it about that *H* entertains that *p*.
>
> **(c)** H entertains that *p* as a result of (b).

Leaving aside certain details, the core idea of Buchanan's criterion is that communication succeeds just in case the hearer recognizes the speaker's communicative intention in the intended way.

It is clear how ib-features recognition is supposed to explain the Loar cases. Namely, the Loar cases are cases in which *what is said* is grasped but not through the *intended ib-feature*, which is not recognized. For the sake of concreteness, let me very briefly illustrate with a previous example. In KASPAROV, Bob thinks of the correct referent, but does so *not* in the way intended. Anna intends Bob to identify the referent by using the information that the referent *is depicted on the front page of the New York Times*. But Bob looked at the man in front of him instead. Hence,

communication is lucky, not successful.

One difference with the standard Fregean solution in terms of descriptive MOPs is that on the intentionalist Gricean view, ib-features are not part of *what is said*. In other words, the intentionalist Gricean view is a *referential* view of communication.[6] Moreover, ib-features can be grasped non-explicitly: they need not be entertained as descriptions, but can involve qualitative states, and be targets of a non-propositional *de re* awareness, not unlike the sort of awareness you have of, say, being angry or cold : you don't have to explicitly predicate of yourself that you are angry or cold in order to know it (see section 4.4 for an elaboration on this notion). However, despite these differences with the standard Fregean view, one may suspect that both views are similar enough to stand or fall together with respect to the threat of the Super-Loar cases (*pace* Buchanan 2013). This is what I argue in the next section.

3 Problem with ib-features recognition as an anti-luck condition

In the previous section, we have considered referential intention recognition (including the inference based features) as what could make the interpretative path from *character* to *what is said* non-lucky. Here I will argue that ib-intention recognition is still compatible with a Loar-case. I do so by presenting two cases from the literature on communicative luck.

3.1 A Super-Loar case of communication of first-person thoughts – Tayebi 2013

The first counterexample I put forward is proposed by Tayebi (2013), who echoes an example from Frege (1918). It is depicted in Figure 2.1:

[6] Referential views are also called *Russellian* in the literature, because at one time Russell (see e.g. Russell 1903) supported the view that singular terms contribute only their referents to the proposition expressed by utterances of sentences in which they occur. See Kaplan 2012 for a good presentation of Russell's ideas at the relevant time.

Figure 2.1 – Tayebi's example. Dr. Lauben is depicted by the man whose image is reflected in the mirror. Leo Peter is depicted by the man with black hair.

> TAYEBI: Suppose that Dr. Lauben [...] is in a room and intends to communicate his thought that he himself has been wounded to Leo Peter, who is in the same room. Unbeknown to [Peter], there is a large mirror in the room, and he mistakenly believes that the space he perceives in the mirror is another part of the room. Peter sees Lauben's image in the mirror too, and, because of his ignorance of the mirror, does not realize that he is just seeing the man sitting next to him. While Lauben is talking to him, Peter is highly curious about the man in the mirror and tries to find out what he is uttering. So Peter has two different (and unlinked) [MOPs] of Lauben: one is $[\text{HE}]_{\text{Lauben}}$, based on his direct perception of the man sitting next to him, the other is $[\text{HE}]_{\text{Nebual}}$, based on perceiving Lauben in the mirror.
>
> This is the moment when Lauben utters the sentence "I have been wounded". Peter hears this utterance; but, due to his focus on the person in the mirror, he takes this as an utterance by that man, whom he does not take to be identical to the man sitting next to him. So he adds the property of *being the utterer of this token of "I have been wounded"* to the content of just one of his two [MOPs] of Lauben, i.e.$[\text{HE}]_{\text{Nebual}}$. (Tayebi (2013): 214)

TAYEBI is designed to be a counterexample to Recanati's theory of first-person communication; for this reason, the example is formulated using the terminology of mental files. But we need not worry about this now. What is important to note, as I shall explain, is that Leo Peter seems to recover the right referent *in the intended way*, namely, by using the information that the intended referent *is the utterer of this token of "I have been wounded"*.[7]

[7] Of course, *this token* refers to the token Dr. Lauben is uttering (in the described situation).

In TAYEBI, Dr. Lauben is saying to Leo Peter that he has been wounded by uttering "I have been wounded". In doing so, Dr. Lauben not only intends Leo Peter (the hearer) to recognize to whom he is referring by "I", he also intends that Leo Peter recognize the referent as a result of its being common knowledge between them that the referent is *the utterer* of the utterance in question. In this example, the intended ib-feature has to do with the conventional meaning of "I". However, Leo Peter does seem to think of the referent in part on the basis of his awareness of the conventional meaning of that expression-type. Recall the definition of successful communication of the sophisticated Gricean:

Successful communication*
A speaker S successfully communicates that p to a hearer H iff:

(a) S intends to communicate that p with u and ib-feature Ψ of u;

(b) H recognizes that u has Ψ and that S intends $u - cum - \Psi$ to bring it about that H entertains that p.

(c) H entertains that p as a result of (b).

In TAYEBI, the conditions (a)-(b)-(c) seem to be all satisfied, and yet intuitively, communication fails. So TAYEBI is a counterexample to the claim that ib-features recognition is what eliminates communicative luck in every case.

One might object that it is not clear that the relevant ib-features have been recognized *in full* by Leo Peter in TAYEBI. The objection is that the intended ib-features might be *richer* than what the diagnosis that makes it a counterexample assumes. In effect, the relevant ib-features in TAYEBI might include other features than the one having to do with the conventional meaning of "I" (and whose appreciation by Leo *does* constrain his identification of the referent). For example, it might include the information *that the speaker is sitting next to the audience*, or something in the vicinity. Clearly, Leo is not using that sort of information in inferring what the speaker was referring to.

I record this response to the counterexample claim as an interesting one. While it is not plausible that only an audience sitting next to Dr. Lauben would be able to understand what he is saying, it might be part of the way he addresses Leo Peter in particular. If this line of thought makes sense, then for any utterance u, we should distinguish the conditions for understanding related to the *specifics* of the communicative exchange between the *actual* speech participants as they are related to each other in the situation, and the conditions for understanding of u in general, for any possible audience – i.e. such that a third party could understand u in *overhearing* the communicative exchange between the actual participants. (I articulate this distinction further in section 5.1 below). However, I also note the potentially *ad hoc* character of this objection, in the following sense: it seems that it will always be possible to postulate richer intended

ib-features to guard against counterexamples. This is so especially because, in principle, any feature of the context of an utterance may be recruited as an ingredient of the ib-feature. For example, it might be open to a rich conception of ib-features to claim that it is part of the ib-feature of Dr Lauben's utterance that the hearer *should feel empathy* for pain for the speaker. A sound methodology thus requires to control the pragmatics of the example and, in particular, the richness of the intended ib-features. What is more, for any plausible intended ib-feature one hypothesizes, I do not see why one could not adjust the case in such a way that the hearer has a gettierized recognition of the referential intention, using the Tayebi recipe illustrated above.

Another related objection is that the picture of referential intentions (and their recognition) assumed in the diagnosis that accepts TAYEBI as a counterexample, is too *externalist*. Let me explain. In TAYEBI, what makes us think that the referential intention of Dr. Lauben *has been* recognized by Leo Peter despite the fact that Leo Peter believes the referent is *not* Dr. Lauben, includes the assumption that referential intentions are *de re* intentions. What is assumed is that, in uttering a claim of the form '*o* is *F*', a speaker intends to communicate of *o* that it is *F*, that is, intends the audience to token a thought *of o*, to the effect that *it* is *F*. But in TAYEBI, the external relation between the MOP Leo Peter deploys in order to interpret the speaker's utterance, and the referent, is such that Leo Peter is *in fact* thinking of Dr. Lauben, even if Leo Peter would *reject* the sentence "Dr. Lauben is the utterer of this token of *I have been wounded*".

However, one might have a more *internalist* picture of referential intentions and their recognition, on which the intention Leo Peter thinks to recognize is *not* the intention expressed by Dr. Lauben. (If, by chance, you have the intuition that Leo does not recognize Dr Lauben's communicative intention in TAYEBI, then you might have an internalist conception of the sort). Reasons to think this has to do with the fact that Leo Peter falsely believes that the intention to communicate that one has been wounded is from a person *other* than Dr. Lauben. According to the way Leo Peter mentalizes the person who expresses this fact about himself, that person possess mental states which are related to the fact that (according to Leo's perspective) he is not Dr. Lauben, e.g. that he is at a certain non-zero distance from Dr. Lauben, etc. A more internalist picture of intention recognition will be sensitive to such facts about the way Leo mentalizes (perhaps tacitly) the speaker. As an upshot, the view according to which intention recognition requires a *de dicto* recognition, might *explain* the intuition that Leo Peter *does not* recognize Dr. lauben's communicative intention in TAYEBI. If this view is correct, then TAYEBI does not constitute a Loar case for this conception of successful communication.

I believe that the internalist picture of intention recognition articulated in the last objection makes intention recognition too stringent. Clearly, a person who overhears Dr Lauben's utterance might successfully understand his utterance, even if that person has no other information about the speaker.[8] Therefore, I believe that it is not plausible to claim an audience must

[8]But see the distinction above between conditions for understanding an utterance related to the *specifics* of the

represent *de dicto* the content of the relevant referential intention in order to be said to have recognized it.

Still, I find it intuitive to claim that Leo's false identity belief is what prevents him from genuinely understanding Dr. Lauben's utterance. Sticking to an intentionalist picture of communication; instead of saying that one must represent *de dicto* the content of a communicative intention in order to fulfill it, we might try to say rather that the audience *must be at least disposed* to have this more fine-grained representation of the referential intention. But it seems that Leo's false identity belief prevents just that.[9]

One may object on more fondamental grounds still to the alleged counterexample, as follows. Given a speaker S, a hearer H, an utterance u, a proposition p and some ib-feature Ψ attached to u, and calling q the following proposition:

(q) S intends H to think that p based on the recognition that u has Ψ.

If H achieves to recognize the communicative intention of S in producing u, then the relation between H and q is knowledge. (Again, I don't mean that the state of intention recognition need to be explicit. For example, it might be a state of *de re* awareness).[10] So the objection is that there is a mental state of believing q, and there is a mental state of knowing q, but there is no intermediate mental state of believing truly that q (as a williamsonian might say, see Williamson 2000: 27-28). As a result, it is simply not the case that Leo Peter has recognized the relevant communicative intention.

Put more simply, the objection is that *intention recognition* is not a gettierizable state, for the following reason. If successful communication is the recognition of the communicative intention, and if the purpose of successful communication is the transmission of knowledge, and we should not try to capture necessary *and sufficient* condition for knowledge, then we should not try to capture sufficient conditions for intention recognition either. The lesson to draw might be that one could define communication in terms of knowledge transmission without reductively defining knowledge.

I believe there is *substance* to this objection. When ones tries, by appealing to intuitions about cases, to analyze knowledge and in particular what it is that eliminates epistemick luck, we decide whether safety obtains by first deciding whether *knowledge* obtains. Safety has got to be understood only in terms of knowledge. But it is dubious that safety so understood can serve in an analysis of knowledge, because we do not have a grasp of safety independent of our grasp of whether *knowledge* occurred (see Williamson 2009 cited in Ichikawa & Steup (2018)).

exchange between the actual participants vs. generic conditions for understanding the utterance.

[9] An adequate way to respect these contradictory externalist–internalist intuitions may lie in reconciling an externalist criterion with phenomenological or internal dimensions of meaning – an option I explore in chapter 5.

[10] See e.g. See Korcz 2021, SEP on this notion, and section 4.3.1 of this chapter.

The analysis of communication seems to be in the very same predicament as the analysis of knowledge regarding what might eliminate luck. It is true that, as already argued, in some cases where communication is lucky, we do have the intuition that communication succeeds nevertheless (see my discussion of the Unssteisson cases in the previous chapter). But if my explanation of such intuitions in terms of the sensitivity of knowledge-of-*what is said* attributions to contextual features such as the purposes of the conversation, the linguistic context, etc, is correct, then such intuitions do not show that we have a grasp of the notion of communicative success *independent from whether knowledge of what is said occurred*. Therefore, I want to keep an open mind on the possibility that one could define successful communication in terms of knowledge without reductively defining knowledge.

Although I am impressed by the theoretical justification for a primitivist, anti-reductive approach to what eliminates communicative/epistemic luck, as already announced, I will try pursuing the analysis by appealing to factive mutual states such as joint attention, in the hope of articulating a criterion for communicative success. The element of joint attention, if we manage to analyze it sufficiently, provides a non-circular criterion. Or so I will argue.[11] Appealing to a mutual factive psychological state is already a kind of concession to the anti-reductive argument. I now turn to my second counterexample to ib-features recognition as an anti-luck condition.

3.2 Ascribing too much ib-intention – Peet 2016

Consider the following case, presented in Peet 2016 (it is depicted in Figure 2.2):

> PEET: Smith and Jones are unaware that the man being interviewed on television is someone whom they see on the train every morning. Smith says 'He is a stockbroker', intending to refer to the man on television; Jones recognizes that Smith is drawing upon their common knowledge that there is a salient man on the television screen; but, seeing the similarity between the man on the television and the man whom they often see on the train, he thinks that Smith, *who he assumes also recognizes the similarity*, is talking about the man whom they see on the train. Now, Jones, as it happens, has correctly identified Smith's referent, since the man on television is the man on the train; but he has failed to understand Smith's utterance. (Peet 2016: 3; italics mine)

[11] With their notion of shared sense non-reductively defined *in terms of knowledge upon full understanding*, the sophisticated Fregean discussed in the previous chapter have a definition for which the circle is narrower, and borders on explicit circularity. See my discussion in the last section.

Figure 2.2 – A case where ib-feature recognition is not enough to eliminate luck

In saying 'He is a stockbroker', Smith intends Jones to retrieve the referent as a result of its being common knowledge between him and Smith that there is a salient man in the television screen. And Jones (the hearer) *is* using that information in inferring what Smith is referring to. So far, so good. However, Jones does not stop here. He believes that it is common knowledge between him and Smith that the man on television looks like the man on the train, and thinks Smith intends him to use this similarity to infer that he is referring to *the man on the train*. As a result, Jones deviates from the intended inferential path, and communication turns out to be lucky: it is a coincidence that the man on the train is the man on television.

One solution suggests itself: we may require that the hearer uses the intended ib-feature, *and only this*, in inferring to the intended referent. Let us keep in mind this idea for later.

In the next section, I examine a candidate for the relation of communicative safety which stands as if halfway between the desiderata that *intention recognition* should be non-gettierizable, and the legitimate demand for analysis: joint attention to the intended ib-features.

4 Joint attention on ib-features

Referential intention recognition seemed to be a good candidate for explaining how communicative luck is eliminated. But if my argument by intuitions on cases in the previous section is correct, then intention recognition is still compatible with a Loar case. Here I explore something close, namely, joint attention to ib-features. What motivates this new approach is that joint attention is not gettierizable, thus giving hope for an additional level of communicative safety. In particular, if two persons A and B do not know that they are jointly attending, then they *are not* jointly attending. The reason is that, as I will argue drawing on Peacocke 2005, joint attention has a fixed point character. Joint attention thus provides coreferential safety. By requiring joint attention to every contextual information used in interpreting the utterance, we make sure that the speech participants have common knowledge that the hearer has recovered the correct interpretation. Still, subjects may have the mistaken impression of being in joint attention, when in fact they are not (as it will sound familar by now). Joint attention really is a mutual factive state. As a result, the condition that participants jointly attend on the ib-feature is an *externalist* condition on the causal transmission chain.

The requirement of joint attention on the ib-feature might sound very demanding. But I will try to show that it is less demanding than it sounds. I first introduce the phenomenon of joint attention using the classical notion of *common knowledge*. Then, following Peacocke 2005, I try to propose a characterization that is less intellectualist and more faithful to the perceptual/attentional texture of the phenomenon, the aim of which is to understand *how* participants arrive at common knowledge through joint attention, rather than to explain the latter in terms of the former.

Interestingly, joint attention seems also more *basic* than Gricean intention recognition: indeed, in at least some cases of demonstrative communication, joint attention *constitutes* the recognition of the referential intention. I distinguish two kinds of referential communication: *deictic*, where the object is present in the discourse situation; and *non-deictic*, where the object is not present or not observable in the discourse situation. I explain how joint attention might be used to analyze communicative success in both kinds of communication.

4.1 Joint attention

Joint attention[12] is a central phenomenon in interpersonal psychology. In the context of this chapter, I limit myself to a minimal presentation of the phenomenon. Figure 2.3 depicts a situation of joint attention between two agents:

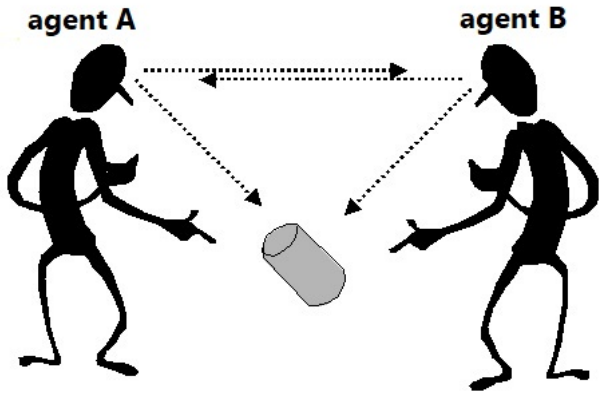


Figure 2.3 – A situation of joint attention

Both agents attend to the cylindrical object placed between them (as depicted by the oblique arrows going from the eyes of each agent towards the object). Moreover, both agents are aware of each other's attention to the cylindrical object placed between them (as depicted by the two-way horizontal arrows between the participants). All this attention (to the object, to each other's attention) is fully transparent to the agents. To a first approximation, putting aside the issue of ascribing an infinite number of embedded mental states for the moment, we may represent schematically the fact that it is 'fully out-in-the-open' to A and B *that there is a cylindrical object* (call this proposition, *p*) as follows:

It is fully out-in-the-open that *p* between A and B:

- A knows that *p*
- B knows that *p*
- A knows that B knows that *p*
- B knows that A knows that *p*
- A knows that B knows that A knows that *p*
- B knows that A knows that B knows that *p*

  . . . etc. etc. *ad infinitum*

[12]What psychologists and philosophers mean with "joint attention" is perhaps better expressed with *joint perception*. Perhaps one interesting exception is O'Madagain & Tomasello 2019 with their notion of 'joint attention to mental content' which is obviously not perceptual. Perception is not the same as attention. I can focus my attention on a belief of mine, but I cannot perceive a belief of mine (in the ordinary sense of "perceive"). And two thinkers may focus jointly on a belief content, but cannot jointly perceive a belief content. For this reason, I find the use of the expression "joint attention" to mean *joint perception* misleading. In contrast to this (I think bad) practice, when I use "joint attention", I mean what I say: joint *attention*. That being said, analyzing *joint perception* will be my starting point. Thanks to Constant Bonard for bringing this to my attention.

We may summarize the infinite list of mental states above in saying that it is *common knowledge* to A and B that there is a cylindrical object placed between them.[13]

One might wonder why we need an infinite iteration in order to capture that it is "fully out-in-the-open" to A and B that there is a cylindrical object between them. A sees the object, and sees that B sees the object. And vice versa with respect to B. Why bother with, say, the condition that *A knows that B knows that A knows* that there is a cylindrical object between them, in order to characterize the shared knowledge at issue?

To see why, contrast the situation of joint attention depicted in Figure 2.3 with the following situation presented by Peacocke:

> GLASS-BARRIER: Consider two people who are standing facing each other, separated by a thick pane of glass. Suppose each person falsely believes that this glass is a one-way mirror, allowing him to see the other, but preventing the other from seeing him. So each really sees the other, while believing the other cannot see him. This is far from having the openness of [joint] attention. Similarly, we can suppose that in this situation, both are attending to something – an animal, say – in their common field of view, off to one side of the glass between them. Each may have a genuine perception of the other attending to exactly the same thing as he is attending to, viz. the animal. But because each believes that the other cannot see him, this too is far from having the openness present in our paradigm cases of joint attention. (Peacocke 2005: 299)

What is missing from GLASS-BARRIER but present in the situation of joint attention above? In GLASS-BARRIER, person 1 knows that person 2 knows that there is an animal; person 2 knows that person 1 knows that there is an animal. However, because each participant falsely believes that the other cannot see him, it is not the case that person 1 knows that person 2 knows that person 1 knows that there is an animal, and vice versa. Therefore, the condition that *A knows that B knows that A knows* that *p* is a crucial ingredient of the situation of joint attention above.[14]

The situation depicted in Figure 2.3 looks very simple. However, perhaps surprisingly, it is not straightforward to characterize the structure of the shared knowledge as it is involved in this type of intersubjective situation. Let us call common knowledge as it is involved in joint attention, *perceptual common knowledge* (PCK in short).[15] What is it? Providing an adequate characterization of joint attention would involve integrating data from developmental and social psychology together within an adequate epistemic-logic formalization, a task beyond this

[13] See Schiffer 1972 for such a recursive approach to common knowledge.

[14] Another nice way to illustrate the need of the whole infinite hierarchy is to consider cases of coordinated attack, where it is clear that nothing less than the infinite hierarchy would be enough for appropriate joint action. See Fagin et al. 1999, Lederman 2017, Campbell 2005, 2017.

[15] Seemann 2019. Let me reiterate, however, that joint attention is not always perceptual. Hence *Attentional common knowledge* might be a more suitable label. I leave this aside for now.

chapter section. Instead, here I would like to flag issues having to do with such a characterization, and provide elements of analysis for further work.

The most immediate idea is to say that PCK just is common knowledge as represented by the infinite hierarchy of knowledge attributions above (an approach endorsed by Schiffer 1972). There are several problems with such a suggestion. A well-known problem is that this approach seems to commit us to ascribing an infinite list of mental states to joint-attenders.

Schiffer tries to answer this probem by showing how the infinite iteration may be obtained via a finite base involving ordinary reasoning only (Schiffer 1972). Similarly, Lewis and others argue that all that is required by the infinite iteration is the fact that agents *could* infer the full iteration from what they currently believe. In other words, the thought of the attributed believed content in a complex embedding does not have to enter the consciousness of the agent for the attribution of the belief to be correct, on this dispositional reading.[16]

However, even if we manage to interpret the infinite hierarchy of knowledge states in terms of *dispositions* to compute recursively from a *finite base* of knowledge states, *infants* might not meet even that finite basis required for common knowledge. But it is well documented that infants can enter in states of joint attention (otherwise they could not acquire their lexicon). In effect, joint attention has been shown to support early word learning.[17] Why, we may ask, is joint attention so helpful for early word learning? One aspect that clearly contributes to bootstrap the child's early lexical acquisition is that if an infant and parent are jointly attending to an object, it is manifest to each what the other is attending to, hence clear to the child which object is being referred to with a novel word (Tomasello 1998, 1999, 2008, O'Madagain & Tomasello 2019). The "Gavagaï" problem is actually not much of a problem for early word learners (Quine 1960 vs Carey & Bartlett 1978, Medina, Snedecker, Trueswell & Gleitman 2011).[18] This suggests that joint attention is a ground for referential intention recognition, and common knowledge of coreference.

To recap: articulating joint attention in terms of *common knowledge* does not do justice to its

[16] As Lewis puts it, the infinite structure of iterated attitudes in common knowledge may be construed as "a chain of implications entailed from our beliefs, not of steps in anyone's actual reasoning" (Lewis 1969: 53).

[17] On the role of joint attention in early word learning, see e.g. Scofield & Berhrend 2011, Pusiol & al 2014, Williams 2016, Akhtar & Gernsbacher 2007, Bruner 1983.

[18] The "Gavagai" problem refers to Quine's example (Quine 1960) of an anthropo-linguist in the field trying to understand the radically foreign language of local native speakers. In particular, the linguist is exposed to an utterance of "Gavagai!" made when a rabbit is present; he has to decide whether the native speaker's utterance means *rabbit*, *undetached rabit parts*, *rabbit time-slices*, *rabbit-hood*, etc. Quine thought this example illustrates that the available evidence about speakers's linguistic behavior *underdetermines* facts about the reference of their utterances.

The problem of assigning the correct meaning of a word-utterance, given that the input a child is exposed to actually *underspecifies* the meaning of the word, has been studied under the name of *The mapping problem*. We now have empirical models how the mapping problem may be solved by children (see references above). Specifically, it has been shown that perceptual social cues having to do with the speaker's gaze are critical for early word learners, e.g. Frank, Tenenbaum & Fernald 2012, Brooks & Meltzoff 2008.

perceptual/attentional and finite nature, in virtue of which it is able to bootstrap the child's early lexical acquisition. What if we try to spell out the shared knowledge in joint attention in its original format, as it were, namely perception?

Even assuming there is a way to analyze knowledge states in the infinite hierarchy in a dispositional way, this is clearly not an option when it comes to *perceptual states*. This is because to perceive something is to be in some *occurrent* state. For a start, perceptual content is clearly not closed under logical operations on its propositional contents: from the facts that A sees that p and A sees that q, and the fact that $p\&q$ manifestly entails r, it does not follow that A *sees* that r (Peacocke 2005: 312). As Peacocke says,[19]

> Someone perceives something to be the case only if in the actual world he is in a conscious state with the representational content of what he perceives to be the case. And it is quite implausible that in all cases that display the openness of joint attention, subjects are in [...] perceptual states with arbitrarily complex embeddings of the *perceives that* operation [...]. (Peacocke 2005: 301)

So the content of perception cannot be plausibly iterated under the 'perceives that' operator in the way that the content of epistemic attitudes for which we have a dispositional notion might be iterated under epistemic operators. For instance, from the fact that A knows that B sees that A sees that B sees that p, it does not follows that A *sees* that B sees that A sees that B sees that p.[20] The *full-in-the-open*-ness caracteristic of joint attention consists in a set of finite perceptual facts, which makes it possible for *infants* to be in joint attention. A infinite list of knowledge states is not the best way to render this.

Joint attention as open knowledge - Peacocke 2005

Peacoke (2005) proposes a characterisation of joint attention which tries to do justice to its perceptual/attentional and finite nature. He proposes that the way in which joint attention episodes are shared involves a kind of reciprocal recognition of the other's recognition of one's own attention. This kind of reciprocal recognition he characterizes as *open-ended*. This is one of the key features in his account, as I understand him. He characterizes joint attention as follows:

> x and y are jointly attending to o iff:
>
> **(a)** x and y are attending to o;
>
> **(b)** x and y are aware that this attention is open-ended;
>
> **(c)** x and y are each aware that they are jointly attending to o, i.e. that this shared complex state of awareness (a)-(c) exists.

[19] In this quote, Peacocke seems to overlook the existence of unconscious perception like blindsight. I would not associate perception and consciousness as he does.

[20] Reduce the iteration to two levels if it does not seem plausible as is. That one might question the plausibility of the three-level iteration of "sees that" is of course another way to make the point I am trying to make in this paragraph.

Condition (a) is obvious. But condition (b) is obscure as long as what it is for an attentional episode to be open-ended is left undefined. So let us try to grasp what this means. (I will comment on condition (c) in due course). What is subject to 'open-ended-ness' is, strictly speaking, not the episode of attention itself, but rather what Peacocke calls its *availability*. What is it for an episode of attention to be available? Peacocke says the following:

> If the obtaining of the state of affairs, and the operation of perceptual and attentional mechanisms in the two subjects, bring it about that one of them perceives that the state of affairs obtains, or bring it about that one of them perceives that the other perceives that it does, or brings it about . . . , then the state of affairs (thus brought about) of his so perceiving is available for the other to perceive. (Peacocke 2005: 302)

Before I try to extract the proposal contained in this passage, let me propose a naive but hopefully useful paraphrase of the passage as I understand it.

Mutual open-ended perceptual availability of a state of affairs between two agents

Let us name p any state of affairs constisting of an entity having properties or standing in relation to other entities, and such that an agent can see (to the naked eye, let's say) that p. Here, very informally, a state of affairs means something like a *possible and perceivable fact*. For example, p could be the state of affairs *that there is a bottle on the floor*. But p could also involve a perceiving agent, for example, in the state of affairs *that an agent sees that there is a bottle on the floor*. I come to the definition.

A state of affairs p which has *mutual open-ended availability* to two agents A and B is such that:

> p obtains, and two agents A and B are co-present to p, in such a way that p, together with the normal workings of perception and attention *in A and B*, cause at least one of the following attentional state of affairs:
>
> - A sees that p – call this state of affairs, q_1 – or,
> - B sees that p – call this state of affairs, q_2 – or,
> - A sees that B sees that p – call this state of affairs, q_3 – or,
> - B sees that A sees that p – – call this state of affairs, q_4– or,
>
> etc. (not *ad infinitum)*
>
> The situation is such that, if q_1 obtains, then B can see that q_1; if q_2 obtains, then A can see that q_2; if q_3 obtains, then B can see that q_3, and so forth (substitute 'can be occurrently aware' for 'can see' when required). In other words, each of the attentional state of affairs $q_1 - q_n$, whenever they obtain, are *available* for the other to perceive/ to the other's occurrent awareness.

> When such conditions are met, we may say that *p* has mutual open-ended availability to the agents A and B.

Still in other words, a state of affairs *p* which has open-ended mutual availability to two agents is such that *p* and the perceptual/attentional mechanisms in the two participants *cause* the fact that one of the participants is attending to *p*, or, that one of the participant is aware that the other is attending to *p*, etc., and *that* latter fact is such that the other participant can attend to it.[21]

Now, according to Peacocke, it is not enough for a situation to be one of joint attention that the attention of each participants on the target (or on the attention of the other on the target, etc) be mutually and open-endedly available. In addition, participants must be aware of just *this* fact.

What is it to be aware? Appreciating this point will enable us to see why the Peacockian characterization is infans- inclusive. *Awareness* in the sense at issue (as I understand it), is a non-propositional state, the sort of awareness one might have of being cold or hungry before thinking *I'm cold* or *I'm hungry*. This awareness is such that one need not put it into words in order to have it.[22]

So a situation between two thinkers is jointly attentional only if the state of affairs of one thinker's attending to the target is mutually and open-endedly available, and both thinkers are aware of that. Why not stop here? Why does Peacocke think we should add condition (c) to a definition of joint attention, to the effect that the participants are aware that they are jointly attending?

This is because joint attention has a so-called *fixed-point* character.[23] Said differently, full joint attentional awareness *is* the awareness of the full joint attentional awareness. Because of this, joint attention is transparent to the participants, in the following sense: if A and B are jointly attending to *o*, then A and B are aware that they are jointly attending to *o* (where *being aware* is of course factive). This does not mean that participants cannot fail to detect that they are *not* in joint attention. In other words, two participants *may* believe that they are jointly attending to some object when in fact they are not (as it will now sound familiar).[24]

[21] I use "fact" merely to make the text easier to follow.

[22] I compare this notion with what Schroeter 2012 calls "mutual appearances of meaning sameness" in chapter 5. Awareness in the sense at issue might be a functional notion e.g. like *access consciousness* (Block 1995). For recent attempts to operationalize joint attention involving "social" robots, see e.g. Huang & Thomaz 2010, Chevalier & al 2020, Huang 2010 *ms*.

[23] In mathematics, a fixed point of a function is an element of the function's domain that is mapped to itself by the function. That is to say, c is a fixed point of the function f if $f(c) = c$. On the notion of a fixed point as applied to common knowledge, see Harman 1977 and Barwise 1988 cited in Peacocke 2005. See Lederman 2018 for a survey on theories of common knowledge.

[24] John Campbell proposes the following idea:

> (...) when there is another person with whom you are jointly attending to the thing, the existence of that other person enters into the individuation of your experience. The other person is there, as co-attender, in the periphery of your experience. (Campbell 2005: 288)

Observe the important difference between *mutual open-ended availability* and *full joint attentional awareness* in Peacocke's definition of joint attention. Mutual open-ended availability can occur in a situation without awareness of its occurrence. However, joint attention cannot occur without awareness of its occurrence (Peacocke 2005: 303). This is condition (c) in Peacocke's definition.

Where does this leave us? In virtue of the fixed-point character of joint attention, we get what Peacocke (2005) calls *open knowledge*, namely:

open (perceptual) knowledge

> x and y have open perceptual knowledge that p iff:
>
> **(i)** x and y both perceive that p;
>
> **(ii)** x and y are both aware that their perceptions that p are mutually open-ended;
>
> **(iii)** x and y are aware that they are both aware of this very awareness (i)-(iii) (*as per the fixed-point feature)*.
>
> (Peacocke 2005: 313)

Open knowledge does not have to be perceptual: it can take as targets contents arrived at by inferential means. We get open knowledge *simpliciter* by adding to the open perceptual knowledge of participants what they mutually know about each other's inferential procedures and epistemic capacities. For instance, if A sees that B sees that p, then A *knows* that B sees that p. This kind of knowledge is fully perceptual. But given this, if A grasps (perhaps in some non-propositional state) that *seeing* is a form of knowledge, then via an easy inference, A knows that B *knows* that p.

The differences between open knowledge and common knowledge, and the implications of these differences, would require careful consideration – something beyond the scope of this

As I understand him, Campbell wants to commit to a form of disjunctive theory of joint attention, according to which, roughly:

> **Disjunctivism about joint attentional experiences:**
> Conscious experiences that are involved in cases of joint attention cannot have the same nature than conscious experiences involved in solo attention.

However, it is a common experience to believe that one is in joint attention, whereas in fact one is a solo attender. I start watching a movie with a friend. Five minutes later, I turn to him to share a furtive connivance about the movie scene. But my friend is no longer here. In this scenario, it seems to me that my conscious experience of the movie, at the moment when I mistakenly thought I was in joint attention with my friend, would have been the same if my friend had been there. In particular, I find the following claim plausible: The counterfactual conscious experience (if my friend had been there) would have had the same phenomenal character, and the same intentional content, as my actual conscious experience. We don't have to disjoint them.

section (but see Peacocke 2005: 309-316). One obvious difference is that common knowledge is (at least partly) dispositional, whereas open perceptual knowledge is *occurrent*. Another feature I would like to emphasize is that open perceptual knowledge, unlike common knowledge, is directly given by the perceptual/attentional texture of joint attention. Figure 2.4 depicts how these different sorts of shared knowledge relate.

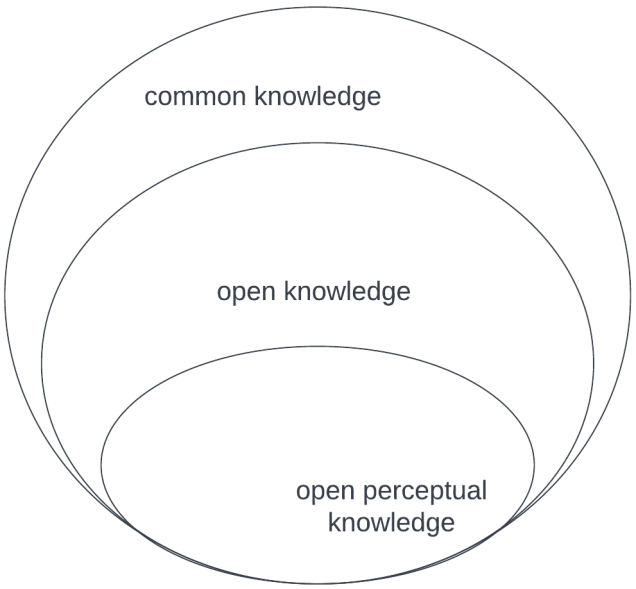


Figure 2.4 – Open perceptual knowledge, open knowledge, common knowledge

This is the end of my presentation of joint attention as open knowledge. Let us see how we may apply this characterization to linguistic communication, and the problem of communicative luck.

4.2 Joint attention on the referent

Joint attention has a clear role in demonstrative communication, i.e. communication in which speakers perform speech acts through the utterance of sentences that contain demonstrative expressions such as "this", "that", "these", "those", "he", "she", "it", "this *F*", "those *F*s", etc., and which may or may not be accompanied by demonstrative gestures such as pointing. When the object under discussion is present in the discourse context, it is available for demonstrative reference to the extent that *its presence in the context* has mutual open-ended availability to the speech participants (reusing Peacocke's useful terminology). Here is an example of demonstrative communication (depicted in Figure 2.5):

BUTTERFLY: **Bob** Look at this butterfly! How beautiful it is!

Anna Yes! It is a morpho!

Figure 2.5 – A case of demonstrative communication

In BUTTERFLY, the participants need not be jointly attenting to the object *prior* to the conversation. We may suppose that *Bob* was able to initiate an episode of joint attention on the object through his utterance. Nevertheless, the presence of the object had *mutual open-ended availability* to the participants prior to Bob's utterance. How should we describe Bob's communicative intention? And what does it take for Anna to recognize it?

In uttering the complex demonstrative "this butterfly", Bob manifests his belief that the object he is demonstrating *is a butterfly*, thereby helping Anna to identify the demonstrated object. The property of *being a butterfly* is thus rendered salient to Anna for tracking Bob's demonstrative intention. Remembering the sophisticated Gricean view of intention recognition, we may say that Bob intends Anna to think of the object on the basis of her recognition that *there is a butterfly in common view*.

Even if Bob is not actually pointing at the object he is referring to, the direction of Bob's *gaze* is part of the information Anna must be sensitive to in her recognition of Bob's demonstrative intention. It is an attentional requirement for understanding Bob's utterance; the direction of Bob's gaze is part of the intended ib-feature of Bob's utterance, as it were. However, the monitoring of the direction of Bob's attention by Anna need *not* involve personal-level perception and judgement about Bob. She may succeed in synchronizing her attention with Bob's without *explicit* intention or planning. This is a sense in which joint attention to the referent may be

said *more basic* than Gricean intention recognitions.[25]

What is it that makes joint attention an *anti-luck* mecanism here? Let us break down the joint attention episode into its constituent elements. First, Anna is successfully monitoring the direction of the speaker's attention. In bringing her attention into line with Bob's attention, she is letting (again, perhaps subpersonally) her own attention to be *causally controlled and sustained by Bob's attention*. As a result, it is not a coincidence that Anna and Bob's demonstrative thoughts corefer: their thoughts refer *together*. Because they are aware that they are jointly attending to the butterfly, Anna and Bob know in common which object Bob is intending to refer to and communicate about: communication is successful.

As soon as Anna recognizes Bob's demonstrative intention, she thereby achieves *open perceptual knowledge* with Bob that there is a butterfly in common view, namely:

- Anna and Bob both perceive that there is a butterfly;
- Anna and Bob are both aware that their perceptions that there is a butterfly are mutually open-ended;
- Anna and Bob are both aware that they have open perceptual knowledge that there is a butterfly.

In a sense, Anna and Bob's open perceptual knowledge described above *just is* the fact that the communication of 'Look at this butterfly!' was successful. It is not a further stage in the communicative episode, even if we can distinguish them intellectually.

Anna's utterance ('Yes! It is a morpho!') extends Anna and Bob's open knowledge about the butterfly (assuming Bob is endorsing what Anna says), to the effect that it is a morpho. Here too, Bob and Anna have common knowledge that Anna's thought and Bob's thought corefer. Anna and Bob are disposed to trade upon the coreference of their respective token singular terms, and they are warranted in doing so.

When open perceptual knowledge is achieved by two thinkers as in BUTTERFLY, it might be tempting to say that the thinkers *share a distributed file* on the object. Clearly, such distributed files exist. In the context of demonstrative communication, what it takes for a distributed file on a given object o to exist is that there be mental files referring to o and whose thinkers have open perceptual knowledge that they are perceiving o. However, if my argument against the sophisticated Fregeans in the previous chapter is correct, such distributed files are not transparent in the sense that thinkers may wrongly assume that they are in joint attention or joint awareness on an object. As we shall see[26], this lack of transparency associated with distributed files is even more dramatic in the case of communication involving proper names, where the

[25] See Campbell 2017 where this line of argument is developped further.
[26] See chapters 3 & 4.

anaphoric path spans contexts and unfolds over a very long distance in space and time.[27] Distributed files are very much like concepts construed as *shared vehicles* of the originalists (see Sainsbury & Tye 2012, 2011, Recanati 2016). If our aim is to understand the role open knowledge has to play in communication within ordinary psychological theory, I believe we should be reluctant to allow that part of the explanation of open knowledge involves thinkers' sharing distributed non-transparent mental states.[28]

Sperber & Wilson 1996 criticize the idea that common knowledge is necessary for successful communication, because they think common knowledge is never reached:

> If mutual knowledge is necessary for communication, the question that immediately arises is how its existence can be established. How exactly do the speaker and hearer distinguish between knowledge that they merely share, and knowledge that is genuinely mutual?[29] To establish this distinction, they would have, in principle, to perform an infinite series of checks, which clearly cannot be done in the amount of time it takes to produce and understand an utterance. Hence, even if they try to restrict themselves to what is mutually known, there is no guarantee that they will succeed. (Sperber & Wilson 1996: 18).

However, if the proposed analysis of joint attention is correct, in simple cases like BUTTERFLY, where open perceptual knowledge of the presence of the object is achieved by jointly attending, speech participants are not merely *justified* in assuming common knowledge. They are *aware* that they have open perceptual knowledge of the presence of the object in common view (a factive state). If, like Anna does with Bob, further knowledge about the object is successfully communicated, then they have extended their open knowledge about the object.[30]

Crucially, in simple cases like BUTTERFLY, joint attention is the *ground* of the open knowledge about the object achieved by the participants. As already suggested, when the object is present in the discourse situation, successful demonstrative communication *is* joint attention, because the speaker's demonstrative intention is fulfilled by jointly attending to the target. More cautiously, it seems that, in demonstrative communication, joint attention may explain how communicative luck is eliminated, thereby explaining how common knowledge of coreference is achieved (*pace* Sperber & Wilson).

[27] For a systematic comparison of proper names with pronouns, see Cumming 2007.

[28] However, as I shall explain in chapter 5, we can, and perhaps we must, recognize a *normative* role for these distributed mutual mental states in a theory of samethinking.

[29] It is clear from this quote that what Sperber & Wilson mean by "mutual knowledge" is what I mean by "common knowledge".

[30] It might be thought that insisting on factivity as I do push me towards the disjunctive theory of joint attention I reject. However, one motivation I have to reject that a sense is shared in joint attention is precisely that senses so defined are not transparent. I am thereby resisting the disjunctive view. More on this in due course.

4.3 Two kinds of referential communication

We may distinguish two kinds of referential communication. We have already met a first kind of referential communication in this chapter, namely, communication about an object which is present and observable in the discourse situation. We may call this kind of referential communication, *deictic communication*. In deictic communication, as we have seen, it is open for us to analyze successful communication by appealing to joint attention (and the related notions of *mutual open-ended availability* or *open knowledge*) on the object under discussion — typically, a salient object in common view.

But this kind of communication does not exhaust, far from it, human referential communication. Indeed, a great feat of the human language, as opposed to animal communication, is that it may be used to refer to what is not present or observable in the communicative situation (Descartes 1646, Chomsky 1966). This is a second kind of referential communication; we may call it *non-deictic communication*. This label is merely a label. In particular, I don't mean to deny that there are indexical aspects to what I am calling *non-deictic* communication: first and foremost, ib-features.

Non-deictic communication (in the sense at issue) is widespread. Thus, humans communicate about dead people (such as Plato, for us now), far way referents (both in space and time), non-observable object or kinds such as *blockchain* or *quark*, fictional referents such as *Sherlock Holmes*, or abstract referents such as *numbers*, social or cultural kinds,. . . , and even more unassignable referents.[31] Regarding communication about existent ordinary objects, but absent from the situation of discourse, it should be noted that even in this case, a token singular term may still be causally related to the referent — or (in the case of fictional object) to the author(s) of the fictional referent — via causal-historical communicative chains (on which more in chapters 3, 4 and 5).

Could we use joint attention to explain communicative success in non-deictic singular communication, namely, in which the intended referent is not present or observable in the discourse situation? I examine this in the next subsection.

4.4 Joint attention on ib-features

We are seeking an account of communicative safety in referential communication in general, not just in cases where the object is present. It would be nice to keep the generality of the intentionalist approach, while trying to have the safety of the approach in terms of joint attention. On the face of it, the role of joint attention in making communication non-lucky (i.e. safe) seems limited to discourse situations where the presence of the intended referent has mutual

[31] It is controversial whether communication about all in this (non-exhaustive) list counts as (pseudo)-referential communication. In this thesis, I mostly focus on ordinary objects or substances such as Noam Chomsky, a bottle, or water.

open-ended availability between the speech participants in the discourse situation. This might not necessarily require the presence of the object *in the flesh*, as e.g. a photo of the object may also provide mutual open-ended availability of the referent and be the target of a joint attention. But people may be jointly attending to a photo and yet, identify it differently (perhaps identifying different people).[32] So it is not straighforward which role joint attention could have in communication where the referent *itself* is not present.[33]

However, the gap may not be as wide as one might think between the two types of communication (deictic and non-deictic), for the following reason. In cases where the referent is present, it is not as if the hearer could transparently attend a demonstrative intention of the speaker without *any* kind of instruction. Rather, there will be some kind of e.g. attentional requirement to monitor the direction of the speaker's gaze, or, in the cases of deictic communication which involve a complex demonstrative, a requirement to track the property expressed by the nominal, and so on and so forth. This much is familar due to our discussion of ib-features recognition.

If ib-features are involved even in deictic-demonstrative communication, then *a fortiori* they are involved in non-deictic communication. If that is correct, then I suggest that joint attention on ib-features may be used as a uniform candidate criterion for communicative success in both types of communication.[34] We may call it *IB joint attentional criterion*. I consider this criterion in this section, and the next.

The idea behind the proposal is that, by requiring joint attention on every element of contextual information used in interpreting the utterance, we get speech participants to have common knowledge that the hearer is recovering the correct interpretation, and luck is eliminated. Further evidence for the IB joint attentional criterion comes from the linguistic theory of the *Givenness Hierarchy* (GH), which I present (for a different purpose) in Chapter 3. This theory

[32]The same remark applies with respect to objects in the flesh, however. If it is true that one may entertain a *de re* thought on an object based on the perception of a photo of it, then at least in some cases it won't matter that different thinkers may construe a jointly given photo differently. The same is true of proper names (which, if a causal theory of proper names is correct, are not unlike photos).

[33]Drawing inspiration from Dretske 1969, 1995, we may distinguish between joint *simple* perception, and joint *epistemic* perception. Now consider the following thesis:

> **Joint epistemic perception conceptualism**: For any object x and any property F, two subjects may jointly perceive that x is F only if they both have concepts of x and F, and deploys those concepts in the joint perception episode.

If *Joint epistemic perception conceptualism* is true, then two thinkers may jointly perceive a state of affairs in the *simple* sense but fail to jointly attend to it in the sense of joint epistemic perception. Moreover, it is an interesting question whether joint epistemic perception involve antecedent conceptual sharing, or something weaker.

One may try to appeal to *Joint epistemic perception conceptualism* to account for *meaning* perception. The thought would be that, for example, a non-French speaking person *NF* cannot jointly attend that *S* said that *p* with a French speaking person *F* if *p* was said in French, as *NF* does not possess the relevant lexical concepts. *NF* might still be able to jointly attend with *F* to the utterance *qua* string of sounds i.e. in the sense of joint *simple* perception.

[34]See Peet 2016 where this suggestion is mentioned. See also Peacocke 2005: 314-316, and Campbell 2017, where a similar idea is articulated (with various differences between them, and between each of them and the account I will propose here).

supports the existence of an interesting class of *inference-based* features attached to linguistic expressions (such as pronouns or determiners) that indicate whether a referent is present in the common ground and its degree of accessibility in the memory/attentional states in the hearer's mind—as assumed by the speaker. Therefore, (GH) supports the notion that discourse participants can have common knowledge that the hearer is recovering the correct interpretation as a result of jointly attending to these IB-features.[35]

Before I introduce the joint attentional criterion I have in mind, I would like to flag, roughly, some foundational questions about psychological properties related to ib-features cognition: how ib-features are represented, what are their contents, and whether ib-features cognition is always perceptual.

Earlier I said that ib-features can be about anything, from the fact that a speaker uttered such and such words, with such and such meanings, in a given shared language, to virtually *any* feature of the extra-linguistic context of an utterance (see e.g. Schiffer (forthcoming a,b), and Buchanan 2013).

What unites this otherwise very diverse class of possible cues is that an ib-feature, whatever its nature, must be accessible to the subject for use and guidance in recovering what is said. Hence ib-feature recognition is always occurrent. However, it is implausible that ib-features of all sorts be *explicitly* represented in consciousness. In particular, the idea that *purely semantic* ib-features are *typically* a part of the experience of interpreting an utterance is highly dubious, as dubious as the claim that speech participants typically experience minimal propositions of sorts when interpreting utterances. Here is a schema to illustrate this difference (Figure 2.6).

In characterizing the nature of ib-features cognition, and in particular the type of transition between thoughts it instantiates, we may draw inspiration from the work of epistemologists trying to characterize the relation which holds between a reason and a belief if and only if the reason is a reason for which the belief is held[36]. As far as we are concerned, the reason is *the occurrent grasp of an ib-feature*, and the belief arrived at is *the grasp of what was said*. Note the difference between the aforementioned relation (sometimes called *epistemic basing relation*) and epistemic justification: a thinker may epistemically base a belief on another, without being epistemically justified in doing so. I favor a certain causal theory of the aforementioned relation, due to Moser 1989:

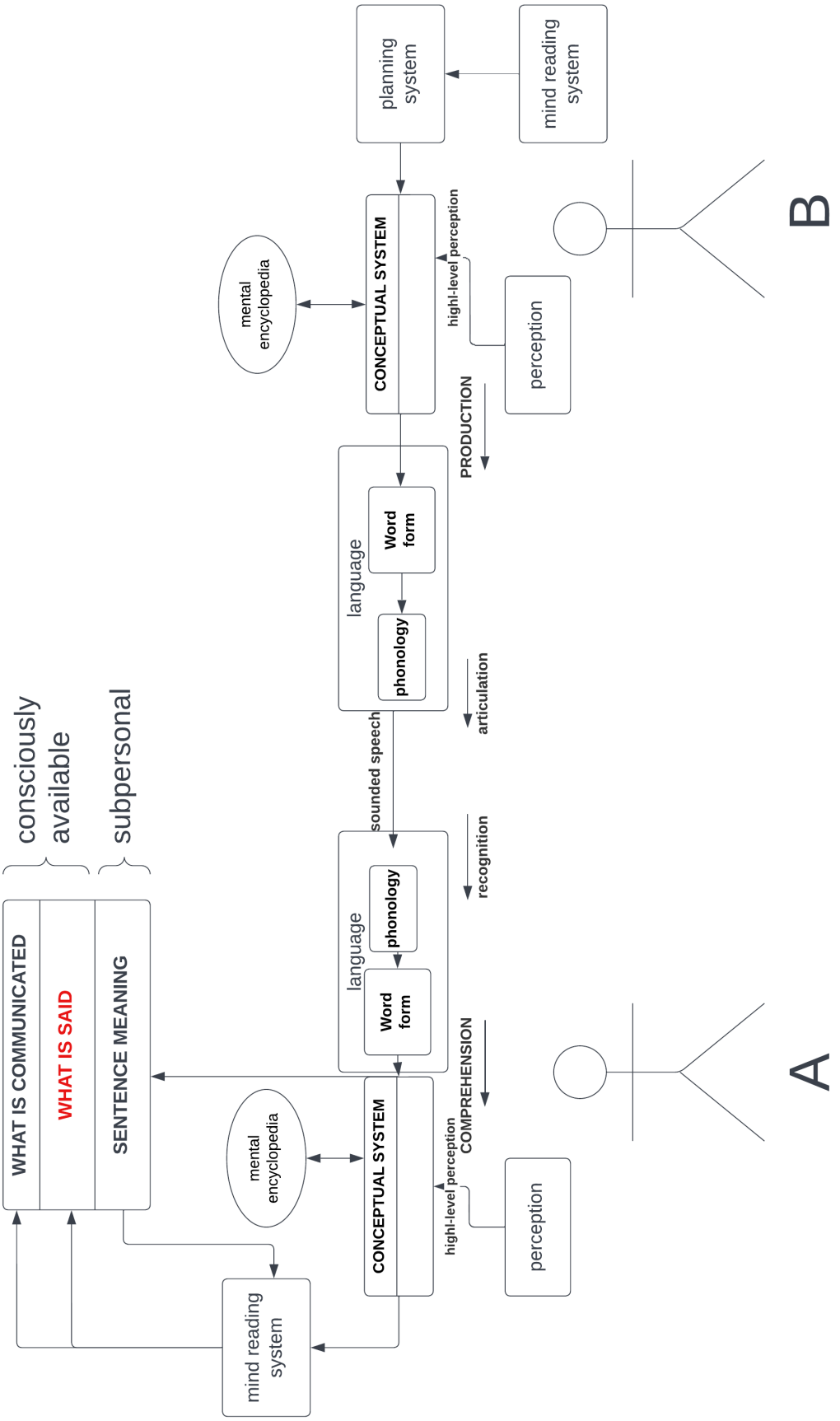
> S's believing P is based on justifying reason Q [=is grounded and justified by] iff S's believing P is causally sustained in a nondeviant manner by his believing Q, and by his associating P and Q.[37]

[35]The conception of the CG offered by (GH) is thus much more rich and structured than the standard Stalnakerian conception, introduced in section 3 of the general introduction. This is of course in line with my proposal.

[36]See Korcz 2021 for an overview.

[37]Such accounts are generally criticized because "in a non deviant manner" has proven difficult to spell out. Although see Peet 2019 for interesting comments on this issue.

Figure 2.6 – A model of communication with cognitive architecture from B to A (I draw inspiration from Recanati 2004, Hickok 2013, and Harris 2019)



. . . where the association relation is defined as follows :

> S occurrently satisfies an association relation between P and Q iff **(i)** S has a *de re* awareness of Q's supporting P and **(ii)** as a nondeviant result of this awareness, S is in a dispositional state whereby if he were to focus his attention only on his evidence for P (while all else remained the same), he would focus his attention on Q.
>
> (Moser 1989: 141-142 modified by me, cited in Korcz 2021)[38]

I suggest that the basing relation as defined is what we need for capturing the relation between the grasp of an ib-feature and the output of the interpretation process. I only consider the occurrent version of the association relation (see above), because in any episode of communication we may suppose that an ib-feature will be either represented in short-term memory or else at the current centre of attention. Having mentioned issues about some aspects of ib-features cognition, I now come to the definition that I would like to put forward.

4.4.1 IB joint attentional criterion of communicative success

I will propose two versions of the criterion. The first version is in terms of intention recognition. The second version is essentially the same definition as the first but adding a characterization in terms of the mental files. Here is the first version

> **Successful communication****
> In face-to-face communication, a hearer H understands a singular term v as used by a speaker S iff
>
> **(a)** S intends to refer to o by v and ib-feature Ψ of v;
>
> **(b)** H and S jointly attend to v and ib-feature Ψ of v;
>
> **(c)** H interprets v through the jointly attended ib-feature Ψ, and only the ib-feature Ψ;
>
> **(d)** H thinks of o as a nondeviant result of (b) and (c).

I have already reviewed condition (a) in the section on ib-features recognition (section 2.2), indeed both accounts have this condition in common. Likewise, conditions (c)+(d) are very much like conditions (b)+(c) of the criterion in terms of ib-features recognition (section 2.2). Note two important differences, however: in order to address the PEET type of Super-Loar cases, I add the condition that the indended ib-feature, *and only it*, must govern the hearer's interpretation. Moreover, joint attention is factive, whereas ib-feature recognition is gettierizable (or so I have argued in section 3 above; the intentionalist is of course free to define ib-intention recognition in terms of joint attention, and more than welcome to do so).

[38] See also Korcz's own proposal in the same spirit in Korcz 2021.

Condition (b) is to be explained. An ib-feature is a way to select a context and use it in utterance interpretation as required by the referential plan of a speaker. The condition (b) demands that speech participants be jointly aware of the relevant way to use the context in the interpretation of the utterance. The criterion is externalist: sometimes, speech participants may fail to distinguish by introspection alone ('from within') which information is commonly known, and which is not. Remember PEET: Jones is jointly attending to the intended ib-feature with the speaker. But then he interprets the utterance using some other information he mistakenly thinks is part of the inferential plan. Sadly, this is the risk of any communication exchange whatsoever. However, in real-life communication, we routinely and smoothly achieve open knowledge of coreference with our interlocutors: referential certainty is not a remote Cartesian ideal, but common practice.[39]

[39]One may object that my criterion *does* make successful communication a remote Cartesian ideal, on the grounds that the IB joint attentional criterion is too demanding. In the words of Cappelen and Lepore (2005: 123 cited in Perini-Santos 2009 who himself responds to Cappelen and Lepore's objection), the objection is that my account would make (if true) successful communication *miraculous*. There is a lot to be said for why this is not the case. One conception of communication which I have not engaged with here, but which is useful in addressing the miracle objection, is the *action* tradition. Issued from Clark, it views language use as a form of *joint action* (Clark 1996). Perini-Santos 2009 nicely summarizes the vision of this approach:

> While we should grant the robustness of communication, it is not guaranteed by some unchanging conditions, but by different flexible mechanisms that enhance the chances of mutual understanding at a relatively low cost — this is true, in particular, of different feedback mechanisms and of alternative ways to make the same information mutually available. Communication is not a series of successive, individual and independent actions; dialogues are a kind of joint activity in which misunderstandings are jointly repaired by participants as part of the very activity they are engaged in. (Perini 2009: 1)

The related literature on *conceptual pacts* (partner-specific temporary lexical conventions to label objects), and *acceptance cycle* (a collaborative testing of mutual understanding) is strongly relevant here. The classic paper on the notion of a conceptual pact is Brennan & Clark (1996). See e.g. Clark 2020 on the notion of acceptance cycle. There are many relevant elements in Perini-Santos 2009, where the aforementioned notion is described as follows:

> Communication exhibits systems of feedback control and of redundancy of information that help to assure mutual understanding. *Acceptance cycles*, proposed by Herbert Clark and co-workers, are systems of positive and negative feedback: participants in a conversation make efforts to establish the mutual belief that listeners have understood what is meant by the speaker. If the listener doesn't see what object is aimed at by the speaker, she will indicate it, and the speaker is expected to propose a new presentation, until the listener gives an acknowledgement sign, followed by a confirmation by the speaker. (Perini-Santos 2009: 239-239; my italics)

As I see it, this theoretical outlook on communication as joint action is fully compatible with the approach presented here. Integrating it into the current approach would be fruitful; that is for another occasion.

I suggest that we may add substance to the discussion by adding the theoretical commitments of the mental file theory of singular thoughts — a particular view of the non-descriptive MOPs introduced in the previous chapter, which I will discuss at greater length in the next chapters of the thesis. I will not elaborate on these additional file-theoretic commitments in this chapter, but I use them later in the thesis, so it is good to see them as of now. Here is a *mental file version* of the definition:

> **Successful communication**(mental file version)**
> In face-to-face communication, a hearer H understands a singular term v as used by a speaker S iff
>
> **(a)** S intends to refer to the object represented by her file $\mathcal{M}_{\mathcal{S}_\mathcal{O}}$ in using v and ib-feature Ψ of v;
>
> **(b)** H and S jointly attend to v and ib-feature Ψ of v;
>
> **(c)** H interprets v through the jointly attended ib-feature Ψ, and only the ib-feature Ψ, by either retrieving, or opening a file $\mathcal{M}_{\mathcal{H}_\mathcal{O}}$;
>
> **(e)** $\mathcal{M}_{\mathcal{S}_\mathcal{O}}$ and $\mathcal{M}_{\mathcal{H}_\mathcal{O}}$ corefer as a non-deviant result of (b)-(c).

The diagram in Figure 2.7 represents the mental file version of the proposed criterion for communicative success.

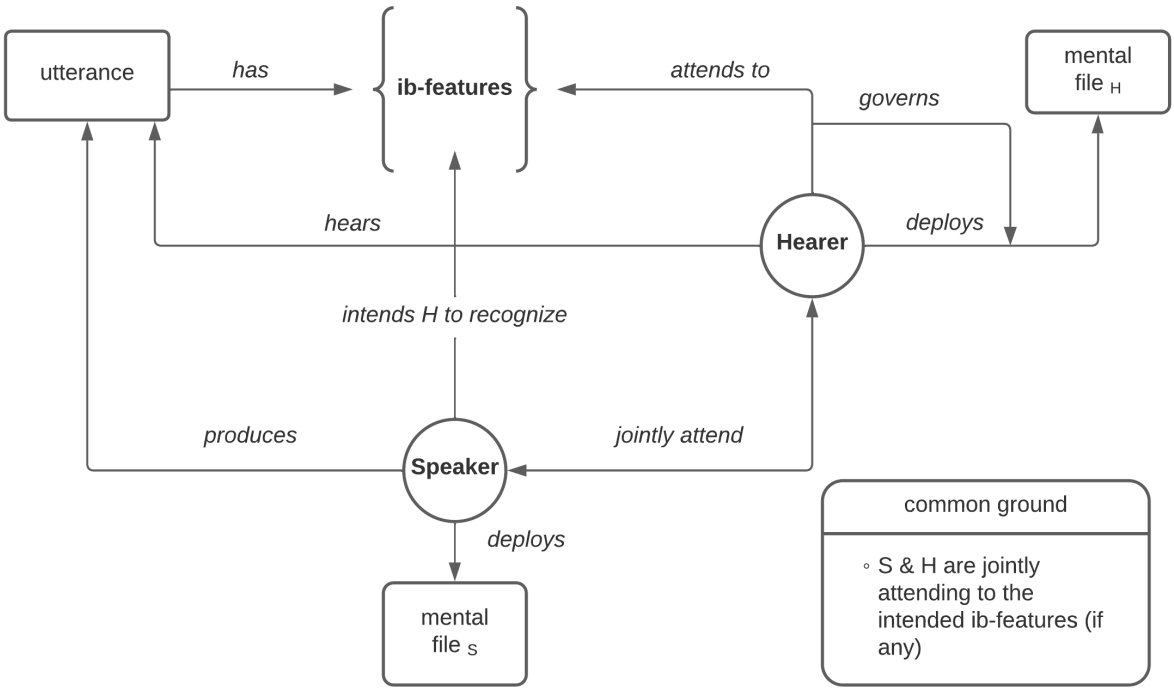


Figure 2.7 – IB joint attentional criterion for communicative success (mental file version)

Let us see how the proposal works on the problematic cases. In TAYEBI (Figure 2.1), Leo Peter attends to the intended ib-feature, namely, the information that *the referent is the utterer of this utterance*. Dr. Lauben also attends to the ib-feature. However, Leo Peter falsely believes he is *overhearing* an utterance which was not intended for him, whereas the utterance is in fact for his attention. As a result, it is not the case that the speech participants are both aware that their attention to the ib-feature is mutually open-ended. Hence they are not jointly-attending to the ib-feature, and communication fails. (The case is similar to my SUPER-LOAR CASE 2 in the previous chapter). In PEET, the joint attentional criterion is violated because it is not the case that Jones' interpretation is wholly governed by the jointly attended to ib-feature. Hence communication is lucky. In the next section, I raise two problems for the joint attentional approach to communicative safety.

5 Problems with joint attention on ib-features as an anti-luck condition

5.1 Non-face-to-face communication

There are many cases of communication for which *joint* attention on the ib-features is not necessary. Perhaps the most obvious case is *written* communication: in typical cases, the author and the hearer are simply not present together. That is true. The proposed criterion only applies to face-to-face communication (possibly including e.g. communication on the phone, by video call, or through avatar characters in the metaverse, and the like). In response, I will say two things. First, face-to-face conversation is, in an important sense, primary. This is through face-to-face conversation that a first language is acquired.[40]Second, even in written communication, it could be argued that we *simulate* as far as possible the contextual elements needed to interpret the written utterances. To illustrate, take Plato's dialogues. Plato believed that writing in the form of dialogues, and staging more or less concretely the situations of utterance, was better than normal prose for transmitting thoughts to people. Plato used written text as if it was spoken language, to maximize transmission. When we read the dialogues, we continuously simulate contextual ib-features (not unlike reading theater), by exercising the skills one is used to in oral communication.[41] What is primary is face-to-face communication, where joint attention *is* required.

However, even in some cases of *oral* communication, joint attention is not required. I have in mind oral communication which is not face-to-face, as when a subject overhears someone

[40]Here is a relevant quote from Perini-Santos 2009 citing Clark 1996:

> The basic setting for language use, as Herbert Clark says, is face-to-face conversation: "it is universal, requires no special training, and is essential in acquiring one's first language."—Clark (1996): 11.

[41]Another widely used kind of philosophical communication in ancient Greece was the epistle, usually directed at an acquaintance or friend of the writer.

soliloquizing, who is not aware of the hearer's presence. Consider the following example, presented in Campbell:

> Suppose that I am hiding in the bushes as you come out into the moonlight, and as you look around, you soliloquize. You use demonstratives in your soliloquy, referring to, for example, 'that star'. There seems to be no reason why I can't understand what you are saying and know what you are thinking; I can see the star myself, and I know that it is what you are talking and thinking about. Of course, there is a sense in which your perspective on the star will be a bit different from mine, since you are seeing it from a different position; but I can compensate for that, either by imagining how it would look from your perspective or by explicit reasoning about what you can see. (Campbell 2005: 290)

I think there is a noteworthy distinction between face-to-face communication, in which a speaker addresses a *particular* audience in a *particular* way, and a situation like the one described in the quote. Among what characterizes the *specifics* of a face-to-face communicative exchange is the ib-feature for the attention of a *particular* listener. For any utterance u, I suggest, we may distinguish the conditions for understanding related to the *specifics* of the communicative exchange between the *actual* speech participants as they are related to each other in the situation of use, and more 'generic' conditions for understanding u, with respect to any possible audience who may overhear the utterance – i.e. conditions on which a third party could understand u in *overhearing* the communicative exchange between the actual participants (or the soliloque as in the quote) without jointly attending to the ib-feature. In TAYEBI, the hearer is not understanding the speaker's utterance, because he believes he is overhearing an utterance which was not intended for him (where, in fact, what he is interpreting is an utterance for his attention). This illustrates that one may understand an utterance *in the generic sense* while failing to understand it *in the specific sense*.[42]

[42] I find a similar distinction in Perini-Santos 2009, who writes:

> The hearer can either be overhearer or a certified participant. If the hearer is an overhearer, it may be the case that he easily misunderstands what is said, but it has no consequence to the understanding of what takes place in a dialogue, since, ex hypothesi, he is not a party to it. If the hearer is a participant, many of the conditions of mutual understanding will be assured in the dialogical activity itself, and both he and the speaker will make efforts to assure that mutual understanding do take place. (Perini-Santos 2009; opening page)

5.2 Comparison with the Sophisticated Fregeans

On the face of it, the *IB joint attentional criterion* looks very similar to the criterion in terms of *shared senses* put forward by the sophisticated Fregeans discussed in the previous chapter — a criterion I have argued against. Recall Dickie & Rattan's view:

> **Individuating Principle for Senses** — The sense of v (used by S_v at t_v with MOP M_v) = the sense of μ (as used by S_μ at t_μ with MOP M_μ) iff the engagement-relevant factors *in the situation of use* generate the possibility of the immediate extension of knowledge upon full understanding. (Dickie & Rattan 2010: 150; italics mine)
>
> **The Fregean view** — Speakers can communicate using a term iff they attach the same sense to it.

It is true that the approach in terms of the quoted *individuating principle* and the approach in terms of joint attention on the ib-feature are very close to each other. Thus, Dickie & Rattan 2010 explicitly cite joint attention as one of the factors involved in determining whether trading upon coreference is warranted for the thinkers (in an externalist sense). However, there are important differences between Dickie & Rattan's view and the *IB joint attentional* view. I will argue that theses differences are in favor of the criterion I have presented here.

One difference is that on my view, shared contents finer-grained than reference such as 'senses' *are not needed* in order to explain the coreferential safety required by communicative success. By constrast, the sophisticated Fregean view is not a referential view of communication. Sophisticated Fregeans believe we *need* to postulate *shareable senses* individuated in terms of the possibility of *the immediate extension of knowledge upon full understanding*. They individuate them in terms of the equivalence classes of MOPs collected by (an abstract interpretation of) the external relation of rational engagement, as per their individuation principle. But two thinkers may be *rational* in mutually presupposing that their thoughts co-refer, even if their presupposition turns out to be unwarranted because the world does not cooperate. As already noted in the previous chapter, there is an ambiguity attached to what Dickie & Rattan call *rational engagement* between the MOPs of two thinkers. On one construal of this notion, 'rational engagement' refers to the disposition in both thinkers to presuppose that they are currently thinking about the same object. I proposed to call the notion of rational engagement interpreted in this way, COORDINATION. This relation is needed for making sense of thinkers' linguistic and non-linguistic behavior. When COORDINATION obtains between the thoughts of two thinkers, even if their trading upon coreference turns out to be *not* valid, we can rationalize their behavior. Hence COORDINATION is a perfectly sound notion of *rational engagement*, which enables to rationalize psychological explanation.

COORDINATION, unlike the relation of rational engagement construed in a way that involves

external factors such as whether joint attention obtains (which I called SUCCESSFUL CO-ORDINATION in the previous chapter), can be reduced to the conjonction of the thinkers' individualistic states. To determine whether COORDINATION obtains between two thinkers, there is no need to determine whether joint attention occurs. Dickie and Rattan seem to be mixing the relation necessary for rationalizing psychological explanation (for the explanation of which there is no need for shared senses), with a world-involving external relation. But as already argued in the previous chapter: if the relation of *same-sense* is external, the sophisticated Fregeans will have to buy the idea that introspection is merely reliable with respect to identity and difference of sense. However, rationality does not have to do with reliability. It has to do with the *a priori* and transparency, which are needed in order to rationalize psychological explanations. Or so it seems to me.[43] [44]

Now, importantly, the *IB joint attentional criterion* partially vindicates the Fregean idea that coreference is not enough for communicative success: the participants' thoughts must in addition be suitably related to each other. But the explanation in terms of joint attention to ib-features is consistent with a referential view of content, because it is a claim about meta-semantics. According to the useful taxonomy Dickie and Rattan provide at the beginning of their article, the *IB joint attentional criterion* could be said to fall under what they call the *Moderate Fregean view*, namely[45]

> **Moderate Fregean View:**
> Speakers can communicate using a term iff **(a)** they take it to stand for the same thing, and **(b)** they attach appropriately related [MOPs] to it.

[43] As the point is made by Boghossian, quoted in Recanati (forth) :

> We (...) ascribe thoughts to a person (...) for two related purposes ; one the one hand, to enable assessments of his rationality and, on the other, to explain his behavior. As these matters are currently conceived, a thought must be epistemically transparent if it is to play these roles. Without transparency, our conceptions of rationality and rational explanation yield absurd results. (Boghossian 1994 : 39, quoted in Recanati (forthcoming))

Recanati (*forth*) proposes an argument which converges with mine against the idea that *shared senses* have a place in the theory of content. The argument is that we do not need *shared senses* to account for trading on identity in the interpersonal domain, because trading on identity takes place intra-personally and at a time in every case.

[44] I find D&R's argument for shared senses as equivalence classes of MOPs difficult to follow, however, here is the argument I can extract from their paper:

> **(1)** On the Moderate Fregean view, the contrast between an intersubjective situation in which there is rational engagement, and one in which there is no rational engagement, is *not* marked in terms of sameness and difference in sense (by definition of the Moderate Fregean view [see definition above])
>
> **(2)** But difference in sense can explain why there isn't rational engagement in the intrasubjective domain (i.e. in Frege cases) iff rational engagement is explained in terms of sameness of sense in the intersubjective domain (Thesis of the excessive focus on the '*multiplying role* at the expense of the *consolidating role*')
>
> **(3)** Therefore **(a)** the Moderate Fregean view is unstable, and **(b)** there is intersubjective rational engagement iff there is shared sense.

As far as this reconstruction of their argument is correct, in the premiss 2, D&R are equivocating between the two different construals of 'rational engagement' I have pointed out: the first occurrence expresses COORDINATION whereas the second occurrence expresses SUCCESSFUL COORDINATION. See 5.1 of the previous chapter for an articulation of the two construals.

[45] D&R classify Heck 2002, against which they argue, as a Moderate Fregean.

As the formulation exhibits, the Moderate Fregean does not feel the need to talk of 'shared senses': it only appeals to MOPs (which can be construed as mental files, as suggested above). Accordingly, I believe that we should understand the requirement of coreferential safety as a condition on the causal transmission chain, not as a condition to coordinate on finer-grained contents. According to the criterion I have proposed here, the only content that is shared is reference. If my view is correct, the putative finer-grained shared content of the sophisticated Fregeans has no explanatory value in the theory, because it is not transparent, and we don't need it to explain *rational engagement* understood as COORDINATION.[46] Moreover, by relativizing *shared senses* to situations of use, sophisticated Fregeans threaten to trivialize the notion of sense, which becomes overly local and context-bound. This is because their proposed criterion is not applicable *across* situations of use, as a result, senses cannot be shared across situations of use on this criterion. Why go to all that trouble?

The sophisticated Fregean might reply that we need this notion of shared sense outside of communication. But where could such narrowly relativized senses be used, and for what? In the next chapters, my plan is to argue that shared senses are not needed in order to account for *communication with proper names*, *agreement and disagreement* without interaction, *attitude reports*, and I suggest the same is true with respect to the validity of *psychological laws*. The only notion of content finer-grained than reference we need has to do with private MOPs individuated by Frege's constraint.

6 Conclusion

According to the model of referential communication I have defended in the chapter, luck is eliminated on the path from *character* to *what is said* provided that (roughly) speaker and hearer are jointly aware of the intended contextual information for retrieving the referent; the hearer's interpretation is governed by this awareness and no other contextual information; and as a non-deviant result, both participants deploy coreferential files. When this relation obtains between the speech participants, we may say that their thoughts — in particular, the respective files they use to think of the referent — are *successfully coordinated* on the referent at issue, and not merely coordinated. (Thoughts (files) are *merely coordinated* whenever the thinkers presuppose they are occurently thinking about the same object, but their thoughts actually target distinct objects or else corefer by luck). I also argued that the relation of successful coordination should not be allowed to interfere with the individuation of MOPs, otherwise we loose transparency — a desirable feature when it comes to MOPs deployed in utterance production and interpre-

[46] See Heck 2002 for a similar objection, and (as already mentioned) Recanati (forthcoming).

tation.[47] [48]

I have argued that joint-attention on ib-features of utterance is what eliminates communicative luck. This condition should be understood as an externalist condition on the causal transmission chain. Joint attention is a mutual factive state, such that if two persons are not aware that they are jointly attending, then they are not in that mutual factive state. I have tried to show that this criterion is not as demanding as it seems, by providing a pre-reflective and finite characterization of joint attention, meant to be psychologically plausible, even for infants. Open (perceptual) knowledge is less intellectualist than common knowledge, and in some ways more basic than the recognition of Gricean referential intentions. The *ib joint attentional* criterion, by appealing to a mutual factive state, is a concession to the idea that one cannot define successful communication in terms of knowledge by reductively defining knowledge. However, I have tried to provide a plausible analysis of the psychological substrate of the process which terminates with the state of shared knowledge required for communicative success in face-to-face communication.

Successful coordination in communication is not all there is to samethinking, for samethinking occurs outside of communication. As already mentioned, one manifestation of samethinking occurring outside of communication is agreement and disagreement between thinkers who do not interact; another manifestation of this phenomenon is when a thinker successfully ascribes a belief to an agent which is not present in the situation of discourse, namely, in attitude reports. When thinkers do not interact, there is no ib-features to appeal to, and thinkers cannot be jointly attending. (Again, the sense of 'interact' as used here possibly applies to people who communicate by phone or video call and the like — and who may jointly attend to, for example, the sound of an explosion in the situation of either of the participants).

What does it take for two thinkers who do not interact to samethink? In the next chapter, I examine a promising proposal for generalizing the criterion of communicative success to non-interacting thinkers.

[47] See Dummett 1978, Recanati 2012, Schroeter 2007 all mentioned in Wikforss 2015. See Murez (2022) for an interesting weaker version of *functional transparency* and a program to generate empirical hypotheses about the putative transparency of MOPs construed as mental representational vehicles. In this thesis, following the sort of methodology found in Perry (e.g. 2012) and Recanati (e.g. 2012, 2016), I am mostly dealing with MOPs *qua* intuitable units of intentional content individuated by *a priori* constraints having to do with MOPs' role in rationalizing psychological explanation (such as Frege's constraint), as opposed to vehicles psycho-functionally individuated. How these two notions of MOPs — semantic appearances vs vehicles — *actually* relate is full of suspense.

[48] I don't mean that the activation of lexical MOPs is always access-transparent. The well-documented phenomenon of *Semantic priming* suggest that it is not, see e.g. Dehaene 1998.

Part II

Samethinking outside of communication

3 From *alignment* to *pragmalignment*

Abstract

This chapter has two parts.[1] The first part, using Onofri 2018 as a guide, takes the reader from a putative criterion of thought identity in communication, to an individuation criterion for shareable thoughts. The proposal is simple: drawing inspiration from Kripke's causal-historical chains, it construes the *same-thought* relation in terms of the membership in communicative and mnesic chains.

I show that this relational criterion is not admissible: interpersonal thought continuity along communicative chains, when construed as thought *identity*, is too *coarse-grained* to account for thoughts' cognitive significance and transparency. I propose a modification to Onofri's criterion that technically solves the problem to some extent, but which I find stipulative. The modified criterion seems to arbitrarily exclude agents from communicative chains, just for the sake of restoring a compatibility with Frege's constraint. But some argument is required to convince us that the stipulated clause – excluding relevant Frege cases from the communicative chains so as to individuate shareable thoughts of suitable granularity to account for Frege's constraint – is an appropriate one, *necessary for explaining communicative success*. Although not satisfactory, my proposal points to a view in the vicinity which may provide the theoretical support sought. As I explain on an example of communication with proper names, we might have *prima facie* reasons to want a more stringent criterion in order to explain communicative success.

The second part of the chapter presents and discusses this envisioned additional constraint on communicative success. According to this constraint, referential communication is successful only if agents have their singular concepts *aligned*. Very roughly, alignment requires that the lexical conventions which relate the speaker and hearer's concepts, relate them in a *one-to-one* manner.

One purpose of my discussion is to decide whether we should accept alignment as a constraint on communicative success, and more generally, samethinking. I point out that alignment-based *shared content* is not transparent, because alignment is an external relation. I examine whether misaligned coordination is always defective, and I argue that if the standards of communicative success are context-sensitive, then alignment is too stringent. My discussion culminates in an attempt to provide a pragmatic twist to the constraint of alignment — in particular, I propose to make it sensitive to the cognitive statuses of mental symbols in the minds of the speech participants. I call the resulting notion, *pragmalignment*,

[1] Large portions of the first part of this chapter also appear (in a different version) in *Inquiry: An Interdisciplinary Journal of Philosophy*, Bourdoncle 2022.

and illustrate how it works.

1 Introduction

1.1 From communication to the individuation of shareable thought

I am now looking for an account of the intersubjective relation of samethinking as it occurs outside of communication. To do this, I start by examining Onofri (2018), an attempt at individuating thoughts of different thinkers, or a single thinker at different times, in terms of the relation which holds between them in communication and memory. This relation he calls *linking*, and analyses it as the relation of *mutual knowledge of coreference*. I examine Onofri's criterion by focussing specifically on communication (and thoughts) involving proper names. This focus on communication with proper names is for heuristic and expository purposes, and is not an intrinsic limitation of Onofri's criterion. I do so for the sake of balance, as my previous two chapters were mostly concerned with demonstrative communication, and I want to make sure that the type of communication I study does not bias my theorizing on communicative success.

Communication with proper names differs from demonstrative communication. The use of a name typically spans across contexts, sometimes over a very long distance in space and time. But the use of pronouns is typically restricted to one-shot situations. Having studied successful coordination with demonstratives, we can hope that studying communication with proper names may shed a new light on the structure of successful coordination. For example, factors determining successful coordination with a name between agents typically involve not one, but several situations of use (unlike context-bound pronouns). In order to understand successful coordination with names, we need a holistic point of view on name-use, one that spans discourse situations (e.g. Kripke 1980, Chastain 1975).

The last remark constitutes a deeper methodological reason to focus on name-involving referential communication, given present purposes.[2] While the cross-contextual nature of name-use possibly complicates a theory of the factors that go into the successful interpretation of a name, it is also an opportunity for theorizing about samethinking between agents who do not interact. For example, If you know the name 'London', then it is because it was transmitted to you through an utterance, oral or written. The transmission chain leads back to an initiating use of the name to label the referent. In typical cases (disregarding descriptive names), a name N refers to an object o because o received the name N by someone in direct cognitive contact with o. Now, if you share the name 'London' with me — if we are both part of the causal network of deference and use of this name — then, even if we do not know each other, we may be said to have common knowledge of London. In fact, if you are competent with the name 'Lon-

[2]While the remark may not be specific to names, names seem to be a paradigm.

don', you can be said to have common knowledge of London with the whole 'London'-using community. (What competence with a name consists of is one of the organizing questions of the chapter.) We may think of this piece of common knowledge as a collective file distributed across the network of 'London'-users (Recanati 2016, Kamp 2015, 2019, 2022).

Because names are like *ultra-long-distance anaphora* (Cumming 2007), they may provide a causal-historical model for generalizing the relation of successful coordination to non-interacting agents. This chapter aims to assess the relevance of such a model for theorizing about non-interactive samethinking, and how this model could be developed.[3] The structure of name-using practices is thus an important theme in what follows.

A word on Onofri (2018)'s own agenda. Onofri wants a model of samethinking in terms of thought identity that respects cognitive significance. Accordingly, one constraint that governs Onofri's proposal is that the *same-thought-as* relation must be compatible with *Frege's constraint*. At this stage, I cannot rule out that an adequate theory of samethinking might *require* fine-grained thought sharing after all. What we want is to explain samethinking outside of communication, and we don't know how to do it yet. Accordingly, I will follow Onofri in his attempt to individuate shareable fine-grained thought.

Going back to the *linking* relation, it is restricted: it is reflexive and symmetric, but not transitive (if only because communication trivially fails to be transitive: that A communicates with B, and B communicates with C, do not jointly entail that A communicate with C. Otherwise, the fact of you talking to one person would entail you talking to virtually any speech participant on earth — making you a very social person[4]). So the relation of *linking* is by definition local and context-bound.

But Onofri considers the smallest relation that contains the relation of *linking* and is transitive.[5] (This is essentially the idea of representations sharing through causal-historical chains that I mentioned in connection with names above). He thinks this relation can serve in an analysis of fine-grained shareable thoughts. It is *this* aspect of his attempt that will interest me in the first part of this chapter. On the face of it, Onofri's attempt is a promising way to define samethinking outside of communication by capitalizing on a criterion of communication we already possess.

[3] I am following a venerable tradition at that, e.g. Kripke 1980, Donnellan 1974, Evans 1982, Kaplan 1990, Perry 2012, Schroeter 2012, Recanati 1993, 2012, 2016, Kamp 2015, 2020.

[4] Assuming the graph of all earthy communication is connected. In the second part of the chapter, I consider a dispositional version of the communicative relation, which relies on speakers' *dispositions* to interpret and produce signals. The dispositional version does not trivially fail to be transitive, unlike the *occurrent* version. One drawback of the dispositional version is of course that it is a semantic/competence-level relation, leaving pragmatics out of scope.

[5] By "smallest" I mean the one that have the fewest related pairs. I am simply relying on the definition of "transitive closure" in mathematics here, but the idea is very intuitive. I make it as much evident as possible using diagrams below.

Whether one can somehow make the relation of linking *transitive* is one question. Whether the generalization *thus brought about* is general enough to capture samethinking outside of communication *simpliciter* is another question. On the face of it, extending a communicative relation to communicative chains does not take us out of communication. But not all thoughts are engaged in communication. A general theory of samethinking should be able to compare even non-communicated thoughts with respect to the samethinking relation (think of perceptual or recognitional MOPs that are not deployed at any time in utterance interpretation or production), or so it seems. So one task that we face is to ensure that the generalization of the linking relation make every thoughts comparable, including the non-communicated ones. Still another task will be to ensure that the *same-thought-as* relation thus brought about is of suitable granularity to respect *cognitive significance* and (relatedly), transparency. For example, the relation of coreference, being coarse-grained, is in this respect a bad candidate for the *same-thought* relation (e.g. Frege 1892, Taschek 1998).

Let me say a word on what links this chapter to the previous ones. Onofri does not provide an analysis of the relation that underlies *mutual knowledge of coreference* in communication. He treats this relation as a "blackbox", as it were (not unlike the sophisticated Fregeans I have discussed in the first two chapters). Here, I suggest the ib-joint attentional criterion I have proposed in the last chapter may serve in an analysis of the relation that sustains mutual knowledge of coreference in communication.[6] [7] Joint awareness on the ib-feature of utterance, because it generates the common knowledge that the hearer is recovering the right interpretation, provides the speaker and hearer with the mutual knowledge that their thoughts corefer. In short, SUCCESSFUL COORDINATION underwrites the linking relation. We thus plug the gap in Onofri's theory; in addition, if Onofri's attempt at generalizing the relation of linking succeeds (or can be modified to succeed), then we may generalize our criterion.

1.2 Chapter plan

The chapter has two parts. In the first part (sections 2-6), I examine Onofri's criterion as an individuation criterion for thoughts. I argue that Onofri's proposed *same-thought* relation is too coarse-grained to account for cognitive significance and transparency: the proposal fails to satisfy Frege's constraint.[8] I then go on to propose a modification to Onofri's proposal to fix the problem. In section 6, I examine this modified criterion, and conclude that it is not satisfactory as it stands. Then I pause to consider the theoretical fork we face. The first option is to abandon

[6] I take the opportunity to revisit this model by considering a linguistic theory (the *Givenness Hierarchy*) according to which referential communication involves implicit assumptions about the 'cognitive statuses' in attention states that the intended referent has in the minds of the conversational partners, assumptions which are made manifest through a particular class of ib-feature. This is in sect. 9.1

[7] The criterion is of course silent on memory-based samethinking, a non-intersubjective relation outside the scope of this dissertation. See e.g. Burge 1993, 1998, Christensen & Kornblith 1997.

[8] This diagnosis, I believe, applies to virtually any causal-historical externalist account of representations sharing (e.g. Devitt 1981, Kaplan 1990, Sainsbury & Tye 2012).

the idea that successful communication involves the identity of thoughts (namely the principle I call Shareability). The second option is to keep Shareability in making the relation involved in communicative success more stringent.

The second part of the chapter identifies the candidate relation to keep Shareability and accommodate Frege's constraint: *alignment*. It discusses whether alignment is, in fact, a necessary condition for successful communication, and identifies central commitments and limitations of this conception of *shared content*. I shows that the alignment-based *same-content* relation fails to be transparent, and I discuss whether misaligned coordination is necessarily always defective. My discussion culminates in an attempt to provide a pragmatic twist to the constraint of alignment. I call the resulting notion, *pragmalignment*, and illustrate how it works.

2 The *indirect linking* relation

We would like thoughts to be shareable (i.e. such that they can be, and often are, type-identical across agents) and we would like them to play a role in rationalizing psychological explanation. For the latter, any two thoughts such that it is possible for a rational subject to endorse one while rejecting the other should be counted as different. This is, roughly, Frege's constraint (already introduced in the previous chapter). A consequence of **(FC)** is that the *same-thought-as* relation must be finer than referential equivalence. On the other hand, for thoughts to be shareable, the *same-thought-as* relation must yield a partition properly coarser than the partition into singletons. Thus, **(FC)** and shareability pull the granularity of thought individuation in opposite directions. As a result, it is not straightforward how to individuate thoughts so that they are both shareable and fine-grained enough to satisfy **(FC)**. In particular, one may wonder: Does shareability require an individuation criterion that is properly coarser than the one required to satisfy **(FC)**? If the answer is yes, then it would turn out that shareability is not compatible with **(FC)**.

As I announced above, Onofri proposes an individuation criterion for thoughts that purports to satisfy both constraints. His proposal is in three steps. First, Onofri defines what he calls the *linking relation*, as follows[9]:

Linking Relation (*L*) Two thoughts t_a and t_b stand in L iff the thinker of t_a and the thinker of t_b know that t_a and t_b ascribe the same property to the same object.

Then Onofri provides a first-pass individuation criterion in terms of (L), as follows:

(IC, first pass) A thought t_a is the same thought as a thought t_b iff t_a and t_b stand in L.

[9]Onofri's notion of linking is similar to Recanati (2012)'s except that the latter is restricted to the intrasubjective domain.

As it turns out, L is not a transitive relation. To show this, here is an example.[10] Consider three thinkers **A**, **B** and **C**. **B** knows that Superman = Clark and **B** knows that Superman = Kal-El. **A** only knows that Superman = Clark. **C** only knows that Superman = Kal-El. In a context in which the identity of Clark and Superman is common ground, **A** tells **B** 'Clark can fly'. Call t_A, **A**'s thought on that occasion and t_{B_1}, **B**'s thought on that occasion. Then t_A and t_{B_1} stand in L.[11] In a context in which the identity of Kal-El and Superman is common ground, **C** tells **B** 'Kal-El can fly'. Call t_{B_2}, **B**'s thought on that occasion, and t_C, **C**'s thought on that occasion. Then t_{B_2} and t_C stand in L. However, it does not follow that t_A and t_C stand in L.

Onofri does not offer a clear diagnosis of why transitivity fails. The failure of transitivity in this case may be unpacked as follows: **A** does not know that the thought she expresses by 'Clark can fly' corefers with the thought **C** expresses by 'Kal-El can fly' *or* **C** does not know that the thought she expresses by 'Kal-El can fly' corefers with the thought **A** expresses by 'Clark can fly'. There are various possible reasons why **A** or **C** may fail to know that their respective thoughts t_A and t_C corefer, which are of varying significance for Onofri's criterion. Let me mention two.

As I mentioned in the introduction to this chapter, an obvious sufficient reason for transitivity to fail in this case is that **A** and **C** may not have been present during each other's utterances to **B**. For all that the example says, **A** may be ignorant of the existence of t_C or **C** may be ignorant of the existence of t_A. In effect, this seems to be a sufficient reason for them not to believe (hence not to know) that the thoughts they express by their respective utterances corefer.

Note the consequence of this for Onofri's criterion. Onofri's criterion is supposed to be *a necessary* and sufficient condition for two thoughts to be the same. Hence, if you are in Paris in 1886 and I am in Paris in 2020 and we both think "2 + 2 = 4", we fail to think the same thought, on this criterion. This is because we don't believe (*a fortiori* do not know) anything of each other's occurrent thoughts. But it seems that type-identity between thoughts should not be thus contingent on e.g. which conversations we have. This is the very reason why we are looking for a samethinking criterion that can be applied outside of communication. As Cappelen and Hawthorne (2009, 60) remark, there is a sense of agreement and disagreement that applies to 'interaction-free pairs of individuals so long as there is some view about the world that they share '. A non-interactive notion of sameness of thought is arguably needed to define a notion of agreement and disagreement in this sense.

[10] I am reusing Onofri's original example, except that I substitute identity thoughts (and utterances of the form '*a* is *b*') for predicative thoughts (and utterances of the form '*a* is *F*'). Since L is formulated in terms of predicative thoughts, I think the example is more straightforward put this way.

[11] As far as I understand the criterion, that t_A and t_{B_1} stand in L is to be unpacked as follows: **A** and **B** both know that the thought that **A** expresses has the same referential content than the thought that **B** entertains when **B** understands **A**'s utterance. I will assume this reading-pattern of L in what follows. (For stylistic reasons, I will keep this reading implicit most of the time from now on). Moreover, in this example, the content is of course pseudo-referential. I ignore this complication, as Onofri does, for present purposes.

To remedy this on the proposed criterion, one could try to define L in dispositional terms instead (an option I will consider in the second part of the chapter, from section 7 on). Alternatively, one could reformulate the criterion so that it is not intended as a *necessary* condition for sameness of thought in general. (The criterion thus modified could no longer be used as an individuation criterion for thoughts, but at best only as a criterion of *sameness* of thought).

Another sufficient reason for transitivity to fail here is that **A** does not know that "Clark" and "Kal-El" corefer. As a result, **A** may fail to know that the thought **C** expresses by "Kal-El can fly" corefers with the thought **A** expresses by "Clark can fly". Similar considerations apply to **C**.

Be that as it may, L is not transitive, but identity is transitive, therefore L is not a candidate for the *same-thought-as* relation as it stands. Hence the third step: to remedy this, Onofri considers the transitive closure of L – which he calls "the indirect linking relation" ($L*$). He then proposes to redefine **(IC)** in terms of $L*$, as follows:

(IC) A thought t_a is the same thought as a thought t_b iff t_a and t_b stand in $L*$.

Here is a graph to illustrate how $L*$ is supposed to help with the failure of transitivity. The graph relation is L, and $L*$ is connectedness on the graph (Figure 3.1).

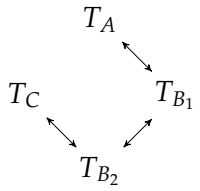


Figure 3.1 – Indirect linking relation

Node T_A and node T_C are not related on the graph, but they are connected (i.e. reachable from each other). This reflects the fact that, although **A** and **C**'s thoughts do not stand in the direct linking relation (L), they stand in the indirect linking relation ($L*$). Hence transitivity is restored by $L*$.

In the next section, I argue that **(IC)** in terms of $L*$ is too coarse to satisfy Frege's constraint.

3 The *indirect linking* relation is too coarse to satisfy Frege's Constraint

For convenience, I will re-use Kripke's well-known Pierre example.[12] Also for convenience I rebaptize Pierre "**P**". **P** (aka Kripke's Pierre) believes that London is not pretty and he also believes of London (under the French name "Londres"), that it is pretty. By hypothesis, **P** is rational.

Now consider two other protagonists: there is **Q**, who, like **P**, is bilingual in French and English. **Q** knows that "London" and "Londres" corefer. There is also **R**, a normal monolingual English speaker competent with the name "London". Here is the example:

> PIERRE: **P** tells **Q** "Londres est jolie". Call t_{P_1} and t_{Q_1}, **P** and **Q**'s thoughts on that occasion, respectively. They are linked (i.e. **P** and **Q** both know that the thoughts related to **P**'s utterance they deploy corefer). At some other time, **Q** tells **R** "London is pretty". Call t_{Q_2} and t_{R_1}, **Q** and **R**'s thoughts on that occasion, respectively. They are linked. Note that t_{Q_1} and t_{Q_2} also are linked, by hypothesis.[13] At some other time, **R** tells **P** "London is pretty". Call t_{R_2} and t_{P_2}, **R** and **P**'s thoughts on that occasion, respectively. They are linked. (Note that t_{R_1} and t_{R_2} also are linked). Of course, **P** disagrees with **R**, for he disbelieves of London, under the name "London", that it is pretty.[14]

P can rationally reject the thought he associates with the utterance "London is pretty" (i.e. t_{P_2}) while endorsing the thought he associates with the utterance "Londres est jolie" (i.e. t_{P_1}) because **P** does not know that t_{P_1} and t_{P_2} corefer. In other words, t_{P_1} and t_{P_2} are *unlinked* from **P**'s perspective. Now recall what **(FC)** says:[15]

(FC) Two thoughts are different if it is possible for a rational subject to endorse one while rejecting the other.

By **(FC)**, t_{P_1} and t_{P_2} are different. By **(IC)**, t_{P_1} and t_{P_2} are the same. I conclude that **(IC)** violates **(FC)**, because **(IC)** is too coarse.

[12] Kripke (1979).

[13] I assume that the respective memories of **P**, **Q** and **R** work properly. I also assume that the protagonists are lexically competent with the adjective "pretty", etc. throughout the episode.

[14] I say "of course", but on reflection, it is not obvious that **P** genuinely disagrees with **R** on this occasion. See my discussion of this issue in section 6 below.

[15] See chapter 1 for two weaker versions of (**FC**). The formulation used here is the classical Fregean version, which entails both of the weaker versions: if two thoughts count as different on the weaker versions, then they count as different according to the classical version.

Here is another graph to illustrate how **(IC)** and **(FC)** clash with each other in the present case. The graph relation is L, $L*$ is connectedness on the graph, and crossed out edges are used to stress disconnectedness on the graph (Figure 3.2).

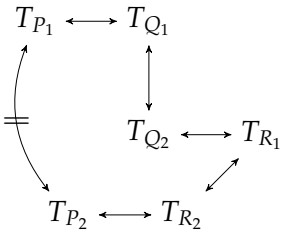


Figure 3.2 – The indirect linking relation is not immune to Frege cases

On the one hand, node T_{P_1} and node T_{P_2} are connected by a path of more than one edge on the graph. This reflects the fact that t_{P_1} and t_{P_2} stand in the indirect linking relation ($L*$). On the other hand, node T_{P_1} and node T_{P_2} are disconnected on the graph, as stressed by the crossed out edge between them. This reflects the fact that t_{P_1} and t_{P_2} are neither linked nor indirectly linked for **P**.[16]

4 Diagnosis

Intuitively, when things go well, $L*$ purports to be a route by which a thought is transmitted from one speaker to another.[17] But in general, we have no reason to assume that $L*$ -continuity

[16]Formally, node T_{P_1} and node T_{P_2} *are* connected by path, as shown on the graph. So why do I say that they are also disconnected? As François Recanati remarks commenting a previous version of this text,

> But what is disconnectedness on the graph? I thought connectedness was the fact that there is an L-path leading from one node to the other. Is this condition not satisfied here? Connectedness and linkedness were supposed to be different, but now the argument you give for unconnectedness is unlinkedness! (personal notes)

When I say that node T_{P_1} and node T_{P_2} are disconnected, what I mean is that they are disconnected in the supergraph representing all Londres- and London- related token thoughts *in Pierre's mind*. The supergraph I am talking about is the graph we obtain by adding all Pierre's token thoughts that are "London" and "Londres" related, and the *linking* relations between them, to the ones of Figure 3.2. This supergraph has two connected components, and these two components are *isolated* from each other. Using the mental files terminology, we may say that one of the two connected components corresponds to Pierre's mental file labelled "London", and the other corresponds to his mental file labelled "Londres". Whenever Pierre interprets an utterance that involves the form "London" (resp."Londres"), he is disposed to sort the incoming information into his "London" (resp. "Londres") -labelled mental file. (I provide a more systematic characterization of the dispositions to use files in utterance interpretation and production in the second part of the chapter (from section 7). These dispositions are an important aspect of, but do not exhaust, a file's cognitive role). Each component is, at any point in time, maximal i.e. such that all nodes are reachable from every other and one could not find another node anywhere in the graph such that it could be added to the subgraph and all the nodes in the subgraph would still be connected. So node T_{P_1} and node T_{P_2} are both connected and disconnected in the supergraph of which the depicted graph is a subgraph, which is the clash I am talking about.

[17]Hence the similarity with Kripke's notion of a causal-historical chain (Kripke 1980), as Onofri himself rightly notes.

will be consistent with the perspective each of the thinkers has on the thoughts they respectively deploy along a chain. For how thoughts are linked and unlinked for a thinker may be idiosyncratic, and may change over time.

By extending the linking relation (L) via communication chains, one thus naturally exposes the individuation criterion based on it to Frege cases.[18] This is another instance of the familiar observation that shareability and **(FC)** pull the granularity of thought individuation in opposite directions. Hence, the described counterexample should come as no surprise. If one wants to individuate thoughts by membership in the ordered sets corresponding to $L*$-routes, one will violate **(FC)** as soon as a route includes two thoughts of a single thinker that are unlinked from the perspective of the thinker to which they belong.

5 An attempt to fix the problem

If the diagnosis offered in the previous section is on the right lines, then a way to solve the problem suggests itself.

I repeat the diagnosis: **(IC)** violates **(FC)** whenever two coreferential thoughts are deployed by a single individual along a $L*$ -route, if these thoughts are unlinked from the perspective of that individual. Therefore, to respect **(FC)**, we want a $L*$ -route that is compatible, instead, with how thoughts are linked and unlinked from the perspective of their thinkers.

In other words, if we want to satisfy **(FC)**, we do not want to link thoughts belonging to one individual on a $L*$ -route if these thoughts are unlinked for that individual. For to do so goes against **(FC)**. Instead, to respect **(FC)**, I suggest that we should consider as linked as many thoughts as possible in the way of $L*$, while refusing to link thoughts of a single individual on a $L*$ -route if these thoughts are unlinked from the perspective of that individual.[19]

Here is a version of the indirect linking criterion that incorporates the point just made:

(Indirect Linking modulo FC) Two thoughts t_a and t_b stand in the indirect linking relation iff there is a tuple $< t_a, \ldots, t_n, t_b >$ such that:

(i) each member stands in L to its successor;

(ii) no thoughts of a single individual that are unlinked for that individual belong to $< t_a, \ldots, t_n, t_b >$.

This redefinition of the indirect linking relation looks stipulative as it stands. This is because

[18] Causal-historical chains can *fork*, as we may say using Cumming's parlance I will introduce below.

[19] Fine 2007: 113 proposes a similar idea with his notion of *coherent referential path*. However, Fine's notion is not meant to provide an individuation criterion for thoughts. I use Fine's notion later in the thesis (in chapter 5).

(IC) together with clause (ii) essentially says "count the thoughts as the same unless there is a Frege case along the $L*$ -route". But there is a less stipulative-sounding formulation in the vicinity. Instead of explicitly ruling out the Frege-cases, we may impose that all members of the ordered set belonging to a single thinker be linked for that thinker. That is to say:

(Indirect Linking modulo FC) Two thoughts t_a and t_b stand in the indirect linking relation iff there is a tuple $< t_a, \ldots, t_n, t_b >$ such that:

(i) each member stands in L to its successor;

(ii)* all thoughts of a single thinker in $< t_a, \ldots, t_n, t_b >$ are linked for that thinker.

(ii)* essentially requires that any thoughts of a single thinker that are interpersonally linked should also be linked from the perspective of their thinker. Since **(IC)** defines sameness of thought in terms of linking, **(IC)** together with clause (ii)* validates a version of the *Tranparency Constraint* for thought: for any two thoughts they deploy, a thinker should be in a position to know that the thoughts are the same, if they are the same.[20] If we define Onofri's individuation criterion in terms of this definition of the indirect linking relation, the criterion is rendered compatible with **(FC)**.[21]

6 A more stringent *linking* relation required?

My "less stipulative-sounding formulation" is still stipulative. One way to see this is the following. The relation underwriting communicational success on the present proposal (what Onofri calls the *linking* relation L, that is, mutual knowledge of coreference) by itself does *not* ensure that the condition (ii)* will hold.[22] Nor does the transitive closure of L (i.e. the *indirect linking* relation $L*$). The definition does not provide any explanation as to why transitivity should fail in the absence of clause (ii)*. Consequently, this raises the worry that shareability *so construed* is only an artificial construct which performs no genuine explanatory role: (ii)* excludes speakers that are in a relevant Frege cases from the L-path; but it does so *by fiat*.

Another way to highlight the stipulative nature of the proposed definition in terms of (ii)* is the following. There *is* a sense in which one might doubt that the disagreement between **P** and **R** is *genuine* when **R** asserts "London is pretty". This is because **P** is *also* disposed to accept the sentence "Londres est jolie", whereas 'Londres' is a perfectly admissible way to translate 'London' in Pierre's speech community (albeit, of course, not for Pierre). As a result, we have a reason to suspect that the communicative exchange between **P** and **R** might exhibit *some degree of failure*. But my modified criterion still allows **P** and **R**'s thoughts to be the same when the

[20] See e.g. Boghossian 1994 for a discussion of this notion.

[21] I use a similar solution applied to idiolectal names in the final chapter, without Shareability/shared content.

[22] This claim can be challenged, see my discussion below.

communicative chain is suitably restricted.[23]

It should immediately be noted that there is an alternative way to construe the communicative path situation depicted on the graph of Figure 3.2, and the issue whether **P** and **R** genuinely disagree. It may be that communication succeeds *at every step*, but that communicative success does not amount to *thought identity*. Here, it seems that there is a *theoretical choice* to be made.[24] I see two main options. Either one thinks that successful communication necessarily involves identity of thoughts, and then one will be happy with the verdict that the exchange presented above between **P** and **R** fails (at least to some degree). Or, one is willing to claim that successful communication does not necessarily require the identity of thoughts, and one will note that there is nothing obviously problematic in the communicative episode between **P** and **R**.

In support of the second line of thought, it is plain that there does not seem to be anything wrong in the communicative exchange between **R** and **P** *per se*. Rather, the sole reason to think this way is *holistic*, taking into account the whole communication chain of Figure 3.2. It is not as if **R** (the speaker) was using sometimes "London" and sometimes "Londres" in the *same* discourse situation, with the intention of talking about the same thing, and Pierre failed to realize that the speaker means the same. (Moreover, even in the latter scenario, it may be that the failure to recognize that the same city is referred to twice should not be blamed on Pierre). Here, in the context of the exchange with **R**, it does not seem to be a *requirement* on Pierre's part to recognize that the same city is referred to with another name, at different occasions of use, in other discourse situations.

Another way to make the point is in terms of the *individuating principles of senses* of the sophisticated Fregeans. Imagine that **P** replies to **R**: "But London is very polluted". It seems that **P** and **R** could have a productive disagreement. **R** might reply in turn, "But there are lots of nice parks in London", etc. I am suggesting that *in the situation of use*, the engagement-relevant factors do generate the immediate extension of knowledge. **P** and **R** apparently understand what the other says. And either of them would be warranted in moving to the conclusion: "Something is both very polluted and has lots of nice parks". In this hypothetical occasion of use, and considering only it, it seems that **P**'s use of "London" and **R**'s use of "London" must share a sense, as the sophisticated Fregeans would say. Indeed, no extra step of establishing that the same object is in question for both speakers is required here.[25] (Of course, the sophisticated Fregean

[23] For example, according to one admissible *same-thought* partition (by condition (ii)* above), the sequence $< t_{P_1}, t_{Q_1} >$ is one distributed thought, whereas the sequence $< t_{Q_2}, t_{R_1}, t_{R_2}, t_{P_2} >$ is another distributed thought. According to another admissible *same-thought* partition (by condition (ii)* above), the sequence $< t_{P_1}, t_{Q_1}, t_{Q_2}, t_{R_1} >$ is one distributed thought, whereas the sequence $< t_{R_2}, t_{P_2} >$ is another distributed thought. Hence the criterion is not very systematic; there is a felt arbitrariness with the proposed solution. See the section 8.1.3.

[24] See Figure 3.4 in the next section below.

[25] Dickie & Rattan 2010 might strongly disagree with my way of using their principle. In the 2010 paper, they are not dealing with factors associated with rational engagement for proper names, so it is not easy to tell what they might have in mind. But Cumming (2013a, 2013b) is a sophisticated Fregean, and he is very explicit about those factors. I present and discuss his view below.

migh reply that **P** lacks full understanding of the singular term "London", but that is question begging). As a result, claiming that communication fails between **R** and **P**, just because of how the communicative chain looks elsewhere, might be *as much stipulative* as my proposed criterion.

However, in support of the first line of thought (according to which communication fails between **P** and **R**), it should be noted that the *linking* relation — of which it is claimed that it underlies communicative success — seems to be blind to, indeed does not capture the fact that the disagreement between **P** and **R** (when **R** asserts "London is pretty") might not be *genuine*, all things considered. I suggest that this is a *prima facie* reason to look for a more stringent criterion of communicative success than the *linking* relation. For how can communication be said to be successful at every step, if it has in fact the structure of a game of Chinese Whispers? (I identify this structure as *forking* in 7.3.1).[26]

In fact, one could even doubt that the relation of *mutual knowledge of coreference* obtains in the communicative exchanges between **P**–**Q** and **P**–**R** in PIERRE, for the following reason. Because **Q** knows the identity of Londres and London, the thought **Q** deploys when **Q** understands **P**'s utterance is such that it could be equally expressed with 'Londres est jolie' and 'London is pretty' – a thought we might doubt **P** knows it corefers with the thought *he* deploys, which is associated with 'London est jolie' but *not* with 'London is pretty'. That is, **P**'s false identity belief might be a *defeater* for **P**'s knowing that the thought he expresses corefers with the thought **Q** entertains when **Q** interprets his utterance. (A similar consideration applies with respect to the thought **R** deploys).[27] Another way to make the point is to remark that the identity of London and Londres seems to be a straighforward consequence of the two pieces of knowledge of coreference attributed to **P** in the epidode. If **P** knows that the thoughts **P** and **Q** deploy ascribe the same property to the same object, and knows that the thoughts **P** and **R** deploy ascribe the same property to the same object, how can he fail to know that his two thoughts corefer? But *given* that **P** does not know the identity, how can he be said to have the relevant pieces of knowledge of coreference in the respective episodes?

Note that this latter line of thought is making assumptions one can reject. Specifically, it is

[26] *Chinese whispers* is also called *The versatile serial reproduction paradigm* in the context of cultural evolution research. It is used to identify the type of information that is more easily passed on from one agent to another. See e.g. Barrett, J. L. and M. A. Nyhof (2001), Bartlett, S. F. C. (1932/1995), Morin 2013. As Lerique (2017) explains this experimental paradigm:

> Similar to a game of Chinese Whispers, people participate in a chain along which content is transmitted; the experimenter gives a first participant initial material, typically a picture or a short piece of text, with instructions to read or memorise it; that participant is later asked to recall or reproduce the material, and the experimenter uses their output as input for the next participant, thus constructing a chain of successive memorisation (or perception) and recollection (or reproduction) of the initial material. Participants may or may not know that they are part of a chain. The setup approximates the transmission and change process that happens in everyday life. (Lerique 2017: 23)

[27] On this construal, alignment (to be introduced below) is in effect a background condition for *mutual knowledge of coreference*. For a characterization of alignment and the rationale for introducing it, see 7.3–7.4 below.

assuming that semantic facts are closed under logical consequence. But this assumption is not compulsory. In fact, this closure property is implausibly strong: if we accept it, we loose the difference between coreference *de jure* and coreference *de facto*.[28] So one might want semantic facts to be closed under a *stronger* relation than logical consequence instead. For example, it can be argued that the fact that 'London' and 'Londres' share a referent is not a *manifest* consequence of the referential facts pertaining to 'London' and 'Londres' respectively. In virtue of this fact, we can explain why 'London' and 'Londres' corefer albeit not *de jure*.[29] As Fine proposes:

> (...) It may be semantically required that ["London"] refer to the object *c* and also be semantically required that ["Londres"] refer to *c* and yet not be semantically required that the two names refer to the same object, since it might not be manifest that the object *c* in the two requirements is the same. (Fine 2007: 108)

If this idea is correct, then the fact that **P** fails to realize that 'London' and 'Londres' corefer cannot be used to deny that **P** has mutual knowledge of coreference in the relevant communicative exchanges.

Although I lean towards the second line of thought (the one that denies that communication fails between **P** and **R**, and denies that successful communication requires the identity of thoughts), I can't decide on this theoretical choice at this point. More work needs to be done to do so. The intuitions on the case are not so clear, and we need fully developed theories to make the relevant predictions. For one thing, we do not yet have an acceptable criterion for successful communication in terms of thought identity that is able to exclude, in a non stipulative and non arbitrary way, agents in relevant Frege cases from the successful communication business. We don't know yet what this notion of interpersonal sameness of thought I have used in my characterization of the main options to support intuitions about cases is. This notion appears to be sensitive to the fact that one speech participant, but not the other, is in a Frege case about the object under discussion (like **P** and **R**). How and why it is sensitive to such facts is among the things to be clarified.

In the next part of the chapter, I explore such a notion of interpersonal sameness of thought, using Cumming 2013a, 2013b as my fellow traveller. This alternative adds additional stringency to the *direct* linking relation itself (more precisely and as I shall explain, to a dispositional version of it), thus strengthening the condition required for successful communication, if one assumes that communication requires shared thoughts. Such a criterion is designed to be sensitive to the kind of putatively problematic feature in the communicative episode between **P** and **R** above, to wit., that the disagreement between **P** and **Q** might not be genuine, all things considered. It predicts that something is wrong, because of how **P**'s thoughts are related to the other agents

[28] I provide an informal presentation of the distinction between coreference *de jure* and *de facto* in the introduction of the dissertation.

[29] See Fine 2007 for an elaboration of the notion of *manifest consequence*; see also the similar notion of a chain of *explicit* coreference in Taylor 2003.

in the communicative chains. My goal is to identify central commitments of this view and its limitations. The main purpose of my discussion will be to assess whether the additional stringency on the criterion for communicative success – alignment – is one we should accept.

7 The *alignment* relation as a way to reconcile Frege's constraint and Shareability

I start by presenting the core idea of the proposal, situate it on the shared thoughts granularity problem I mentioned at the beginning of the chapter, and outline the dialectic between the competing views. Recall the tension between Frege's constraint **(FC)** and shareability we are trying to resolve. Frege's constraint pulls the granularity of thoughts towards the fine-grained. Shareability (the notion that thinkers routinely entertain type-identical thoughts or MOPs, in communication, when they agree or disagree, or ascribe beliefs) pulls the granularity of thoughts towards the coarse-grained. In the previous part, I have examined one candidate *same-thought-as* relation: $L*$ or the transitive closure of the *linking* relation, namely, the smallest transitive relation that contains the linking relation and constituted by overlapping causal-historical memory-links and communicational-links of mutual knowledge of coreference. I argued that this relation is not a good model of thought identity, because it fails to account for **(FC)**. I proposed a fix by adding an extra-constraint on this *indirect linking* relation in order to exclude the class of agents who deploy thoughts that are different from their perspective, but indirectly linked along the *L*-path. But I found the fix not independently motivated. *Alignment* is a more radical attempt to implement and provide theoretical support for the fix I proposed – the condition (ii)* above– at the level of the criterion for communicative success itself. More specifically, it requires a 1-1 mapping between thinkers' coreferential concepts as a background condition for successful communication.[30]

A consequence of alignment is (very roughly) that agents that are in a so-called *Frege case* with respect to an object *O* (such as Pierre with respect to Londres/London) cannot successfully communicate singularly about *O* with people that are not in the relevant Frege case with respect to *O* (like **R** in PIERRE).[31] In short: alignment entails that misaligned agents cannot communicate successfully.[32] For example, it predicts that communication fails between **P** and **R** in PIERRE.

It may be that this consequence of alignment is part of a good picture of communication with proper names, one that shed lights on the nature of name-using practice and successful coordination. Or it may be that this consequence is simply inadmissible, and can be argued against

[30]Cumming seems committed to Strawson's (1974) 'merging' model, on which someone who knows the relevant identity will express the same mental symbol with e.g. 'Hesperus is bright' and 'Phosphorus is bright'. For a recent defense of the 'merging' model, see Recanati 2020. Another prominent earlier defender is Millikan 1997.

[31]See the main introduction of the thesis for a definition of 'Frege case'.

[32]I use 'misaligned agents' as a shorthand for *agents whose relevant concepts are not aligned*.

by *modus tollens*, as follows: alignment entails that misaligned agents cannot successfully communicate. But misaligned agents can (sometimes) successfully communicate. Therefore, alignment should be rejected. One central aim of the remaining of this chapter is to decide the issue.

Now I will point out that if a certain conjecture is correct, then what is at stake is the validation or rejection of Shareability. Recall the question I raised at the begining of the chapter: Does shareability require a *same-thought* relation that is properly coarser than the one required to satisfy **(FC)**? The conjecture I want to put forward is that alignment is the only *same-thought* relation which satisfies both **(FC)** and Shareability. It will take me a section to etablish the conjecture, but I introduce it now to clarify the dialectic.

The conjecture says that the only partition of the set of MOPs or thought tokens that accommodates both **(FC)** and **(SHAR)** is determined by a *Cummingian* relation on the set of MOPs or thought tokens. A *same-MOP-as* relation is *Cummingian* just in case any equivalence class in the partition induced by that relation intersects on *at most one element* with any equivalence class from the partition of mental symbols into agent lexicons. In other words, the conjecture is that all the *same-MOP-as* or *same-thought-as* relations that are immune to Frege cases are *Cummingian*. The conjecture is depicted in the following diagram (Figure 3.3), which exploits the fact that the *finer-than* relation on the set of *partitions* of the domain of MOPs or thought tokens is a partial order, so that we may represent and examine all the candidate *same-MOP* or *same-thought* relations at once by employing a partition lattice (its least upper bound is the partition induced by referential equivalence and its greatest lower bound is the maximally fine-grained "each and every one" partition in which no MOP or thought is ever shared):

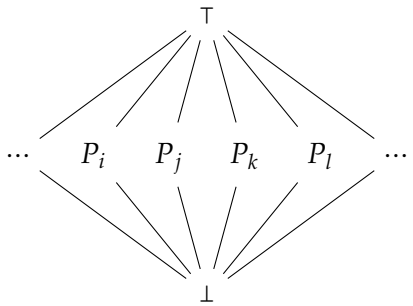


Figure 3.3 – *Cummingian* partitions on the set of thought tokens

If the conjecture is correct, then any relational individuation criterion must be Cummingian in order to define shared thoughts of suitable granularity to account for cognitive significance. Even if the conjecture is true, the validation of Shareability is not thereby settled. In addition, we need to know whether alignment is *in fact* a necessary condition on successful referential communication. I recap the dialectic in Figure 3.4 below.

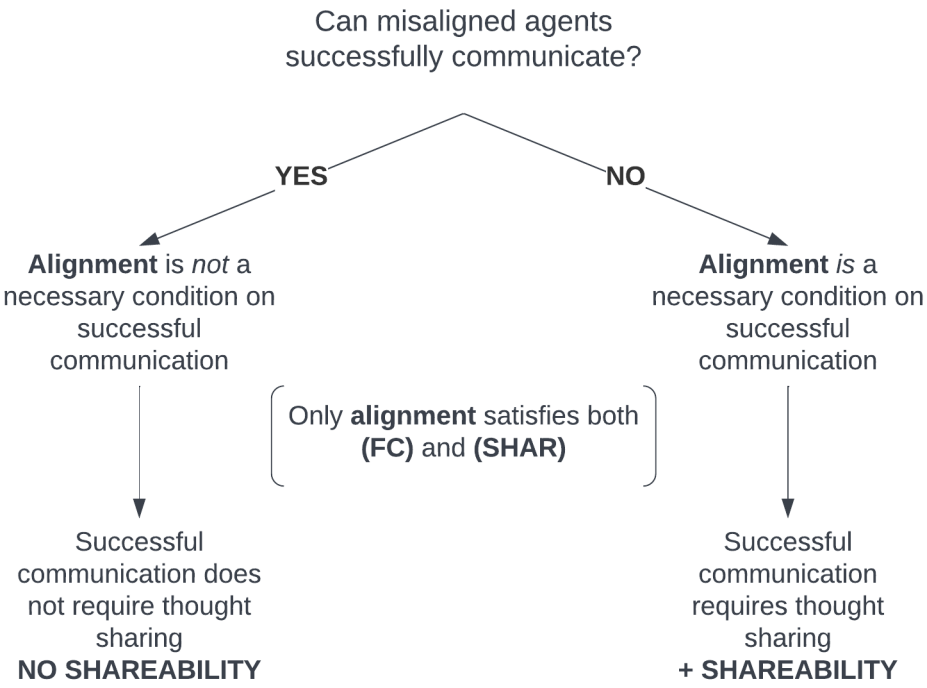


Figure 3.4 – Deciding whether thoughts are shareable

Let me now present the alignment relation in more detail. For expository purposes, the structure of Cumming's theory may be seen as consisting of two parts. There is a ground relation — call it R — that connects the thoughts or MOPs of different agents. And there is the relation of *content sharing* proper — call it $R*$ — that is a refinement of R: it is obtained from R, by adding an extra-constraint on it.[33] Intuitively and very roughly, R gives us communication while $R*$ gives us successful, perfect communication. Communication can be said to track content only in this latter sense (i.e. when it is underwritten by $R*$), as we shall see. I now turn to the first level of Cumming's theory.

7.1 The ground relation ($\rightharpoonup$)

To introduce the ingredients of Cumming's proposed answer to the problem of shared content, we need to switch to a *strategic* point of view on verbal communication.[34] Let us assume, then, that when we communicate, we use utterable signals (e.g. words) *strategically* to signal the mental symbols that we choose to express. '*Strategically*' means that one's choices as a speaker are conditioned by the choices that a hearer will make in interpreting what one says. (In this respect, communicative strategies are not unlike Gricean communicative intentions and their recognition). Hence communication may be seen as a coordination problem. A speaker's strategy is an algorithm to convert private symbols $s_1,\ldots,s_m$ into utterable signals $\sigma_1,\ldots,\sigma_n$.

[33] These are my labels, references to Cumming will follow.

[34] In so doing, Cumming is following an important tradition tracing back to Lewis (1969), who proposed that communication involved particular kinds of *coordination problems* amenable to straighforward mathematical, game-theoretic analyses.

Formally, it is a mapping from mental symbols into signals. A speaker acts according to a strategy *Fs* if for each mental symbol s_i, whenever they choose to express it, they act according to a mapping $Fs(s_i)$.[35] A hearer's strategy is an algorithm to convert utterable signals $\sigma_1,\ldots,\sigma_n$ back into private symbols $h_1,\ldots,h_m$. Formally, it is a mapping from signals to mental symbols. The hearer acts according to a strategy *Fh* if, for each signal σ_k in the domain of *Fh*, she executes $Fh(\sigma_k)$ if she observes that the speaker produces σ_k. Here is a sample of speaker's strategies:

Fs1:
If you want to express s_1, produce σ_1;
If you want to express s_2, produce σ_2.

Fs2:
If you want to express s_1, produce σ_2;
If you want to express s_2, produce σ_1.

Fs3:
If you want to express s_1, produce σ_1;
If you want to express s_2, produce σ_1.

Fs4:
If you want to express s_1, produce σ_2;
If you want to express s_2, produce σ_2.

Here is a sample of hearer's strategies:

Fh1:
If you observe that σ_1 is uttered, token h_1;
If you observe that σ_2 is uttered, token h_2.

Fh2:
If you observe that σ_1 is uttered, token h_2;
If you observe that σ_2 is uttered, token h_1.

Fh3:
If you observe that σ_1 is uttered, token h_1;
If you observe that σ_2 is uttered, token h_1.

Fh4:
If you observe that σ_1 is uttered, token h_2;
If you observe that σ_2 is uttered, token h_2.

[35] I don't mean that *Fs* is always a function; see below.

Table 3.1 – **Successful vs. Unsuccessful strategy profiles**
Two strategy profiles are *equilibria*.

		Hearer	
		Fh1	Fh2
Speaker	Fs1	1,1	0,0
	Fs2	0,0	1,1

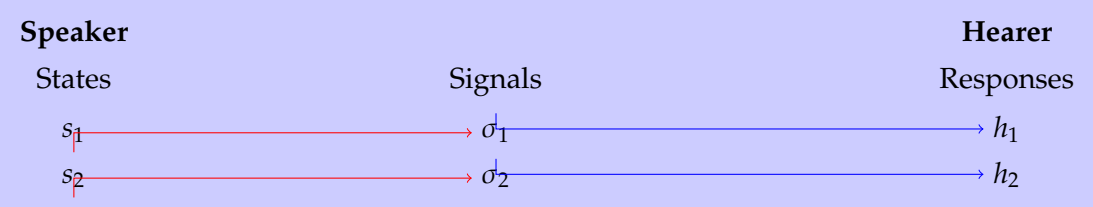


Figure 3.5 – The strategy profile < $Fs1, Fh1$ > is an equilibrium

A *strategy profile* is a combination of a speaker's strategy and a hearer's strategy. It may be thus construed as a *joint policy* governing communication between a particular pair of agents. When a state of *equilibrium* is reached via strategy profile, neither agent can benefit from being the only one to change their strategy (cf. Figure 3.5 and Table 3.1). Cumming calls a speaker's strategy that can suitably interact with a hearer's strategy to reach an equilibrium, *competent expression*. And he calls a hearer's strategy that can likewise suitably interact with a speaker's strategy to reach an equilibrium, *competent construal*. Both types of strategies are communicative policies taken in an *internalist* sense, an extension of their I-LANGUAGE — i.e. the cognitive system comprising an agent's cognizance of linguistic rules:[36]

> . . . *competent expression* and *competent construal*, as the terms are used in this paper, are efforts that match the internal policies belonging to the cognitive system of the relevant agent, and are not defined in terms of the preservation of content. (2013b p. 383)

> (...) an important component of linguistic strategies of communication is the agent's *linguistic competence*, or grammar in the Chomskian sense. An agent's grammar connects utterable signals (e.g. words), via its lexicon, to mental representations, but is itself opaque to introspective reflection (2013a: 8)

Hence, both types of strategy are subpersonally implemented algorithms. One might be tempted to read Cumming as subscribing to a form of what Sperber & Wilson call *the code model* of human verbal communication here. Sperber & Wilson define the code model as follows:

> According to the code model, communication is achieved by encoding and decoding messages [=representations internal to the speaker]. (S&P 1996: 24)[37]

[36] See Chomsky (1986) quoted in Cumming 2013a
[37] As Sperber & Wilson further explain:

Here is a diagram from Sperber & Wilson depicting the code model applied to human linguistic communication (Figure 3.6):

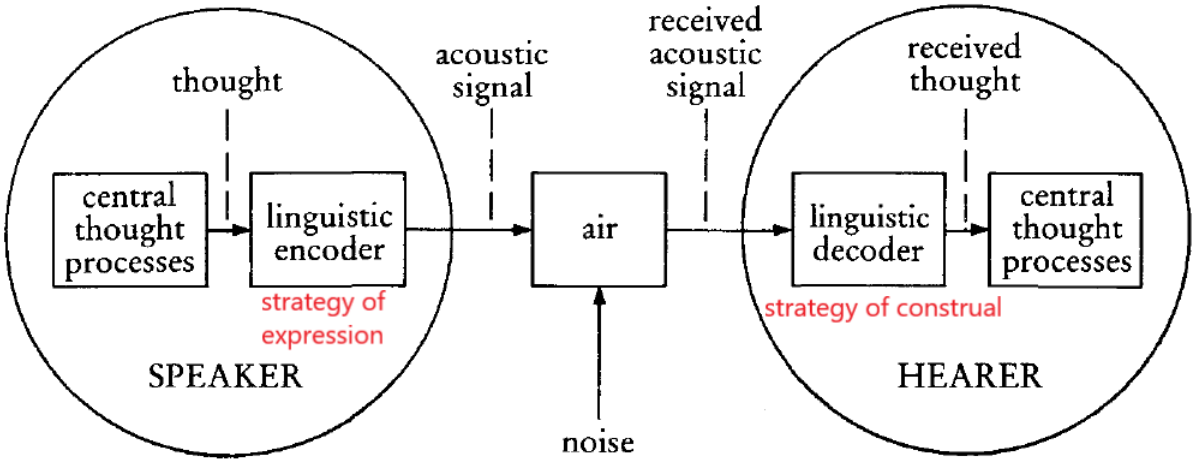


Figure 3.6 – Code model of human verbal communication (from Sperber& Wilson 1996, slightly modified in red)

One may wonder whether and how pragmatics could impact strategies as defined. For one thing, there is no model of grammar which would pair a signal with a *speaker* meaning, that is, a meaning as *the speaker* intends to use it. At a minimum, strategies should be able to recruit representations of *intentions* to count as pragmatics. For, as I have argued in the first two chapters, construals typically appeals to complex background information about the context.

The *code model* is not an adequate reading of what Cumming wants with the strategies. An agent's grammar is only *one* component of the communicative strategies. Agents have rich and flexible communicative dispositions, and can use different linguistic devices to express their symbols, depending on the context.[38] So we better understand the notion of a mutual policy of expression and interpretation at a more abstract level, as being agnostic on the nature (possibly pragmatic) of the coordinating strategy. Understood in this way, strategies could involve e.g. *ad hoc* signals, and would not necessarily imply a code model where the message-signal pairs are *pre-established*.

Going back to Cumming's ground relation, it is essentially the relation that obtains between two mental symbols of different agents just in case these mental symbols are connected via

> A code (...) is a system which pairs messages with signals, enabling two information-processing devices (organisms or machines) to communicate. A message is a representation internal to the communicating devices. A signal is a modification of the external environment which can be produced by one device and recognised by the other. A simple code, such as the Morse code, may consist of a straightforward list of message-signal pairs. A more complex code, such as English, may consist of a system of symbols and rules generating such pairs. (S&W 1996: 4)

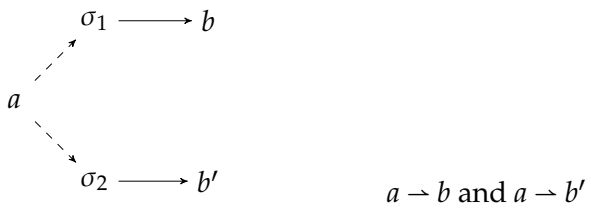
[38] See my model of communication with cognitive architecture of the previous chapter.

strategy profile and this strategy profile is an equilibrium. Said in the terminology of Cumming (2013b); given the mental symbols a and b (belonging to agents A and B):

> $a \rightharpoonup b$ iff the pair of strategies consisting of {the strategy of A according to which A expresses a} and {the strategy of B according to which B interprets by b the signal sent by A to express a}, is an equilibrium. (2013b: 383-384, modified by me).

Another feature of these strategic interactions is that they need not be a function (2013b: 384). That is, Speaker's strategy may well associates one of Speaker's mental symbols with *two* different signals, while Hearer's strategy associates two distinct symbols with these distinct signals[39]. For instance, agent A might express his symbol a as either 'attorney' or 'lawyer', while agent B interprets 'attorney' as b and 'lawyer' as b'. Here is a diagram to illustrate (Figure 3.7):

Figure 3.7 – The relation ($\rightharpoonup$) need not be a function



Crucially, agents need not actually communicate for their mental symbols to be strategically connected (i.e. to be in the relation that I called "the ground relation"). A communicative policy (i.e. a strategy of expression or construal) is a competence. As a result, agents need not interact to combine their strategies of expression and construal. So Cumming's ground relation is a dispositional relation (unlinke the *linking* relation we have examined above). This allows to relate speakers that otherwise do not interact.

That does not imply that the relation of strategic connection ($\rightharpoonup$) is transitive over its domain: that is, given three mental symbols a, b, and c belonging to agents A, B, and C, that $a \rightharpoonup b$ and $b \rightharpoonup c$ does not entail $a \rightharpoonup c$. For instance, A could be a monolingual English speaker, C a monolingual French speaker, where B is bilingual in English and French. As a result, A's and C's strategies of competent expression and competent construal cannot interact in the right way, because the signals that enter into A's strategy of expression are not in the domain of the strategy of

[39]It is unclear to me why Cumming does not want to conceive of this coordination game as a model in mixed strategies (that is, with probabilistic ones). Assuming there are m mental symbols and n signals (I assume that $n \geq m$), a convenient way to depict the probabilities that the speaker will send signal σ_j when choosing to express s_i is in terms of the $n \times m$ matrix (let's call it Z), whose entries z_{ij}, denote the probability that the speaker will send the signal σ_j whenever they chooses to express s_i.

interpretation of *C* (and likewise, the signals that enter into *C*'s strategy of expression are not in the domain of the strategy of interpretation of *A*). Still, we *want* to say that there is a strategic path between *a* and *c*, because agent *B* is a competent mediator between *A*'s and *C*'s strategies of competent expression and competent construal. I now turn to Cumming's way to address this issue.

7.2 Communicative paths

Consider the following graph (Figure 3.8):

Figure 3.8 – A communicative path

$$a \longrightarrow c \longrightarrow d \longrightarrow b$$

The mental symbol *b* is not in the relation ($\rightharpoonup$) to the mental symbol *a* (perhaps the signals that enter into *A*'s strategy of expression are not in the domain of *B*'s strategy of interpretation). This is reflected in the fact that node *a* and node *b* are not related on the graph. However, node *b* is *reachable* from node *a*: there is a path between them, via *c* and *d*. The reachability relation is what Cumming calls the relation of COMMUNICATIVE PATH.[40] It is the transitive closure of the relation ($\rightharpoonup$) – since the latter relation is not transitive, its transitive closure is a different relation. Said differently,

> a COMMUNICATIVE PATH from *a* to *b* is a sequence of length *n* (n ≥ 2), $< x_1, \ldots, x_n >$ such that $a = x_1$ and $b = x_n$ and, for each *i* $(0 < i < n)$, $x_i \rightharpoonup x_{i+1}$. (2013b p. 384)

To take my previous example again, with agents *A* and *C* monolingual in English and French respectively, and *B* bilingual in French and English, the relation of COMMUNICATIVE PATH allows us to say that a message can originate at mental symbol *a* and terminates at mental symbol *c* without divergence of policy. That will be so if *A* first speaks to *B* (where *a* gets translated as *b*), and *B* passes the message to *C* (where *b* gets translated as *c*). Ultimately, *a* gets translated as *c* in this indirect way.

Neither the relation of communicative path nor the relation of competent expression + competent construal — what I called the 'ground relation', in Cummingian symbolism: ($\rightharpoonup$) — are sufficient for sharing content finer-grained than reference.[41] Accordingly, on Cumming's view, neither one nor the other relation is sufficient to underwrite successful communication. (Cumming is thus a sophisticated Fregean). I illustrate why Cumming thinks so with examples in the next subsection.

[40] A very similar relation, called 'strategic path', is presented in the other paper (2013a) of the diptych I am engaging.

[41] In the following, when there is no ambiguity, I often write 'content' for 'finer-grained than reference content'.

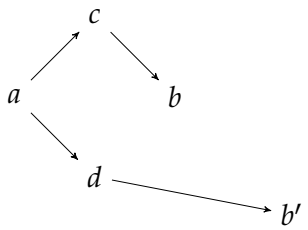
7.3 The need for an extra constraint

There are essentially three types of configuration in which the relation of COMMUNICATIVE PATH fails to transmit (fine-grained) content. If one thinks that successful communication *must* preserve fine-grained content, then one will think that cases involving a configuration in which the relation of communicative path fails to track content cannot be cases of successful communication. This is Cumming's view on the matter (I will later consider an alternative view). Let me present these types of configuration in turn.

7.3.1 Forking

Consider the following situation (Figure 3.9):

Figure 3.9 – Forking



Agent *A* expresses her symbol *a* with 'Hesperus' or 'Phosphorus'. Agent *C* interprets 'Phosphorus' with *c* (and expresses *c* with 'Phosphorus'). Agent *B* interprets 'Phosphorus' with *b* but interprets 'Hesperus' with *b′* (where *b* and *b′* are two distinct symbols). Agent *D* interprets 'Hesperus' with *d* (and expresses *d* with 'Hesperus'). This is the configuration of communicative policies depicted on the graph. This type of configuration is called 'forking' because the communicative path terminates at two distinct symbols belonging to the same agent, namely *b* and *b′* (it 'forks').[42]

Why are situations of forking *problematic* from the point of view of content sharing? This is because a chain that forks terminates at two *distinct* symbols belonging to the same agent. In the Representational Theory of the Mind that Cumming assumes, it is axiomatic that two distinct symbols belonging to the same agent have different contents, because content is individuated by syntax at the intrasubjective level.[43] So, at most one of *b* or *b′* can have the same content as *a*. However, we have no non-arbitrary way of choosing which symbol among *b* and *b′* should be put in the *same-content* relation to *a*. Therefore, this is a case where the relation of COMMUNICATIVE PATH fails to track content.

I am following Cumming in describing the forking configuration at the subpersonal level, that is, in terms of the arrangement of communicative strategies. But we may also construe forking

[42] Note that my forking example is not a minimal example: forking is possible with only two agents.
[43] See Fodor 1975 for what might be the first formulation of this claim.

configurations at the personal level. Communicative paths that fork are communicative paths that terminate at the mental representations of an agent in a Frege case with respect to the referent of her symbols. In the example I give, and more specifically, agent *B* has one mental representation affected by (and deployed when) encountering tokens of the name "Hesperus", but not affected by (and not deployed when) encoutering tokens of the name "Phosphorus". And *B* has another, distinct mental representation for which the converse is true. Hence, what we find is that the forking configuration is an instance of the clash between Frege's Constraint and Shareability (**(FC)** vs. **(SHAR)**).

Cumming seems happy to allow such a move from the personal to the sub-personal level and vice versa.[44] For instance, he formulates the *Transparency Constraint* – what he calls 'a constraint of *perspicuity* on content', mentioning Frege (1892) and Evans (1982) – in subpersonal terms:

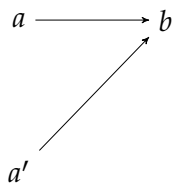
> Perspicuous contents are analogous to readable symbols. The mind acts on them in a manner of a symbolic processor, combining them into complex thoughts and copying them from one cache to another (e.g. converting a long-term goal into a current intention). *The Fregean theory of perspicuous contents is a natural match for the computational theory of mind*, which is itself recommended for its potential to give a naturalistic account of the mind's information-processing and problem-solving capacities. (2013b p. 380, my italics)

I will follow Cumming in moving seamlessly from the subpersonal-computational level to the personal level. In this thesis, I am construing mental symbols as intuitable units of intentional content individuated by *a priori* constraints having to do with their role in rationalizing psychological explanation (such as Frege's constraint). However, if the goal is to study mental symbols *qua* psycho-functionally individuated mental representations, then I don't think we should take the *natural match* between the computational level and the personal level for granted.[45]

7.3.2 Pooling

Another type of configuration of communicative policies in which the relation of COMMUNICATIVE PATH arguably fails to track content is *pooling*. Consider the following situation (Figure 3.10):

Figure 3.10 – Pooling (1)



[44] This is as expected if pragmatics impacts strategies, for pragmatics is often described at the personal level (e.g. Recanati 2002).

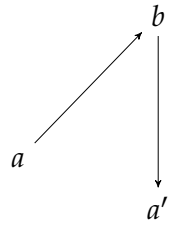
[45] I am following Murez 2022 here.

An example of the arrangement of communicative strategies depicted on the graph is the following. Agent A expresses her mental symbol a with 'Hesperus', and expresses her mental symbol a' with 'Phosphorus' (where a and a' are distinct mental symbols). On the other hand, agent B interprets *both* 'Hesperus' and 'Phosphorus' as b. As before, since a and a' are distinct, they have distinct contents. So we cannot say that the content of b matches that of a and a'. But there is no basis to choose which of these two symbols has the same content as b. This is a configuration of 'pooling' because the communicative path starts at two distinct symbols belonging to one agent, but terminates into another agent's single symbol (it 'pools' the two symbols at which the path starts into another agent's single symbol)[46].

Again, we can describe what's happening here at the personal level, in virtue of the 'natural match' that is axiomatic in Cumming's framework between the personal-level notion of *Transparency* and the computational characterization of the mind. Agent A is in a Frege case with respect to the unique referent of her mental symbols a and a': she has two distinct symbols where agent B has only one (all symbols being coreferential). In particular, A has one concept (a) that is expressed with tokens of the name 'Hesperus' (but not with tokens of the name 'Phosphorus'); and A has another, distinct concept (a') for which the converse is true. By **(FC)**, a and a' cannot be in the *same-content* relation. However, by **(SHAR)**, we should say that a and a' are in the *same-content* relation with b. But this cannot be. So the configuration fails to track content.

A variant of the pooling configuration is as follows (Figure 3.11):

Figure 3.11 – Pooling (2)



Here is one example of this type of situation: agent A expresses a with 'Hesperus', but interprets 'Phosphorus' as a'. On the other hand, agent B interprets 'Hesperus' with b and expresses b with 'Phosphorus'. The communicative path *pools* A's two mental symbols into a unique mental symbol from B. Said differently: agent A has one concept that is expressed with the word 'Hesperus' but *not* with the word 'Phosphorus'. And A has another, distinct concept a' that is activated when encountering tokens of the word 'Phosphorus', but not activated when encountering tokens of the word "Hesperus". On the other hand, agent B has a single concept that is expressed with 'Phosphorus', and that is triggered by the public signals "Hesperus".[47]

[46] For a first formulation of this sort of case, see Crimmins 1992.

[47] This case is for illustrative purposes only; we might not find it in nature. In effect, agents' grammars are

7.3.3 Missing connection

A third way in which communicative paths can fail to track content is when there are holes in the communicative path. Consider the following situation (Figure 3.12):

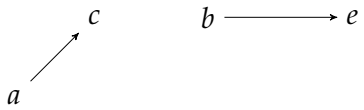


Figure 3.12 – Missing connection

An example of such a configuration is the following. *A* expresses *a* as 'Hesperus' and *C* interprets 'Hesperus' as *c*. Agent *B* expresses *b* as 'Napoleon' and agent *E* interprets 'Napoleon' as *e*. There is no strategy profile (i.e. no combination of existing strategies of expression and of interpretation) relating 'Hesperus' and 'Napoleon' in the population. Hence the missing connection. Here, content is not transmitted from *a* to *e*.

7.4 Alignment (⇌)

We have surveyed three types of configuration in which communicative paths fail to track content (the third case being a trivial case at that, I will just ignore it for now). Pooling and Forking are types of configuration in which a communicative path connects *more than one* symbols belonging to the conceptual lexicon of a *single* agent with other symbols of different agents. What all this suggests is the following:

> The relation of COMMUNICATIVE PATH tracks content only if the path relates at most *one* symbol per agent.[48]

This is, you will recall, the conjecture I formulated in the previous section: only relations that are *Cummingian* can accommodate both **(FC)** and **(SHAR)**. They exclude misaligned Frege cases from the conceptual sharing business. We now have the extra constraint which, when added to our ground relation of COMMUNICATIVE PATH, gives us the intersubjective *same-content* relation that Cumming puts forward in our diptych[49]

generally unbiased with respect to perception or production. See Hurford 1989 for an explanation from biological evolution of the bidirectionality of the mapping between signals and concepts.

[48]This principle is reminiscent of Fine's notion of a *coherent referential path*, which I mentioned above. The major difference with Cumming is that Fine allows successful communication and understanding between misaligned agents. Fine's notion of coordination does not support shared content. Moreover, as he defines it, Fine's notion of interpersonal coordination is consigned to idiolectal names underlying a 'common currency' name in a speech community. Fine says that two idiolectal names are coordinated if they are connected by a coherent referential path, where a referential path is roughly a causal-historical transmission chain along which a 'common currency' name is transmitted from one speaker to another. A path is *coherent* when the sequence does not admit more than one idiolectal name per speaker. Cumming sees the debate with Fine as substantive, and he is assuming something like the difference I indicate (see e.g. 2013b: 386-387; 396; *in press* 10-11). As indicated earlier, I make use of Fine's concept in chapter 5.

[49]He calls it 'alignment' in 2013b and '*coordination de facto*' in 2013a. The two terms denote roughly the same relation (both relations are, in my terms, *Cummingian*). I prefer to use the relation of alignment in this chapter.

Alignment ($\rightleftharpoons$) — the relation that we get with the extra constraint The base relation is the relation of strategic connection, the transitive closure of which is the relation of COMMUNICATIVE PATH, and the target relation is the relation of COMMUNICATIVE PATH *that tracks content*. We have just seen that a communicative path fails to track content when it *forks* a unique mental symbol into two distinct mental symbols belonging to another agent *or* when it *pools* two mental symbols belonging to a single agent into a unique mental symbol belonging to another agent. To get a relational criterion out of the relation of COMMUNICATIVE PATH that is a necessary and sufficient condition for sharing content, we therefore need a relation that is immune to both forking and pooling situations.

This is what Cumming obtains by adding the extra constraint of *1-1 mapping*. The target relation, Cummings calls alignment. Accordingly, mental symbols a and b (belonging to agents A and B) are **aligned** ($a \rightleftharpoons b$) iff:

a. There are COMMUNICATIVE PATHS from a to b and from b to a;

b. a is the only symbol in A's lexicon connected to b by a path in either direction;

c. b is the only symbol in B's lexicon connected to a by a path in either direction.

Condition (a) ensures that a is connected to b via strategies of expression and construal in the agent A, and likewise for agent B. We want the mapping between public signals and concepts to be *bidirectional*. Agents' grammars should be unbiased with respect to perception and production. Condition (b) together with condition (c) is what allows Cumming to ground a notion of *fine-grained content* on communicative paths. This is because (b) + (c) ensures that any equivalence class in the partition defined by ($\rightleftharpoons$) intersects with *at most one element* with any equivalence class in the partition into agent lexicons (see Figure 3.13). In other words, alignment is a one-to-one mapping. In my terms, condition (b) together with condition (c) makes the relation of alignment *Cummingian*.

Given two agents A and B, the most simple case of alignment (see Figure 3.13) is such that

$$a \rightharpoonup b \wedge b \rightharpoonup a$$

and there is no $x \neq a, b$ belonging to the lexicon of A or B such that

$$a \rightharpoonup x \vee x \rightharpoonup a \vee b \rightharpoonup x \vee x \rightharpoonup b$$

Finally, two mental symbols have the same content iff they are aligned:

Symbols x, y (belonging to different agents) have the same content iff $x \rightleftharpoons y$.

Reified naturalized contents The content attached to a mental symbol a is thus the equivalence class of all mental symbols that are aligned with a.[50] In effect, alignment enables to reify contents: we can quantify on them.[51] Let me explain. Given the set of all mental symbols at a time, the equivalence classes determined by the relation of alignment ($\rightleftharpoons$) form the set of *all contents* at that time.[52] We can call this partition (i.e. the set of equivalence classes by the relation of alignment), the *quotient space* of the set of mental symbols at a time by ($\rightleftharpoons$). Calling S_t the set of all mental symbols at time t, the quotient space is written $S_t / \rightleftharpoons$. Given a mental symbol a of an agent A, its content is the equivalence class of all mental symbols aligned with a, namely:[53]

$$\{x \in S_t : x \rightleftharpoons a \}$$

Cumming's alignment-based theory of shared content is thus a recipe to *naturalize* suitably fine-grained and shareable Fregean senses out of ingredients that Russellians accept, and that we better understand. As we saw, the recipe consists in constructing *contents* out of mental representations connected by 1-1 lewisian conventions.[54] Whatever my verdict on the alignment constraint, I endorse Cumming's naturalization constraint: representations sharing must constitute a potential subject of study for the natural sciences.

---

[50] Things are complicated by the fact that alignment is relative to sets of agents. See my discussion below.

[51] An historical note: Quine 1956 famously argued against fine-grained contents (he calls them *intensions*). He thought 'quantifying into names of intensions' was a 'dubious business'. He said the following:

> Intensions are *creatures of darkness*, and I shall rejoice with the reader when they are exorcised. (Quine 1956: 180; italics mine)

Cumming's defense of alignment-based content is an explicit reply to Quine 1956, following Kaplan 1968's paper *Quantifying In* (which may be seen as an ancestor of the mental file theory). Hence the title of Cumming's 2013 paper (a mixed quote), *Creatures of darkness*.

[52] Given that thought is productive, the set of all contents à la Cumming is perhaps better construed as the base allowing us to move from the content of a few primitive concepts at a time to the content of all possible thoughts at that time.

[53] Again, putting aside the relativization of the alignment relation to sets of agents.

[54] The 'recipe' methodology has been famously described by Dretske 1994:

> If you want to know what intelligence is, or what it takes to have a thought, you need a recipe for creating intelligence or assembling a thought (...) out of parts you already understand (1994: 468-469).

Both Dretske and Cumming are naturalizers of content and intentionality.

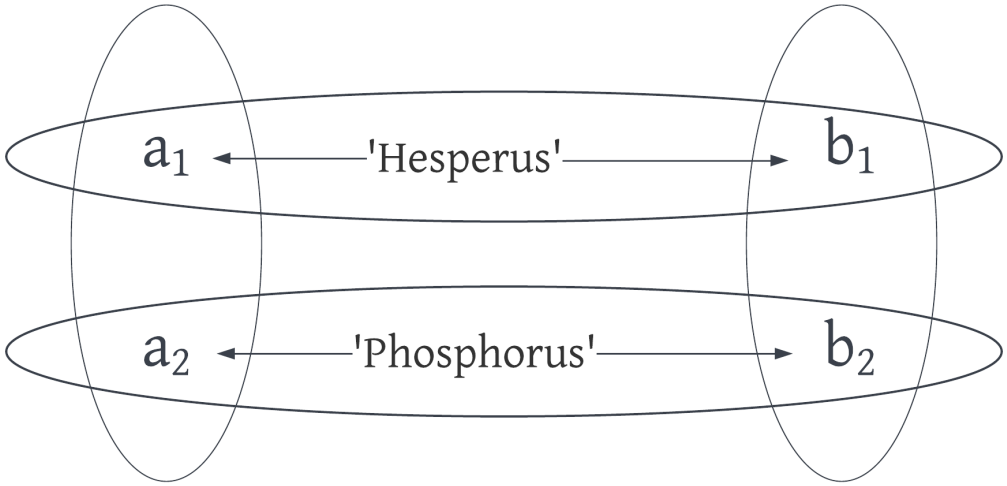


Figure 3.13 – The partition of the set of mental symbols by the relation of alignment is orthogonal to the partition of that set into agent lexicons: any equivalence class in the *same-content* partition intersects on at most one element with any equivalence class from the *same-agent* partition.

As the example of Figure 3.13 shows, agents in Frege cases can successfully communicate with each other to the extent that their mental symbols are aligned. For instance, the Babylonians were able to communicate about Hesperus and Phosphorus, respectively. (A more complex case of alignment would involve a communicative path involving direct connections to other symbols belonging to other agents.)

This is the end of my presentation of alignment and the intersubjective *same-content* relation based on it. I will now examine whether we should accept alignment as a necessary condition for successful communication, and samethinking more generally.

8 The status of misaligned coordination

My critical examination of the alignment constraint elicites intuitions about cases, and accumulate theoretical support from various angles substantiating anti-alignment intuitions. Ultimately, the goal of this section is not to reject the alignment relation altogether, but instead to give it a pragmatist twist.

My discussion has three parts. In the first, I argue that *contents* based on the alignment relation are not transparent. Relatedly, I criticize the relativization of alignment to sets of agents on the ground that it disrupts the individuation of MOPs. In the second part, I argue that alignment does not line up with the intuitive demarcation of communicative success. I provide theoretical support in favor of this judgment, and discuss epistemological counter-arguments in favor of the constraint. In the third part, I point out that alignment involves disambiguated

signals. This is a huge idealization in the model: in real life, ambiguity is prevalent, indeed a normal aspect of our linguistic practice. Relatedly, depending on how linguistic strategies are individuated, I point out that Loar cases involving demonstratives might not even be in the scope of Cumming's alignment-theoretic picture of shared content.

8.1 Alignment-based contents are not transparent

Here I argue that, contrary to what Cumming seems to suggest, alignment-based fine-grained shared contents are not transparent. Furthermore, relativizing alignment to networks of agents seems to disrupt the individuation of MOPs. Let us take these two points in turn.

8.1.1 The alignment relation is not transparent

Here I argue for the following: the alignment relation is external, so it is not transparent, so the *same-content* relation based on it is not transparent either.

I find Cumming's claims on the alleged transparency of his *same-content* relation puzzling. Sameness of content as he defines it is a matter of whether agents' mental symbols are aligned. Now, alignment is a property of the mutual lexicon of a given pair of agents. But, interlocutors do not have direct access to the lexicon structure of one another. So they can be mistaken about whether they share content. For example, it seems that I could falsely believe that e.g. Kripke's Peter has only one symbol for Paderewski, hence falsely believe we share content when we do not.[55] Conversely, it seems that I could falsely believe my interlocutor has a false identity belief about the object under discussion, when in fact she has not. We would share content in virtue of our lexicons' being aligned, but I would fail to know it. Let me illustrate these two points in turn.

It seems to me that agents could have their mental symbols aligned, but fail to know it. Consider the following example:[56]

> FALSEPHORUS:
>
> *Anna falsely believes Bob does not know Hesperus is Phosphorus. Anna and Bob have their symbols aligned.*
>
> **Bob:** Phosphorus is my favorite planet.
>
> **Anna:** Yes, it's so shiny!

In FALSEPHORUS, by hypothesis, our agents have their mental symbols aligned. That is—calling a and b, Anna and Bob's mental symbol for Venus respectively—there is one and only one construal of a in Bob's lexicon (namely b), and one and only one construal of b in Anna's

[55] cf. Fine 2007 p. 109 for a similar observation.

[56] I am assuming something like the 'merging' model of Strawson, a 'parti-pris' I share with Cumming. See Recanati 2020 for a recent defense of the 'merging' model.

lexicon (namely *a*). However, Anna has a mistaken representation of the context. Due to her false belief, she does not know that her interpretation matches in content —in the sense of 'content' at issue— with the belief Bob expressed.[57]

Furthermore, it seems to me that agents could have their mental symbols misaligned, but fail to detect it, just as well. Consider the following case:

> ANNASPHORUS:
>
> *Bob believes Hesperus and Phosphorus are two different planets. Anna believes Hesperus is the same planet as Phosphorus. The identity of Hesperus and Phosphorus is not at issue in the conversation*
>
> **Bob:** Phosphorus is my favorite planet.
>
> **Anna:** Yes, it's so shiny!

Let us name *a*, Anna's symbol, and *b* and *b*′, Bob's symbols for Venus. In ANNASPHORUS, by hypothesis, the mutual lexicon of Anna and Bob is such that *a* is connected by strategic path to both *b* and *b*′. That is, Anna is disposed to interpret both 'Phosphorus' and 'Hesperus' as *a*, and she is disposed to express *a* with both words, where Bob interprets each name with a different symbol and express each symbol with a different name. However, the structure of their mutual lexicon is not transparent to Anna and Bob from the communicative exchange they have. (Of course, by repeating interactions on the conversational topic, they could at some point find out. But they also might not).

On reflection, that alignment fails to be transparent should come as no surprise: why should we expect of communicative dispositions relative to a conversational topic that they be entirely manifest in each interactions where these dipositions are in part exercised?

My point of contention with Cumming comes from the fact that he himself invokes the alleged transparency of the alignment relation[58] as a reason to prefer it to the relation of membership in causal-historical chains — which we have seen is not transparent.[59] For example, Cumming says:

[57] In section 9, I propose to extend the domain of the alignment relation to regular+indexed mental symbols. The pragmalignment criterion (as I will call it), because it is sensitive to the implicit assumptions participants can make about the structure of the relevant segment of their interlocutor's lexicon, seems to better handle cases like FALSEPHORUS than Cumming's alignment relation.

[58] Consider this passage:

> The a priori argument of this section and the information-theoretic argument to follow converge on a same-content relation that is (i) an equivalence relation, (ii) intrasubjectively perspicuous, and (iii) *intersubjectively perspicuous* (i.e. communicative policies between agents track content). (Cumming 2013b : 387; my italics)

[59] See my discussion of Onofri's *indirect linking* relation (sections 2–6 above), which applies to any reachability relation in causal-historical networks.

> Causal-historical connection is not perspicuous. It takes empirical investigation to determine that different nodes belong to the same causal-historical tree. One could easily find oneself in the position of Kripke's (1979) character Peter, who possessed two symbols that, unbeknownst to him, referred via the same causal-historical network to the same individual. If causal-historical connection entailed content identity, then Peter's two symbols would have the same content. No matter how vigilant he was, Peter could end up with conflicting attitudes towards the same content—agreeing and denying at once that Paderewski was a great musician. (Cumming 2013b: 391)

I agree with everything Cumming says in the quote, but (as far as I understand his proposal correctly) parts of what he says in the quote seem to apply to his *alignment* relation just as well. So I do not think he can invoke the property of intersubjective transparency to promote his own view against the causal-historical model.[60] Cumming does consider cases of violation of intersubjective transparency. For example, he says:

> The mechanism of intersubjective perspicuity —communication— can also fail in practice. The speaker might misspeak; the hearer could mishear. This can result in an error: the content of the construal may not match the content expressed. (Cumming 2013b: 384)

However, in mentioning the reasons he cites for the failure of intersubjective transparency (performance errors like slip of the tongue, poor signal hearing), Cumming seems to overlook the possibility that the failure of transparency of his proposed intersubjective *same-content* relation could be rather the norm than the exception, due to the external nature of his *same-content* relation.

As an upshot: If my argument is correct, then Cumming's argument from transparency against causal-historical models of samethinking *does not go through*, because his *same-content* relation equally fails to be transparent.

8.1.2 Relativizing alignment makes the individuation of MOPs unstable

The intersubjective *same-content* relation based on alignment interferes with the individuation of mental symbols, because it interferes with the individuation of intrasubjective content.[61] In effect, alignment is meant to combine with the intrasubjective criterion to produce a general

[60] A noteworthy limitation of my objection is that I am operating with a notion of *access*/higher-order transparency (beliefs about sameness or difference of contents) whereas Cumming might operate with a *functional* (first-order) notion of transparency (see again Wikforss 2015 for the distinction). It seems possible for a thinker to be mistaken at the higher-order level, but there is a match at the first-order level. I won't discuss this issue here.

[61] I move freely from talk of MOPs to talk of mental symbols, on the assumption that MOPs can be construed as mental symbols.

same-content relation.[62] I will argue that this interference makes the individuation of intra-subjective content unstable — an indesirable result. To show this, let's introduce a feature of Cumming's view I have not presented yet.

Alignment is not an absolute notion, in Cumming's framework. Instead, it is relative to networks of agents.[63] Consequently, the *same-content* relation is likewise *relativized* to networks of agents. Two mental symbols may be misaligned relative to the speech community at large, but can be aligned relative to a given subset of agents of that community (Cumming 2013b: 387). For example, given three agents A, B and C, consider the configuration depicted in Figure 3.14:

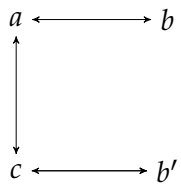


Figure 3.14 – $(\rightleftharpoons)$ is relative to sets of agents. Here, $a \rightleftharpoons_{\{A,B\}} b$ but **not** $a \rightleftharpoons_{\{A,B,C\}} b$

Let me provide an example of communicative sequence structured by the configuration of communicative strategies depicted in Figure 3.14. I will rely on a famous example given by Frege (1918) (We have already examined a version of this example given by Tayebi 2013, see TAYEBI in chapter 2):[64]

> Now if both Leo Peter and Rudolph Lingens mean by 'Dr. Gustav Lauben', the doctor who is the only doctor living in a house known to both of them, then they both understand the sentence 'Dr. Gustav Lauben was wounded' in the same way; they associate the same thought with it. But it is also possible that Rudolph Lingens does not know Dr. Lauben personally and does not know that it was Dr. Lauben who recently said [in his presence] 'I was wounded'. In this case, Rudolph Lingens cannot know that the same affair is in question. I say, therefore, in this case: the thought which Leo Peter expresses is not the same as that which Dr. Lauben uttered. (Frege 1956: 358)

Let's construe this example as follows. The *first* communicative exchange is between Dr. Lauben and Lingens. Dr. Lauben says to Lingens: 'I was wounded'. Since Lingens does not know Dr. Lauben personally, *he does not recognize* Dr. Laubens. Still, intuitively, Lingens understands what his interlocutor says. In the *second* communicative exchange, Dr. Lauben tells Leo

[62] See Cumming 2013b: 386.

[63] Cumming considers the view according to which the relation of alignment is relative to *communicative paths taken*, and give some reasons to reject it. The difference between relativizing alignment to *a path* and relativizing it to *a network of agents* is subtle: paths are *sequences*, whereas networks of agents are *sets*. In a sequence, unlike a set, the same elements can appear multiple times, and the order matters.

[64] Frege's example is quoted in Cumming (*in press*).

Peter: 'I was wounded'. Here Leo Peter, unlike Lingens, is able to identify Dr Laubens as his interlocutor. In the *third* communicative exchange, Leo Peter says to Lingens: 'Dr. Lauben was wounded'. As Frege remarks, here Leo Peter and Lingens understand each other. In particular, they both know that what one refers to with 'Dr Lauben' is what the other refers to with the same expression. However, as Frege also suggests, we cannot say that the thought Lingens entertains when he understands Leo's utterance, is the same as the thought Lingens entertains when he understands what *Dr. Lauben* said to him. Again, this is because, in the exchange with Dr. Lauben, Lingens did not (and could not) recognize Dr Lauben.

Let us call a, Dr. Lauben's symbol for himself. Dr. Lauben is disposed to express a by uttering 'I', and Lingens is disposed to interpret 'I' (when uttered by Dr. Laubens) as b. Moreover, Lingens is disposed to express b as 'you' (in the context), and Dr. Lauben is disposed to interpret 'you' (when uttered by Lingens, in the context) as a. As a result, $a \rightharpoonup b \wedge b \rightharpoonup a$. Now, the same is true with respect to Leo Peter: in particular, Leo Peter is disposed to interpret 'I' (said by Dr. Lauben) as c. So $a \rightharpoonup c$. Likewise, in the third communicative exchange of the sequence, Leo Peter is disposed to express c by uttering 'Dr. Lauben', and Lingens is disposed to construe 'Dr. Lauben' as b'. Crucially, b and b' are different mental symbols. This reflect the fact that Lingens fail to identify his interlocutor as Dr. Lauben in the first communicative exchange. So Frege's example, interpreted as I did, is an instance of the configuration depicted in Figure 3.14.

The alignment relation being relativized to sets of agents, we have $a \rightleftharpoons_{\{A,B\}} b$ but **not** $a \rightleftharpoons_{\{A,B,C\}} b$. So Dr. Lauben and Lingens's symbols — a and b — have the same content relative to the first communicative exchange. But Lingens's symbol — b — has a *different* content from a if we take into account a larger network of agents including Leo Peter. This is because there is a second communicative path through Leo Peter connecting a and b', so we cannot say that a and b share content anymore, or we will have to deny the difference in content between b and b' (a familiar consideration by now).[65]

Let us write the content of a mental symbol x, $\ulcorner [x] \urcorner$. The content of b thus shifts from $[a]$ to something else altogether, depending on which sets of agents we consider. But on the other hand, recall Cumming's individuation criterion of content at the intrasubjective level:

> Symbols x, y, belonging to the same agent, have the same content iff $x = y$

There is something amiss, here. We are told that intrasubjective content is individuated by mental syntax. The mental syntax of a given agent — in other words, a given agent's lexicon — is not relative to which sets of agents we consider. It seems to be an intrinsic property of the agent. So the contents of a given agent's lexicon should not change depending on which sets of agents we consider. But, on the other hand, we are told that *alignment* combine with the intrasubjective criterion to provide a general *same-content* relation which is relative to networks

[65] And it seems to be what Frege had in mind in the quoted passage above.

of agents. These claims do not seem to fit well with each other. If content is individuated by mental syntax at the intrasubjective level, then we cannot combine this intrasubjective criterion with the *same-content* relation by *alignment* relativized to networks of agents. Rather, it seems that two different levels of content are involved here: one corresponds to MOP content proper, the other (individuated by alignment relativized to networks of agent) corresponds to a surrogate content which can be shared. Otherwise, it seems that we disrupt the individuation of the MOPs. More specifically, it seems that Cumming thereby commits to a distinction between MOPs *qua* vehicles and MOPs *qua* contents. But then he cannot individuate MOPs content in terms of vehicle identity after all.

8.1.3 Comparison of condition (ii)* of the modified *indirect linking* criterion with ($\rightleftharpoons$)

PIERRE again As the reader might have noticed, my PIERRE example is not quite an instance of communicative sequence structured by the configuration of communicative strategies depicted in Figure 3.14. Instead, it is the following (let p and p' be **P**'s mental symbols associated with 'Londres' and 'London' respectively, q and r the mental symbols for London/Londres of agents **Q** and **R**, resp.):

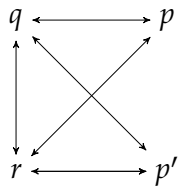


Figure 3.15 – The network of communicative policies in PIERRE

In PIERRE, we have $q \rightleftharpoons r$. We have $p \rightharpoonup q \wedge p' \rightharpoonup q \wedge q \rightharpoonup p \wedge q \rightharpoonup p'$. So p and q are not aligned, no matter which network of agents we choose. (Saying otherwise would trivialize the constraint of alignment). Likewise, we have $p \rightharpoonup r \wedge p' \rightharpoonup r \wedge r \rightharpoonup p \wedge r \rightharpoonup p'$. So p' and r are not aligned, no matter which network of agents we choose. Hence the constraint of alignment predicts that communication fails between **P** and **Q**-**R**.

By contrast, the modified *indirect linking* criterion I have proposed in section 5 —with condition (ii)*— allows that **P** and **Q** or **R** may share a thought in PIERRE. To show this, consider the graph that represents the communicative chain in PIERRE, and the partition of that graph as depicted with the blue line in Figure 3.16:

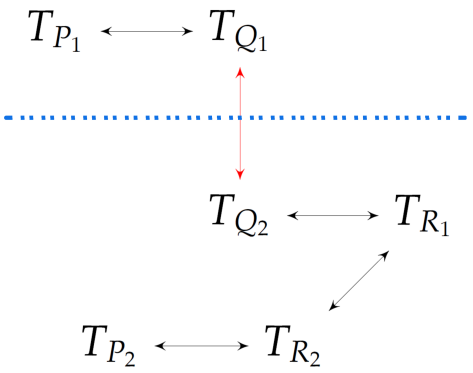


Figure 3.16 – Indirect Linking with (ii)*
NB: the graph relation is the *linking* relation and not (⇀)

The partition depicted in blue is determined by the following **cut-set**:

$$C = (T_{Q_1}, T_{Q_2}).$$

A cut-set is the set of links that have one endpoint in each cell of the partition. The cut depicted in blue is an admissible *same-thought* partition, according to condition (ii)* of the modified *indirect linking* criterion. But there are other admissible partitions.[66] According to the aforementioned partition, the sequence $< t_{P_1}, t_{Q_1} >$ is one distributed thought, whereas the sequence $< t_{Q_2}, t_{R_1}, t_{R_2}, t_{P_2} >$ is another distributed thought. So alignment relativized to networks of agents, and the modified *indirect linking* criterion are not equivalent. The latter allows Pierre to share thoughts with misaligned agents, not the former.

I mentioned PIERRE again to highlight that the relativization of alignment to networks of agents does not trivialize the constraint of alignment. For example, the communicative episodes in PIERRE are ruled out. The same applies to all the examples I am going to discuss in the nextsubsection: they are still ruled out by the relativized constraint of alignment, no matter which network of agents we choose as a parameter.

As an uphshot: if we take the relativisation of (⇌) to sets of agents seriously, it seems that we disrupt the individuation of MOPs content in terms of vehicule identity, because a given vehicle can be attributed different contents depending on which sets of agents one considers. The content of a mental symbol ends up being relative to the network of agents we consider. This commits to a form of semantic *localism*, as we might say. (This problem is reminiscent of the issue sophisticated Fregeans face when they allow MOPs to be collected differently depending on the situations of use.) I will now discuss whether alignment as a necessary condition for successful communication lines up with the intuitive demarcation of communicative success.

[66] Hence the felt arbitrariness of my modified *indirect linking* criterion, as already pointed out. The modified criterion is better understood without Shareability. Or so I argue in the last chapter.

8.2 Communication between misaligned agents

Cumming construes the alignment relation as a background condition for successful communication between a speaker and an audience. A corollary of this is that misaligned agents cannot successfully communicate, which implies the following: it is never the case that context allows misaligned agents to successfully coordinate.[67] I will call this hypothesis, *no pragmalignment*:

> **No Pragmalignment**
> It is never the case that context allows misaligned agents to successfully coordinate.

I will argue that this claim does not line up with the intuitive demarcation of communicative success. To show this, I will rely on judgement about cases of communication with proper names. But judgement about cases are seldom conclusive on their own. As pointed out in chapter 2, given certain relevant factors, we are quite happy to tolerate some degree of communicative failure. 'Good enough' communication is in fact very generally imperfect. So intuitions cannot be deemed to track successful communication: at best, they can be said to track successful *or* good enough communication. What is required, then, is that judgements about cases be given theoretical support. Here, we are not totally helpless. Capitalizing on my argument in chapter 1, I suggest we can motivate a diagnosis of communicative success or failure in a scenario by considering whether there is *knowledge* of what is said in that scenario (where knowledge of what is said precludes luck).

This methodology gives us a recipe for arguing against *(No Pragmalignment)*: we need to find cases such that **(a)** agents are relevantly misaligned but **(b)** the hearer comes to *know* what is asserted by the speaker's utterance, in virtue of exercising her capacity to understand utterances involving the name. Can we find such cases?

8.2.1 Case study: PIANIST 1 & 2

I think we can. I will rely on Kripke's Paderewski case (Kripke 1980). Having overheard conversations in two different settings about Paderewski, Peter comes to the view that there are two persons called "Paderewski": one a musician, the other a politician. As a result, Peter has two different concepts for Paderewski. In fact, he is wrong because there is only one person called 'Paderewski', both a musician and a politician.[68] The other protagonist in my example is Anna. She has only one concept for Paderewski.

Let us call Anna's concept, a and Peter's concepts, p and p'. Peter and Anna's concepts for Paderewski are not aligned. In particular, $a \rightharpoonup p \wedge p \rightharpoonup a \wedge a \rightharpoonup p' \wedge p' \rightharpoonup a$. In other words, there

[67] In what follows, I use 'misaligned communication'/ 'misaligned agents' as a shorthand for "communication between agents whose relevant concepts are not aligned".

[68] Were Peter to realize that his use of 'Paderewski' is in fact wrong/misaligned with the rest of the speech community, he *ought* be disposed to modify his use so as to track the *correct* use. This is an important normative fact. I address the issue how to explain such facts in the last chapter of the thesis.

are two strategy-adherent interpretations of a in Peter's lexicon (namely, p and p'), conversely, there is one and only one strategy-adherent interpretation of both p and p' in Anna's lexicon (namely a) — as depicted below[69] (Figure 3.17). Now consider the following conversation:

PIANIST:

Peter: Paderewski is playing next week. I have been told he is a talented pianist. Would you like to go to his next concert?

Anna: Yes, with pleasure! He is brilliant indeed.

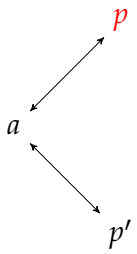


Figure 3.17 – Peter and Anna's concepts are misaligned.
(The red node means the corresponding symbol is the one activated in Peter's mind during the conversation.)

In PIANIST, it seems that Anna is not lucky in retrieving the correct referent. She is able to understand utterances involving the name 'Paderewski'. In this occasion, the intuition is strong that she *knows* what is asserted in Peter's utterance. Anna's use of "he" is anaphoric on Peter's use of 'Paderewski'/'he'. Each of Anna and Peter apparently understand what the other says. Either of them seem warranted in trading upon coreference of their respective uses of singular terms. As a result, we are tempted to say that both Anna and Peter know that the thought Peter expresses with his utterance and the thought Anna entertains when interpreting the utterance corefer. As a result, among the beliefs Anna forms on the basis of understanding what Peter said, the belief e.g. *that Paderewski is playing the week after* amounts to knowledge — at any rate, if the belief does not amount to knowledge, it does not seem due to any *communicative* shortcoming: perhaps Peter was misinformed, or maybe Peter is lying to lure Anna into a trap. But, the inference from Anna's grasp of what Peter said to her aforementioned belief does not seem unreliable. In short, we have reasons to think that communication succeeds. Yet interlocutors have their concepts misaligned, as shown in Figure 3.17. Because of this, the *Cummingian* is committed to the prediction that communication must fail in PIANIST. But this sounds stipulative.

[69] I explain the notion of *activated concept* and the role it has in successful coordination (on my view) in the next section.

One might object that the example is *too easy*. After all, *Peter* is the participant with the unorthodox idiolect, and Peter is not primarily in the position of a hearer. My opponent might object that this feature of the case is making the circumstances overly favorable. But we can switch the roles, and still have a case of successful communication (or so I submit):

PIANIST 2:

Anna: Paderewski is playing next week. I have been told he is a talented pianist. Would you like to go to his next concert?

Peter: Yes, with pleasure! He is brilliant indeed.

In Peter's lexicon, 'Paderewski' is ambiguous between two names, as shown in Figure 3.17.[70] But Peter has no difficulty disambiguating Anna's utterance on that occasion: he retrieves the referent by using the information that the referent *is a pianist*. That is, Peter disambiguates the word form 'Paderewski' by accessing the information *is a pianist* in his relevant file (the one represented by the node in red in Figure 3.17). The other PADEREWSKI symbol of Peter (i.e. p') is not deployed, and not activated in the episode: it is causally inert, because it is irrelevant.[71] The information extracted from Peter's interpretation is immediately sorted into the repository associated with p (i.e. the mental symbol associated with the information *is a pianist*). Accordingly, the intuition is strong that Peter is not lucky in retrieving the referent: he comes to *know* what Anna said. Both participants know that the thoughts they respectively deploy corefer. This might be, I submit, a case where *context allows misaligned speech participants to successfully coordinate*.

I have suggested by intuitions on cases that misaligned agents *might be able* to successfully communicate (namely, when contexts allows). What I mean by this is not merely that misaligned agents might be able to enjoy 'good enough' communication. I mean something stronger: that misaligned agents might be able to communicate in such a way that the hearer can *know* what was said by the speaker's utterance. There is perhaps an ideal of perfect communication that is impossible for misaligned agents to achieve. But this ideal of perfect communication is not always what matters for successful communication, or so I am suggesting.

Cumming might object as follows. When agents are misaligned, the hearer cannot grasp the content expressed by the speaker's utterance, because the concepts that the speaker and the hearer respectively deploy differ in content. As a result, it is not the case that a hearer whose relevant concepts are misaligned with the speaker's may genuinely *understand* the utterance. A misaligned hearer cannot achieve knowledge of *what was said*, because the concepts she uses

[70] Recall that strategies of expression and construal are individualistic extensions of I-languages. See Kaplan 1990 and Fiengo & May 2006 for the claim that Peter is confused about words.

[71] I anticipate on notions I will present in the section 9 of the chapter. Here I merely rely on intuitive observations: Peter does not think at all about the politician in the conversation.

to interpret the utterance do not match in content with those of the speaker.

Let me say why I find this move question begging. What is at issue is whether successful understanding requires fine-grained content sharing. Therefore, it cannot be *assumed* that an audience can come to know what was said by a speaker only if the interpretation of the hearer and the thought of the speaker match in their fine-grained content.

The Cummingian might object on a different ground, as follows. Assuming there is a safety requirement on knowledge, it is not the case that misaligned agents can achieve knowledge of what is said, because misaligned agents do not meet the safety requirement for knowledge of what is said. The objection can be formulated as follows:

Argument from safety

(1) Successful communication requires knowledge of what is said.

(2) Knowledge requires safety.

(3) In particular, knowledge of what is said requires safety.

(4) Utterance interpretation with a misaligned concept is never safe.

(5) Therefore, utterance interpretation with a misaligned concept never constitutes knowledge of what is said.

(6) Therefore, misaligned agents cannot communicate successfully.

Before I illustrate the idea expressed in premiss 4 on an example, let me introduce a similar notion for the analysis of knowledge in epistemology, which I believe can be used to make the argument for alignment even stronger. The idea is that in order to know a proposition, a subject must be able to rule out competing relevant hypotheses (or "Relevant Alternatives") to that proposition. In short, S knows that p only if S is able to rule out *all relevant alternatives* to p. I illustrate how this idea may apply to communication with an example.

> BARRY: Anna is at the library, and bumps into a professor from the philosophy department. The professor knows Anna is a philosophy student, but otherwise knows almost nothing about her. Meeting her eyes, the professor says: "Come see Barry Smith next week, he's giving a lecture." Anna happens to know *two* philosophers named "Barry Smith"(one is an ontologist from Buffalo, the other is a philosopher of language from London).[72] She forms the belief that Barry Smith *the philosopher of language* is giving a lecture. In fact, her belief turns out to be true.

[72]This example is based on facts. One Barry Smith is this philosopher: https://philpeople.org/profiles/barry-c-smith, the other Barry Smith is this philosopher: https://philpeople.org/profiles/barry-smith.

In BARRY, Anna is not able to rule out the relevant alternative that the intended referent is Barry Smith *the ontologist*. Because of this, she was simply lucky that her interpretation turned out to be correct. Contrast with this version of the example:

> BARRY 2: Anna is at the library, and bumps into a professor from the philosophy department. The professor knows Anna is studying the philosophy of language. Anna recognizes the professor, as she was her student in a philosophy of language seminar. Both Anna and the professor know that each of them is only interested in the philosophy of language, and in no other area of philosophy. The professor tells Anna: "Come see Barry Smith next week, he's giving a lecture." Anna happens to know *two* philosophers named "Barry Smith". She forms the belief that Barry Smith *the philosopher of language* is giving a lecture. Her belief is true.

In BARRY 2, Anna associates relevant information with the name employed in the utterance, which enables her to rule out the competing hypothesis that the intended referent is Barry Smith *the ontologist*.[73] As we may say, there is joint awaraness between Anna and the professor that the referent *is a philosopher of language*, information which is part of the intended ib-feature. As a result, Anna's interpretation amounts to knowledge of what is said.

What the pair of examples BARRY 1-2 suggests is that knowledge of what is said requires *being able to rule out competing relevant interpretations*. I find this idea very compelling. But we are now in a position to formulate another version of the argument from safety on behalf of Cumming:

Argument from Relevant Alternatives

> **(1)** Successful communication requires knowledge of what is said.
>
> **(2)** A subject S knows that *p* only if S is able to rule out *all relevant alternatives* to *p*.
>
> **(3)** In particular, a subject S knows that *p* is *what is said* only if S is able to rule out *all relevant alternatives* to *p*.
>
> **(4)** Misaligned agents are never in a position to rule out all relevant possibilities of misunderstanding.
>
> **(5)** Therefore, utterance interpretation with a misaligned concept never constitutes knowledge of what is said.
>
> **(6)** Therefore, misaligned agents cannot communicate successfully.

In both arguments, the crucial premis is premiss 4, which says that utterance interpretation with a misaligned concept is never safe. Now, why should we think that misaligned communication is always unsafe, or that misaligned agents are never in a position to rule out relevant

[73] Provided she does not associate the same information with the other Barry Smith, see Gray 2016 on disambiguation and differential belief.

possibilities of misunderstanding? Let me elicit intuitions on cases. Imagine the following sequel to the PIANIST example above:[74]

8.2.2 Case study 2: PADEREWSKIS

PADEREWSKIS:

Anna I saw Paderewski today. He looks very interesting.

Peter Which Paderewski are you talking about?

Anna What do you mean by "which Paderewski"?

Whereas in PIANIST context allowed Anna and Peter to coordinate, in PADEREWSKIS, communication is going wrong. Peter is unsure 'which Paderewski' is being talk about. We may suppose he has *both* of his mental symbols for Paderewski *activated* when interpreting the utterance.[75] Communication fails. Here, clearly, misalignment has a role in the breakdown in communication between Anna and Peter. But, by the safety requirement on knowledge of what is said, we should be inclined to *re-evaluate* our judgment on PIANIST in light of PADEREWSKIS: are we so sure that conversation was successful in PIANIST after all? One reason for reevaluating our judgement is that *alignment* between Anna and Peter would have ruled out the way in which communication went wrong in the PADEREWSKIS episode. But this suggests that, appearances notwithstanding, Anna and Peter were not able to rule out all relevant possibilities of misunderstanding in PIANIST already! So communication was not successful in PIANIST after all, or so the argument goes.

I believe that there is substance to the worry about safety. In fact, I accept all premisses of both arguments, except premiss 4: I will argue they are too strong. My take on these arguments will point towards a different picture of communication than the one alignment involves. I will argue that, once we undertand the *context-sensitivity* of the safety requirement, we should make the constraint of alignment *context-sensitive as well*. My argument will span until the next section (included), where I propose my own pragmatic version of the alignment relation. In closing this subsection, let me just outline the line of argument I will deploy.

We can accept that successful communication requires knowledge of what is said; that knowledge of what is said requires being able to rule out all relevant possibilities of misunderstanding;

[74] Michael Murez suggested me this sequel to PIANIST 2:

Anna: Paderewski is playing next week. I have been told he is a talented pianist. Would you like to go to his next concert?

Peter: Yes, with pleasure! He is brilliant indeed.

Peter: Wait, no! Actually, I planned to attend a political meeting with Paderewski at that time.

Such scenario urges us to revise our judgment that communication was safe in PIANIST 2.

[75] In the next section, I argue that this feature of the example is critical to explain the breakdown in communication here.

and that alignment between agents' concepts makes communication safer in making it the case that it could not easily have gone wrong, and still find alignment arbirarily too stringent.

For one thing, in "knowledge of what is said requires being able to rule out *all relevant possibilities of misunderstanding*", the expression in italics is context-sensitive, and depends on the conversational context. This should not be controversial, for English quantifiers like 'all' are obviously context-sensitive. But once we observe this, the problem with the alignment constraint becomes apparent. The assumption that misalignment *systematically* defeats knowledge of what is said, *irrespective* of facts about the circumstances of utterance, commits to a form of *infallibilism* applied to communication. The idea behind alignment as a background condition on successful communication may be phrased as follows: if one cannot rule out every possibility of misunderstanding, no matter what the conversational context is, then one cannot achieve perfect communication, so one cannot successfully communicate.

I suggest that this rigid demand on the interlocutors' mutual lexicon overlooks the role of context with respect to the safety requirement. Misalignment is not a *systematic* defeater to knowledge of what is said. It is *sometimes* a defeater. What determines whether misalignment actually defeats knowledge of what is said depends on the conversational context. So here is a version of the argument from safety which I think should replace the previous ones:

Argument from Relevant Alternatives REVISED

> **(1)** Successfull communication requires knowledge of what is said.
> **(2)** A subject S knows that *p* only if S is able to rule out *all relevant alternatives* to *p*.
> **(3)*** In particular, a subject S knows that *p* is *what is said* only if S is able to rule out *all relevant alternatives* to *p* (i.e. relative to the conversational context).
> **(4)*** Depending on the conversational context, misaligned agents may or may not be in a position to rule out all relevant possibilities of misunderstanding.

This version of the argument is compatible with successful misaligned communication. But the argument does not tell us *how* context impacts the standards for knowledge of what is said in connection with the safety requirement. There are various options. Let me mention two.

One option is to say that the stringency of the safety requirement is a function of the purpose of the conversation in the context. In particular, roughly, the more *important* the purpose of the conversation is, the *harder* it is to know what is said (see e.g. Stanley 2005). This option is a version of the dotrine called *Pragmatic Encroachment* in epistemology. Here is a generic

formulation of the doctrine due to Ichikawa & Steup (2018):[76]

> **Pragmatic Encroachment:**
> A difference in pragmatic circumstances can constitute a difference in knowledge.

For example, in PIANIST, the purpose of the conversation is to coordinate on the action plan to go to the concert. This is rather low stakes. By contrast, if it had been a matter of life and death for Anna to understand what Peter said, the standards would have been much stricter. The general idea is that pragmatic factors such as the purpose of a conversation are relevant for determining whether a misaligned hearer's state constitutes knowledge of what is said.

Another, related (and already mentioned) option is to say that the extension of 'all' in *all relevant alternatives* depends on the conversational context. This enables us to say that misaligned agents may be in a position to eliminate the relevant alternatives in some cases, hence may be in a position to successfully communicate. The idea common to both options is to develop an anti-skeptical and faillibilist account of knowledge of what is said, once it is observed that *alignment* commits to some form of infallibilism with respect to knowledge of what is said.[77] If the conversational context allows, a misaligned interpreter may be able to rule out all relevant alternatives *relative to the context*. A hearer may acquire *knowledge* about the bearer of a name through an utterance, even if the interlocutors are relevantly misaligned. Rejecting this latter idea implies a form of scepticism about testimony, given (if I am right) the prevalence of misaligned knowledge-yielding communication.[78]

Either way, here is the claim the endorsement of which enables me to reject premiss 4 of each safety argument. I call this principle *pragmalignment (i)* because it is one of *two* tenets governing the pragmatic version of alignment I will argue for:

> **Pragmalignment (i)** The stringency of the standards related to the safety requirement for knowledge of what is said is not uniform, but depends on the conversational context. There are different standards in different contexts, related to different purposes, on what counts as understanding an utterance.

I would like to conclude on whether misaligned agents can communicate by adopting a moderate position. Which relevant possibilities of misunderstanding a hearer must be able to rule out are not independent of the conversational context. When context allows, misaligned communication *can* be knowledge-yielding. Therefore, I do not quite accept the alignment requirement on communication. In the next subsection, I point out that utterance interpretation is not wholly explained by the relation of alignment (a relation at the competence-level) but involves pragmatic reasoning which are out of the scope of the alignment-theoretic picture.

[76] See Pinillos 2012 for a series of experimental results which give some support to the pragmatic encroachment claim that ordinary people's attributions of knowledge are in fact sensitive to practical interests or stakes.

[77] It is open for the contextualist to say that the relevant contextual features are the practical stakes governing the speech participants' context, in which case, the two options merge.

[78] The phrase *knowledge-yielding* is from Peet 2019

8.3 Performative confusions

Alignment, and the ground relation it is based on, are idealizations. Alignment is a constraint on the paths of equilibrium-yielding communicative policies.[79] But the ground relation already involves a great deal of idealization, because it is a relation at the competence level. Here I point out that alignment cannot account for communicative failure between aligned agents when the communicative failure is due to performance factors.[80]

Alignment is a relation between speakers' idiolects. Given two agents A and B, and two of their mental symbols a and b, a strategy profile —or path i.e. if the connecting path is indirect— between A and B maps a to b or b to a in virtue of a lexical convention that they share (among themselves or through other agents). We may imagine that which signal is used in a given strategy profile between two agents may change depending on the context. For example, when I talk with my medievalist friend about his work, we coordinate our symbols with 'Thomas'. But this is not the signal we use in a broader context – we'll prefer 'Thomas Aquinas' instead.

It may be the case that, in a discourse situation, more often that not, the same signal will realize the same name. So for instance, given a context of utterance, the various occurrences of 'Thomas' will typically realize the same name, that is, actualize a lexical convention that we share. But in general, it is not the case that, more often than not, the same signal in *different* contexts will be the realization of the *same* name.

This is where the need for disambiguation enters the picture. The networks of communicative dispositions with links of equilibrium-yielding reference-preserving joint-strategies are an idealization. They are about competence. Strategic profiles actually link mental symbols as long as there are no *performance errors* in the execution of the strategies that connect agents' symbols. For example, what the speaker says is not always what the hearer hears. But a salient source of 'performance error' is ambiguity.[81] For instance, to take the example of my medievalist friend, even if this versatile lexical convention is in place between us ('Thomas' in private talks, 'Thomas Aquinas' otherwise), our strategy profile *actually* links our symbols only if we disambiguate the signal adequately. If my medievalist friend construes 'Thomas' as referring to our common hellenist friend when the conversational topic is Aquinas (perhaps I changed the subject a little abruptly without thinking of the possible confusion), then communication fails.

[79] The relation *is related by a communicative path to* is the transitive closure of the relation ($\rightharpoonup$) (which is not transitive).

[80] In chapter 5, I define a less idealized notion of joint communicative strategy than Cumming's, which I write ($\searrow$). In particular, ($\searrow$) need not be reference- or subject matter- preserving, unlike ($\rightharpoonup$). For example, *Mrs Malaprop*'s symbols may be ($\searrow$)-related to the symbols of a competent English speaker in virtue of their respective grammars, even if the joint strategy is not an equilibrium.

[81] I don't mean that we typically need to think hard to disambiguate proper names. For instance, among philosophers, we typically interpret the signal 'Aristotle' without considering competing targets.

What the last paragraph shows is that strategies of construal must be equipped with *disambiguation strategies* in order to be operative. That is, alignment involves *disambiguated* signals. But this is a huge idealization in the model. Let me illustrate what I mean on an example.

Case study 3: JOHN SMITH

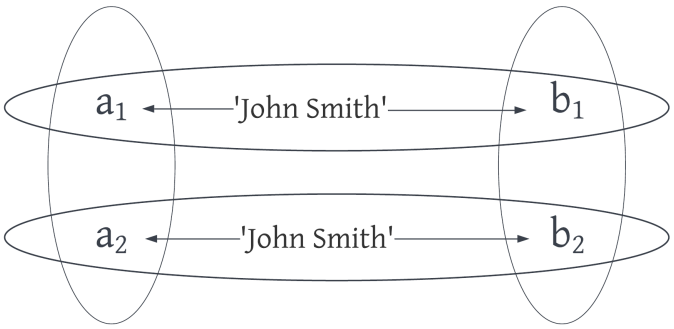


Figure 3.18 – Aligned agents

> JOHN SMITH: Agents A and B know in common two individuals both named 'John Smith'. A and B were both there when they were introduced to these two individuals by name, and both A and B memorized the names with the intention of using them in the normal way. In short, A and B have their relevant concepts aligned. There are both disposed to interpret correctly utterances involving 'John Smith' as meaning, in context, the relevant John Smith. One day, A and B meet. A is very tired, and gets everything muddled up. In particular, A forms beliefs about John Smith_1 on the basis of B's utterances, when B really is talking about John Smith_2.

In such a case, it does not really matter that $a_1 \rightleftharpoons b_1 \wedge a_2 \rightleftharpoons b_2$. It is true that, because their relevant concepts are aligned, A and B are disposed to successfully communicate about John Smith_1 and John Smith_2. Participants are also disposed to interpret in the relevant way in these sorts of circumstances. However, in that occasion, they just don't! In that occasion, because of performance factors, A's behaviour does not conform to his communicative disposition.

This indicates the (boring) idea that alignment is not all there is to successful communication. Even when the right communicative dispositions are in place, speech participants may still fail to communicate successfully (as is well-known, dispositions can always fail to manifest). Of course, in JOHN SMITH the performance error in itself does not call into question the fact that a_1 and b_1 and a_2 and b_2 share content (here, since the context is not hyperintensional, the Cummingian can say that shared content just is reference).[82] It's just that A failed to deploy the

[82] See chapter 2 for a definition of hyperintensional discourse context.

right mental symbol. In terms of the criterion I proposed in the previous chapter, a plausible diagnosis is that *A* and *B* failed to have joint awareness to the relevant ib-features.

I believe that my relatively trivial point about performance errors leads to a somewhat less trivial point about the scope of Cumming's proposal. Communication is a pragmatic phenomenon: which mental symbol we choose in order to interpret a given verbal signal, often depends on complex background information about the context (in particular, speaker's intentions as they are represented by the hearer). Hence strategies of expression and construal often are negotiated *on the fly*. For instance, demonstratives ('that', 'he', etc.) — so-called 'impure indexicals' — are terms such that the linguistic rules which govern their use are not sufficient to determine their referent in all contexts of use. A demonstrative used on a particular occasion refers to a salient individual, but salience belongs with speaker's intentions, not meanings. As Evans says, "there is no linguistic rule which determines that a 'he' or a 'that man' refers to *x* rather than *y* in the vicinity" (Evans 1985: 230). Hence the interpretation of such terms has an *ad hoc* character. We need a model that is sensitive to these pragmatic aspects in order to explain what goes on in the Loar case. For this reason, Loar cases involving demonstratives are not even in the scope of Cumming's framework.[83] Accordingly, in what follows I would like to motivate a more pragmatic conception of the alignment relation. For example, as I will explain, we might want alignment to be sensitive to the cognitive statuses of mental symbols in the minds of the speech participants.

9 Introducing pragmalignment

9.1 Relevant symbols

In Cumming's theory, the constraint of alignment concerns mental symbols as long as they are *possessed*. What is it for a mental symbol to be possessed? Cumming does not really tell us. I take it that, in order for a subject S to possess a mental symbol *s*, S needs to have *s* stored in a certain sector of memory — typically long-term memory. As we might say then, the only cognitive status alignment is sensitive to, is memory. Moreover, the mental symbols concerned by the constraint of alignment must all be linguistically expressible, as they must feature in joint strategies of expression and construal for them to be connected by communicative paths. What this last constraint exactly amounts to is not straightforward in Cumming's framework.

That mental symbols must interact with the language system in order to be in the domain of the relation of alignment, does not mean that they must be *lexicalized*, that is, expressible by an atomic lexical item that is stored. However, as already mentioned, it is not crystal clear how to individuate communicative strategies in Cumming's framework. As a result, it's not clear what

[83] As I understand Cumming, he would count Loar cases involving demonstratives as performance errors: 'a piece of behavior that does not conform to policy'. A hearer can fail to correctly interpret a demonstrative, even though she is disposed to correctly interpret the demonstrative.

the set of departure of the alignment relation is (Does it include *any* mental symbol stored in memory whatsoever which may interact with the language system?). One thing is clear: which public signals can be used in a joint communicative policy may vary with the context. But, as I remarked in my presentation of Cumming's ground relation (sect. 7), Cumming's talk of *grammar* may give the impression that a mental symbol must be expressible by an atomic lexical item that is stored in order to be in the domain of ($\rightleftharpoons$). If this reading is true, then Cumming's theory is exclusively about competence, not about performance — it is semantic and does not deal much with pragmatics. I think we should prefer a conception of the communicative strategies compatible with the fact that agents can make themselves understood in a variety of ways, using pronouns, names, definite descriptions, perhaps *ad hoc* signals. Accordingly, I will assume that a joint communicative policy ensures that there is a reliable way to communicate about some o between two agents, using one signal or another, provided that they both possess coreferential mental symbols wich may interact with their respective verbal system.

The domain of the alignment relation is insensitive to any notion of *relevance* regarding mental symbols. By this I mean that the mere *presence* of a mental symbol in memory storage (i.e. its mere availability) ensures that the mental symbol is in the domain of the alignment relation, provided that this symbol can interact with the verbal system. That is so even if the mental symbol is hardly *accessible* to the agent.[84] But such a conception is implausible. The intuition is strong that merely having a mental symbol stored somewhere in memory but hardly accessible, and which happens to be coreferential with a mental symbol used in a given communicative episode, is not enough to make the communication fail. Consider the following case:[85]

> BURIED MEMORY: Anna has known someone for a long time by the name of 'Robert'. On a certain occasion, she meets someone who looks exactly like Robert, but it seems to her that it is not Robert. She hears someone call this person 'Bob'. For complex reasons, it seems to her that the best explanation of the situation is that Robert actually has a twin brother named 'Bob'. Call the mental symbol she then tokens to think about this guy, a'. Subsequently, Anna totally forgets about this story. Ten years later, she meets Robert with other friends and says:
>
> **(1)** Bob is looking good!
>
> If Anna did enough psychoanalysis, the deeply buried memory about 'the twin' might emerge, but she doesn't think at all about the man she met on that very old occasion when she produces her utterance of (1). In other words, her mental symbol a' is not active in the context. Assume that Anna is disposed to believe that it is common ground between her and the audience that Robert has a twin called 'Bob'. Because alignment applies to mental symbols as long as they are *possessed* and can interact with the verbal system, the mere availability of a' makes Anna misaligned

[84] See Tulving and Pearlstone 1966 for the distinction between availability and accessibility.
[85] I'm indebted to Michael Murez for this style of example.

with Robert and the rest of the audience. For Anna is disposed to express her mental symbol a' with the signal 'Bob', and Robert is disposed to interpret 'Bob', in that particular context, as himself. Consequently, the alignment criterion predicts that communication fails.

Let a be Anna's symbol for her friend Robert, and b, Robert's symbol for himself. As just mentioned, a' denotes the mental symbol Anna tokened to think about 'Robert's twin'. The communicative policies which underlies the context of the utterance of (1) is as follows:[86]

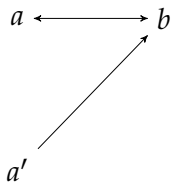


Figure 3.19 – Communicative policies in BURIED MEMORY

It should be intuitive that this deeply buried memory of Anna is *irrelevant* to the communicative episode. BURIED MEMORY suggests that the mere *availability* of a mental symbol in memory storage does not make it *relevant* to a communicative episode. It seems that a mental symbol must be *sufficiently accessible* to the agent, and be able to influence performance *in the discourse context* to constitute a possible defeater. In the mental lexicon of an agent (i.e. the set of the mental symbols she possesses), only a subset of that lexicon is relevant to any given thought episode, communication or reporting event. Which mental symbols are relevant vary with the discourse context. This notion of *relevance* applied to mental symbols, opens the way to contextual formulations of the constraint of alignment.[87]

What is it for a mental symbol to be relevant? In BURIED MEMORY, what made the memory of Anna irrelevant to the communicative episode was its inaccessibility to Anna in that communicative episode. Accordingly, as just mentioned, one idea suggests itself: a mental representation is relevant to a communicative or thought episode if it has *a certain degree of accessibility* in that episode. In other words, the mental symbols that are relevant to any communicative or thought episode might be the mental symbols *activated* in that episode. From now on, I will call *cognitive status* the property of a mental symbol having to do with its degree of accessibility in memory and attentional states. The cognitive status of a mental symbol is

[86] It's not clear that Robert would interpret the signal Anna would use to express a' as b, because Anna might introduce the dref attached to a' as "your twin, Bob. . . ". In which case, $a' \nrightarrow b$ and $a \rightleftharpoons b$. But I take it that this kind of examples exists, and the point goes through.

[87] The notion of relevance I am appealing to should be distinguished from the technical definition of relevance given by Sperber & Wilson 1996, who defines relevance in terms of what has contextual effect:

> *Relevance*$_{S\&W}$:
> An assumption is relevant in a context if and only if it has some contextual effect in that context.

As will be clear later in the text, the notion I have in mind is non-semantic and has to do with the degree of activation of a mental representation (which may be ultimately characterized in neural terms). How the two notions relate is an interesting question.

thus a determinable of which the property of *being activated* is a determinate.[88]

In fact, the idea that the cognitive statuses of mental symbols play a role in various aspects of linguistic communication (in particular, as they are assumed by the speaker) is part of a research program which I mentioned in chapter 2, called *Givenness Hierarchy* (GH) — see e.g. Gundel, Hedberg, & Zacharski, 1993; Chafe 1994, Féry & Krifka 2008.[89] For example, in

(1) I couldn't sleep last night. **It** kept me awake.

the speaker's use of **it** is felicitous only if the speaker is warranted in assuming that her conversational partner has their attention focused on the referent in question.[90] So in using expressions such as **it**, a speaker must monitor the cognitive status of the intended referent in the mind of her audience.

This is reminiscent of the notion of joint awareness on the ib-feature I have introduced in the previous chapter. It is worth making this slightly more explicit. Following Hedberg (2013: 1-2), we may call *referential givenness/newness* the class of ib-feature attached to singular terms that describe the relation between the intended referent of a linguistic expression, and its cognitive status in the memory or attentional states in the audience's mind (as assumed by the speaker). Hedberg mentions the following definition of Féry & Krifka 2008 for this class of ib-feature:

> A feature X of an expression α is a Givenness feature iff X indicates whether the denotation of α is present in the CG [=Common Ground][91] or not, *and/or indicates the degree to which it is present in the immediate Common Ground*.

The *Givenness Hierarchy* may be construed as a theory of meaning for the relevant class of linguistic expressions (*it*, *this/that/this NP*; *the NP*; *a NP*, etc). As Hedberg proposes:

> The Givenness Hierarchy (...) is a set of six 'cognitive statuses' (memory and attention states) in the mind of the addressee (as assumed by the speaker).[92] These statuses are claimed to *constitute meanings* of pronominal and determiner forms, and determine necessary and sufficient conditions on the use of each referring form in discourse. (Hedberg, 2013: 2; my italics)

[88] Of course, while *being activated* is a determinate of *cognitive status*, it might be a determinable of *being activated to degree x* for a given *x*.

[89] This research program has interesting ramifications in human-robot interaction modelling, see e.g. Pal, Zhu & Golden-Lasher 2020, Williams, Schreitter & Scheutz 2019.

[90] The example is presented in Hedberg 2013.

[91] I.e. (roughly) the set of propositions and references presupposed to be already shared between the speech participants. The conception of the CG offered by (GH) is thus much more rich and structured than the standard Stalnakerian conception, introduced in section 3 o the general introduction.

[92] Here is the Givenness Hierarchy:

> **in focus** > **activated** > **familiar** > **uniquely identifiable** > **referential** > **type identifiable**
> *it* > *IT (stressed)/ this/that/this NP* > *that NP* > *the NP* > indefinite *this NP* > *a NP*. (Hedberg op. cit.)

For example, in this framework, the meaning of 'it' is the following instruction: *associate representation in focus of attention (in focus)*. So, if the hearer is not in a position to associate such a representation, an utterance of 'it' fails to be felicitous. The meaning of 'this/that/this NP' is the following instruction: *associate representation in working memory (activated)*. Accordingly, an utterance of this category is felicitous if the hearer can associate an activated representation with the object under discussion. Likewise, the meaning of 'that NP' is the following instruction: *associate representation in memory (familiar)*. An utterance of 'this/that/this NP' is therefore felicitous if the hearer can retrieve a memory representation of the object under discussion. More generally, in the (GH) framework, the meaning of a pronominal or determiner expression is an instruction which directs the hearer to retrieve a certain mental symbol with a certain (assumed) cognitive status for interpreting the utterance. An utterance of this class of expressions is felicitous iff the speaker is warranted in assuming the hearer can associate a representation with the cognitive status assumed of the intended referent.

I don't have to take on board all the details of this (otherwise very interesting) linguistic theory here. Rather, I have mentioned (GH) to motivate the idea that the cognitive statutes of mental symbols play a role in linguistic communication: they do according to linguistic theories. What interests me in (GH) is the hypothesis that the use of a singular term in communication involves implicit assumptions about the cognitive status the referent has in the mind of the audience — an idea I want to incorporate when revisiting the alignment constraint. Relatedly, I have proposed that which mental symbols are *relevant* to any communicative episode are thoses with a certain degree of accessibility in the mind of the speech participants — the activated ones. I believe we have enough material to revisit some examples, and see if we get new light on those.

Recall PIANIST 2:

> **Anna:** Paderewski is playing next week. I have been told he is a talented pianist. Would you like to go to his next concert?
>
> **Peter:** Yes, with pleasure! He is brilliant indeed.

Now consider this sequel to PIANIST 2:[93]

> PIANIST 3:
>
> **Anna:** Paderewski is playing next week. I have been told he is a talented pianist. Would you like to go to his next concert?
>
> **Peter:** Yes, with pleasure! He is brilliant indeed.
>
> **Peter:** Wait, no! Actually, I planned to attend a political meeting with Paderewski at that time.

Another example of bad case was

[93] Due to Michael Murez

PADEREWSKIS:

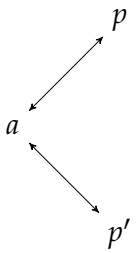
Anna I saw Paderewski today. He looks very interesting.

Peter Which Paderewski are you talking about?

Anna What do you mean by "which Paderewski"?

The underlying configuration of Anna and Peter's mutual communicative dispositions is as follows:

Figure 3.20 – Peter and Anna's concepts are misaligned.



But the contrast I want to point out between the cases can be represented as follows (Figure 3.21):

Figure 3.21 – Left: activated symbols aligned. Right: activated symbols misaligned
(activated symbols represented in red)

Here is the pattern that emerges from the sample of cases: the GOOD case seem to be ones in which the activated mental symbols are aligned. The BAD cases seem to be those in which the activated mental symbols are not aligned. This suggests that we could try to keep the constraint of alignment by restricting its domain to the set of *activated* symbols. Such a constraint does eliminate bad cases, but it is less stringent than ($\rightleftharpoons$), and allows misaligned successful communication when context is favorable. Such a criterion seems to make more intuitive predictions about cases than alignment. That being said, the sample of cases considered was very small, at any rate this diagnosis requires some theoretical support. The notion of an activated concept, and its role as a possible defeater of understanding need to be explained.

9.2 Restricting the domain of ($\rightleftharpoons$) to activated symbols

A mental symbol is deployed in an occurrent thought when it is a *constituent* of that thought. Any mental symbol that is deployed is also activated, is accessible in memory, and is possessed (as depicted in Figure 3.22). But a mental symbol can be activated without being deployed. So a mental symbol is deployed in utterance interpretation if it is activated and *contributes its semantic content* to the interpretation.

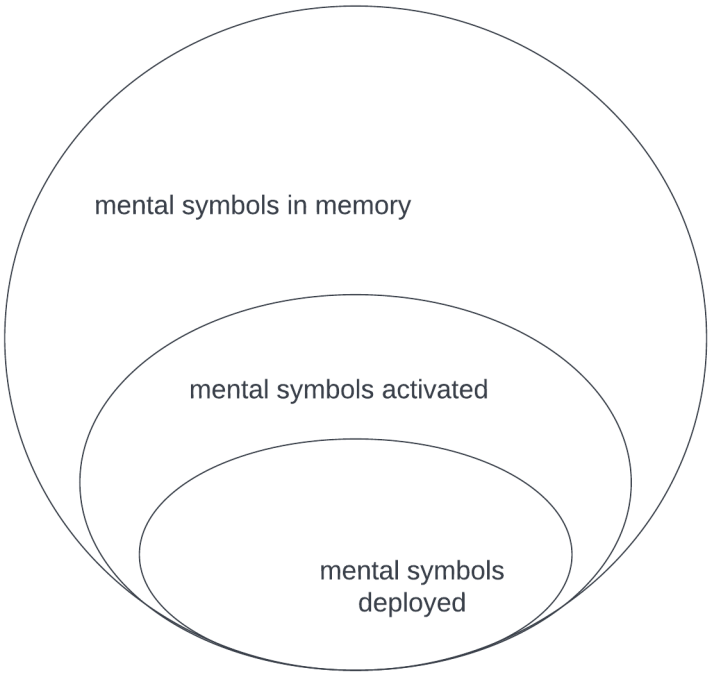


Figure 3.22 – Mental symbols deployed, activated, possessed

A mental symbol is deployed only if it has a sufficient degree of activation. One may cash out the property of being activated for a mental symbol in terms of attention, accessibility, and ultimately in neural terms.[94] Each mental symbol is matched to a repository of information, namely the predicative entries associated with the object the mental symbol is about. When a thinker deploys a mental symbol, she gets (roughly) access to the entries stored at that address. As mentioned, being activated is not sufficient for a mental symbol to be deployed. As Recanati puts it, quoting Fodor & al:

> To be a constituent of a given thought it is not sufficient to be active when the other constituents are. As Fodor and Pylyshyn point out, 'when representations express concepts that belong to the same [thought], they are not merely simultaneously active, but also in construction with each other' (Fodor and Pylyshyn 1988: 25). (Recanati 2016, op).

[94] Recanati, 2016, Preface.

On the *Langue-of-Thought-Hypothesis*-picture I am suggesting, the *deployed* mental symbols are what compose with each other to output complete occurrent thoughts, whereas the repositories of descriptive entries associated with the mental symbols are what explain categorization and induction — that is, concept use as they are studied by cognitive psychologists (see Murez 2021 drawing upon Glanzberg 2018).

If we restrict the domain of the alignment relation to the *deployed* mental symbols of agents, it seems to me that we thereby *trivialize* the constraint of alignment. This is because in most cases, there will be exactly one symbol deployed per intended referent: interpretations, more generally thoughts, are typically non-ambiguous. If the interpretation of an utterance involves more than one mental symbols for a given nominal expression, then that interpretation is arguably faulty, because it is ambiguous: it is not really an interpretation. So relativizing the constraint of alignment to the *deployed* mental symbols is a dead-end.[95]

But let's consider the hypothesis that the relevant mental symbols are the activated ones, that is, that the mental symbols relevant to the success of any given episode of communication are the ones that are *activated* in the interpretation and production of the sequence of utterances that make up the communicative event. Not all activated mental symbols are deployed, so the restriction does not seem to trivialize the constraint of alignment. As we saw, the criterion of alignment of activated symbols rules out *some but not all* cases of misaligned communication.

When a file is activated, the repository of information matched to that symbol is made more accessible (all this being a matter of degree, since activation is a graded notion).[96] When a given signal, such as a name, is ambiguous for the interpreter, disambiguating that signal may require accessing predicative entries associated with one's mental symbols (Gray 2016).[97] But only the deployed mental symbols contribute their referent to the semantic content of a thought. For example, if you deploy SALT, it makes more accessible the associated representations e.g. PEPPER. However this does not imply that your thought about salt has PEPPER as constituent.

[95] But see Gauker 2003 for the intriguing claim that thoughts can be ambiguous. Quilty-Dunn 2021 also argues that thoughts can be polysemous. He proposes that e.g. THERE IS AN OPEN SEAT BY THE DOOR might be entertained without resolving whether the door is the barrier (as in JOHN KNOCKED ON THE DOOR) or the aperture (as in JOHN WALKED THROUGH THE DOOR). Or, I LOVE FRANCE might be thought without having a particular denotation in mind (a piece of land, a government, a population, a culture, etc). I am not sure what I think about these examples, and I want to keep an open mind on the issue. That being said, it seems to me that the examples Quilty-Dunn adduces are more akin to thoughts like RED IS MY FAVORITE COLOR (thought without any specific shade in mind). That is, they are 'polysemous' in the sense that they are abstract/generic. To give another example, if I read and understand an utterance of the sentence "the square is next to the circle" without any accompanying illustration, I will entertain a generic thought (while the accompanying mental imagery I experience may be specific, picking out a favorite/spontaneous model, see e.g. Knauff 2013). This kind of thoughts do not seem to be polysemous in any interesting sense. Be that as it may, even if they are genuine examples of polysemous thoughts, I doubt that they affect the point I am making with respect to alignment. For example, Peter could not think a thought such as I LOVE PADEREWSKI and have a polysemous thought–relative to his lexicon.

[96] That means that communicative success is also a matter of degree, on this criterion. See my discussion below.

[97] We may call a *mental file*, the complex entity composed of the mental symbol + the repository of information it is associated with. See e.g. Perry 2012, Recanati 1993, 2012.

Drawing on Quilty-Dunn 2021, we may provide the following three conditions individually sufficient for use/deployment over mere activation:

(a) figuring in the computational process experienced as the interpretation of the utterance (in focus of attention);

(b) being moved into working memory;

(c) surpassing some threshold of activation.

How is all this relevant to communicative success? The idea I am exploring is that merely activated mental symbols (i.e. activated but not deployed) have a role with respect to the safety requirement for knowledge of what is said. They constitute relevant possibilities of misunderstanding which are not ruled out. Having more than one activated mental symbol related to the unique coreferential mental symbol of another agent is a predicament which reflects that communication could easily go wrong/is going wrong/may easily go wrong in the future.

Whether Peter has more than one of his mental symbols for Paderewski *activated* in a communicative episode matters to whether the communication succeeds. For example, Peter might have both of his symbols activated in interpreting an utterance, because he is unsure 'which Paderewski' is being talk about, as in PADEREWSKIS above. Or, he might have both of his symbols activated in producing a sequence of utterances, because he uses them to target putatively competing referents, as in PIANIST 3.[98]

The idea I am putting forward is that Cumming is *right* that mental symbols must be aligned for communication or reporting to be successful; but he is *wrong* in claiming that the constraint applies to all possessed symbols. If a symbol is not activated, the only sense in which that symbol can disrupt a conversation is remotely counterfactual, and if my view is correct, in many conversational contexts this kind of safety does not matter. Yet, Cumming is right that some information is lost in a misaligned configuration. An agent that has *two* symbols where another has only *one*, introduces spurious information (that is, noise), not matched by the agent who is not in a relevant Frege case.[99] But again, if my line of argument is correct, successful communication is not *always* perfect communication. I am not denying that there may be a notion of perfect communication worth studying. What I deny is that *this* notion is what "successful communication" means *in every context*.[100] I now come to the definition of alignment on activated concepts.

[98]The proposal might be worth relating to 'psychologistic' models of common ground, e.g. Maier 2016a, 2017 and Kamp 2015, 2013, 2019. This task must be left to one side for the present.

[99]See the reconstrual of alignment in terms of information, relying on Dretske 1981, in 2013b.

[100]Relatedly, if I am right, what "successful communication" means depends on features of the conversational context, such as the purpose of the conversation. This is the other tenet of pragmalignment.

9.2.1 Pragmalignment (alignment of the activated symbols)

Let us call *pragmalignment* the constraint of alignment as it applies to *activated* mental symbols in a discourse context.

> Mental symbols a and b (belonging to agents A and B) are **pragmaligned** in a context $\mathcal{C}$ iff:
>
> a. a and b are related by communicative dispositions in both directions;
>
> b. a is the only *activated* symbol in A's lexicon related to b in either direction;
>
> c. b is the only *activated* symbol in B's lexicon related to a in either direction.

It is my contention that *pragmalignment* is a more plausible constraint on successful communication than alignment. More work needs to be done to establish the empirical plausibility of the definition. In particular, I would need to engage psycholinguistic data, and models of the common ground that incorporate mental representations. I leave it to further research to draw the relevant connections to empirical research, and "psychologistic" theories of the common ground.

Since activation is a graded notion, accepting a necessary condition for communicative success in terms of activation implies a graded notion of communicative success. In my opinion, this is a happy result. One prediction it seems to entail is that the degree to which a communicative exchange fails will vary roughly in proportion with the degree to which the disrupting symbols are accessible. Again, some empirical work would be required to assess the soundness of this hypothesis. Another possible issue has to do with semantic priming. It is a well-documented effect that one can prime the activation of particular concepts of a subject e.g. by exposing the agent to visual words masked and presented extremely briefly.[101] We can imagine that in priming a misaligned agent, one may trigger the activation of the agent's disrupting symbols. But, intuitively, semantic priming does not (on the face of it) make such symbols relevant. The right reaction to this issue might be to impose that activated symbols must be *conscious* in order to be relevant, or something along this line. Relatedly, Michael Murez makes the following remark on the proposal:

> It seems like "relevant" could be understood in epistemological terms, but "activation" is a psychological notion. Why couldn't a relevant representation be one that the hearer did not happen to activate? (personal notes)

In response, let me stress the following. The constraint of alignment, however it is characterized, is consigned to sets of coreferential symbols (a symbol non-coreferential with another symbol cannot make the latter misaligned). So the notion of relevance here is very narrow: a mental symbol is relevant in the sense at issue just in case it is a *possible defeater* for knowing what is said or successfully communicating in a misaligned configuration. In this context,

[101] See Maxfield 1997 for review.

accepting that a relevant representation could be one the speech participant did not happen to activate amounts to accept alignment à la Cumming, defined on the set of possessed symbols. But I have argue against the idea that mere availability in memory storage of a symbol was enough to make it a possible defeater for communicative success in any context.[102]

Pragmalignment, I have argued, provides a better picture of samethinking in communication. In the next subsection, I explore another dimension which may be a factor in how speech participants can successfully coordinate in a misaligned configuration. The dimension I would like to bring to bear on the alignment constraint involves the capacity to represent the subjective context of another agent; I now turn to it.

9.3 Extending the domain of (⇌) to metarepresentational symbols

Pragmalignment is the constraint we get when we restrict the domain of the alignment relation to the set of activated symbols (relative to a discourse context). My rationale behind this restriction is that the domain of the alignment constraint was too broad, making (if my arguments are correct) the constraint arbitrarily too stringent. The *relevant* mental symbols are only a proper subset of the mental symbols *possessed*.

However, I also believe that the domain of alignment à la Cumming is in some respect *too narrow*, and that we should enrich the domain of the alignment relation. I have in mind the fact that a speech participant who is aware of the misaligned perspective of her conversational partner, might be able to take this into account when planning or interpreting an utterance. A correct representation of the misaligned perspective of one's interlocutor might play a role in successful coordination. It might play the role of a 'repair' of the mutual lexicon, as it were. If this idea is correct, then one might get a better criterion by including in the domain of the alignment relation those mental symbols we use to think about the perspective of another agent with respect to an object. In what follows, I would like to put forward a version of the alignment constraint that incorporates this metarepresentational dimension. First, let me illustrate the general idea with an example. Here is a variant on PADEREWSKIS:

PADEREWSKIS*:

Anna knows that Peter believes there are "two Paderewskis". She takes this into account when introducing Paderewski in the conversation.

Anna I saw Paderewski "the politician" today. He looks very interesting.

[102] I also note that some epistemologists define epistemological notions in terms of accessibility, or "activation" as I use it. For example, Conee and Feldman 2004 defend the view according to which one's evidence consists exclusively of one's current mental states. Thus, Feldman proposes that 'the evidence someone has at a time is limited to what the person is thinking of at the time' (Feldman). I use this as an example, I am not committed to this kind of radical internalism.

Peter Oh yes, he is a great statesman.

Anna By the way, he is the same person as Paderewski "the pianist", Peter!

In this example, the intuition is strong that communication succeeds, whereas in the original PADEREWSKIS case, it failed.[103] The difference is that Anna is able to correctly represent the perspective of Peter on the intended referent, and is able to use this correct metarepresentation in her referential plan. (The idea that a speaker makes assumptions on the perspective of her interlocutor when using singular terms was already present with the notion of a Givenness feature.)[104]

This distinction between 'normal' or 'regular' mental symbols, and metareprentational ones already exists in the literature on mental files. Here I will draw on the distinction Recanati 2012 makes between regular and *indexed* files. Recanati introduces the distinction as follows:[105]

> An indexed file is a file that stands, in the subject's mind, for another subject's file about an object. An indexed file consists of a file and an index, where the index refers to the other subject whose own file the indexed file stands for or simulates. Thus an indexed file $< f, S_2 >$ in S_1's mind stands for the file f which S_2 putatively uses in thinking about some entity. So there are two types of file in S_1's mind: regular files which S_1 uses to think about objects in his or her environment, and indexed files which she or he uses vicariously to represent how other subjects (e.g. S_2) think about objects in their environment. (Recanati 2012: 183)

One may be tempted to extrapolate a criterion that is sensitive to the accurate representation of the perspective of a misaligned conversational partner, as follows. Communication may be rendered successful in a misaligned configuration, if the enlighted speech participant deploys mental symbols indexed to the perspective of the identity-confused interlocutor in such a way that alignment is restored through the indexed mental symbols.[106] One idea is to enrich the domain of the alignment relation with the indexed mental symbols, and simply reuse the alignment relation template. The new constraint (let's call it *pragmalignment**) would be the

[103] This might be so even if we assume that Peter does not endorse Anna's correction, thus staying misaliged with her.

[104] Wether such implicit assumptions (about the *cognitive statuses* the referent has in the mind of one's interlocutor) in fact involves indexed files, is an interesting question. Recanati (forthcoming) suggests that

> In general, the mode of presentation of objects that characterizes interpersonal communication about these objects involves a deferential/anaphoric component: we presuppose that the object we are talking about is also the object the other person is talking about. (Recanati, forth: 23)

See also Recanati 2016: 159 on this theme. I am following Recanati in suggesting that we sometimes assume that our interlocutor has a diffrerent 'take' on the intended referent, e.g. fails to know relevant identities about the object under discussion, *by deploying indexed files*.

[105] See Recanati 2012, chapter 14 & 15.

[106] Very plausibly, indexed files can do some interesting work with respect to alignment in the context of attitude reports as well. See Recanati *op. cit.* where indexed files are used in attitude reports theorizing. I deal with attitude reporting in the next chapter, and provide a version of alignment using the indexed files.

constraint of pragmalignment whose original domain of departure is extended to the set of activated regular *and indexed* mental symbols in discourse context.

One might worry that the envisioned criterion will under-generate, because it seems that a speaker (i) who is not in the relevant Frege case and (ii) knows that her interlocutor is, will typically have her *regular* file activated in the context, *in addition to the indexed file* she uses to represent the misaligned perspective of her interlocutor. But of course the regular symbol of the enlighted speaker *is* related by mutual communicative dispositions to the ones of the misaligned interlocutor. Assuming that the indexed file will likewise be related in the same way, and that we have to count the regular file and the indexed file as two different files, then pragmalignment* defined in this way will never obtain.

This worry arises only if the indexed file is taken to be a file in its own right about the object with respect to which the indexed file user intends to think the perspective of the mentalized subject. This is not how Recanati 2012 thinks of indexed files. Recanati suggests that we may think of an indexed file as a sub-file of the regular file of the thinker *for the mentalized subject*:

> Since, in order to think about S_2 [the agent whose perspective is represented] and his thoughts, the subject must have a mental file about S_2, we may think of indexed files as sub-files (files within files): the indexed file $< f, S_2 >$ will be a file embedded within S_1's file about S_2 and specifically representing S_2's way of thinking about some entity. (Recanati 2012: 183)

An indexed file to S_2 with respect to object o is not about o, it is about S_2. So it should be understood as a proper part of the file the subject has for S_2. This addresses the problem I raised with respect to the one-to-one mapping condition, because an indexed file won't be related by a mutual communicative disposition to the file the subject to whom the file is indexed uses to think about the object under discussion. So the definition of pragmalignment* above is not vacuously uninstantiated, as the worry had it.

However, the definition given above does not capture the idea that metarepresenting the misaligned perspective of one's interlocutor is a way to successfully coordinate in a misaligned coordination. It does not put any constraint on the use of indexed mental symbols. So we cannot model the intended definition after the alignment relation template. Pragmalignment* really is another constraint than the alignment relation, rather than a variant on it. We want a criterion that puts constraint on the deployment of indexed files in a misaligned context. Accordingly, here is another attempt:

9.3.1 Pragmalignment* (alignment through indexed mental symbols)

Pragmalignment*
In a context $\mathcal{C}$ of misaligned configuration where one conversational partner is identity-confused with respect to an object o (call the agent in this role, X) and the other not (call the agent in this role, Y), the enlighted speech participant Y may *pragmalign** with the identity-confused agent X if:

> **(a)** Y indexes to X as many symbols as required to restore alignment with X [that is, Y indexes n mental symbols to X ($n \geq 2$) iff X has n mental symbols to think about the object o];
>
> **(b)** For every i ($0 < i \leq n$), whenever X expresses the mental symbol s_i with a given signal in $\mathcal{C}$, then Y deploys $< s_i, X >$ namely, the mental symbol that represents X's symbol s_i in Y's mind in $\mathcal{C}$.[107]

This criterion of Pragmalignment* captures the idea that metarepresenting the misaligned perspective of one's interlocutor is a way to restore alignment, and successfully coordinate.[108] Condition **(a)** is the alignment condition: it requires that the enlighted subject make enough distinctions to adequately capture the perspective of the mentalized subject. For example, in PADEREWSKIS*, if Anna had had only one indexed symbol about Peter with respect to Paderewski, she would have failed to meet condition (a). Condition **(b)** is about the proper use of indexed mental symbols in context. It is the 'pragmalignment' condition, as it were. Imagine Peter tells me 'Paderewski is great!' in a context in which it is not clear which attribute of Paderewski is salient. If I have two symbols indexed to Peter with respect to Paderewski (thus satisfying condition (a)), but deploy my 'musician' indexed symbol when Peter expresses his 'politician' symbol, then I fail to meet condition (b).

The definition is of course a rough sketch; I leave it to further research to explore the explanatory potentials of this criterion, and draw the relevant connections to the maturing mental file theory and empirical research. Note that this criterion is not intended as a necessary condition for communicative success in a misaligned configuration. There might be cases where the enlighted speaker does not represent the perspective of her misaligned partner, and communication succeeds (the context being favorable). PIANIST 2 was an example of this sort. But there may be conversational contexts in which *failing* to correctly represent the perpective of one's interlocutor makes communication fail.[109]

Going back to PADEREWSKIS*, the structure of the mutual lexicon, when taking into account Anna's metarepresentational take on Peter's misaligned perspective, might be represented as

[107] I am relying on the notation proposed in Recanati 2012: 183.
[108] I examine a similar criterion with respect to attitude reports in the next chapter.
[109] See for example KASPAROV(c) in chapter 2. KASPAROV(c) is a case where there is alignment in the standard sense, but where there is misalignment through indexed mental symbols. Again, enriching the domain of the relation with metarepresentational symbols seems better able to make the relevant prediction in such cases.

follows (Figure 3.23):

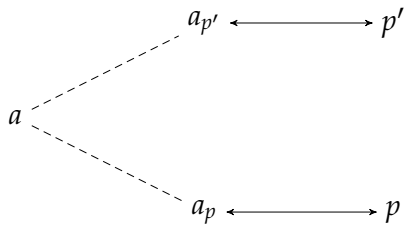


Figure 3.23 – Alignment restored through indexed symbols, where $a_p =< p, Peter >$ and $a_{p'} =< p', Peter >$

As the diagram shows, Anna represents Peter's mental symbols p and p' *as distinct*. In the terminology of Recanati 2013, a_p and $a_{p'}$ are *internally unlinked*, where internal unlinking is a means to represent the fact that the subject whose perspective is represented, is not sensitive to the coreference of her symbols.[110] As the diagram suggests, Anna's correct representation of Peter's subjective context reconstitutes a form of alignment across the relevant segment of their two lexicons. Anna's deployment of indexed symbols serves as *restoring* alignment between her and Peter. We may extrapolate the following algorithm as a guideline for enlighted misaligned speakers:

> **Alignment as an effort to match the internal policy of an identity-confused agent:** Whenever you become aware that your interlocutor is identity-confused (i.e. misaligned with you), split your indexed file into two so as to restore alignment. (Conversely, you should merge your indexed files if you become aware that your interlocutor is not identity-confused the way you thought she was).

The proposed algorithm seems sensible. Observe that I am assuming speakers deploy a default indexed file to their conversational partners in communication.[111] This hypothesis seems to make communication overly metarepresentational. For example, according to the Givenness Hierarchy Theory mentioned above, the use of pronouns (inter alia) involves implicit assumptions about the cognitive statuses of the intended referent in the minds of the audience. But we do not want to postulate metarepresentational symbols for every mutually salient elements in the common ground.

[110] It is not clear why people would have internally linked mental symbols if the merge model is true. (Recall that on the merge model, someone who knows the relevant identity will express the same thought with 'Superman is strong' and 'Clark Kent is strong').

[111] I have already alluded to this hypothesis when I mentioned Recanati's observation that

> In general, the mode of presentation of objects that characterizes interpersonal communication about these objects involves a deferential/anaphoric component: we presuppose that the object we are talking about is also the object the other person is talking about. (Recanati, forth: 23)

One pending question is: how do these 'deferential components' of composite MOPs deployed in communication relate to the indexed files?

More generally, the role of indexed files in communication raises its share of pending difficult questions in connection to the putative architecture that relates indexed files to regular ones. For example: In which cases do agents automatically activate a default indexed file to their conversational partners in communication? In which cases do we have a composite file including a 'deferential component' as opposed to a full-fledge indexed file? What is the role of indexed files in coordination and in establishing the common ground in aligned configurations? However, these issues suggest promising directions for further research.[112]

As a last remark, the aforementioned algorithm suggests a distinction between two notions of alignment. When alignment supervenes on the structure of the mental lexicons of a given set of agents and their communicative dispositions, we may say that alignment is a *state*. Two agents may be aligned in this sense without interacting at all. By contrast, Pragmalignment* suggests a sense of alignment according to which it denotes an *activity*. Pragmalignment* is the outcome of the deployment (in context) of indexed files. It is the end-state of an *effort* to match in context the internal policy (as assumed by the mentalizer) of her identity-confused interlocutor.

10 Taking stock

The question whether alignment is required for samethinking decides whether (**SHAR**) holds. In this chapter, I have argued that misaligned agents *can* successfully communicate in some but not all contexts. My argument was the following. If the standards for communicative success are context-sensitive, then alignment is not a necessary condition on successful communication. But the standards for communicative success *are* context-sensitive. Assuming that knowing what is said involves being able to rule out all relevant alternatives, *which* alternatives are relevant depend on the conversational context. I have suggested two different specific conceptions of this context-sensitivity. Let me recap the pragmatist twist to the alignment constraint I have advocated in this chapter. Very roughly, the modifications on the alignment constraint I put forward may be summarized with the conjonction of the following two innovations:

> **(i)** The stringency of the standards related to the safety requirement for knowledge of what is said is not uniform, but depends on the conversational context.
>
> **(iia)** The constraint of alignment applies to *activated*, rather than merely possessed, mental symbols.
>
> **(iib)** A speaker can successfully coordinate with an identity-confused misaligned agent by *indexing* files to that agent.

In the penultimate section, I have focused on principle (ii), which makes two modifications to the alignment constaint. Observe that one can buy the modifications separately: they are

[112] A more complete outline must deal with the role alignment via indexed files can serve in an analysis of *de dicto* attitude reports, a topic I address in the next chapter (restricting myself to issues of samethinking).

compatible but should be distinguished. The modification (iia) says that the domain of the alignment relation is not the set of mental symbols *possessed*, but the (more restricted) set of mental symbols *activated* relative to a discourse context. I call the resulting criterion, *pragmalignment*. The second modification (iib) says that indexed files are a way to restore alignment between misaligned agents, so we should enrich the domain of the alignment relation accordingly. I proposed a criterion incorporating this idea, and I called it *pragmalignment**.

Principle (ii) singles out two dimensions relevant for communicative success between misaligned conversational partners. On dimension is whether the enlighted speech participant in a misaligned pair correctly represents the structure of the relevant segment of the lexicon of her identity-confused misaligned conversational partner (through mental symbols indexed to the interlocutor, as I have suggested).[113] Another dimension turns on the cognitive statuses of the "surplus" mental symbols (relative to the alignment constraint): i.e. whether there is, in the discourse context, alignment of the *activated* symbols in the minds of the speech participants.

Principle (i) was defended by judgments about cases, and general principles about the context-sensitivity of the standards for knowledge. The principle can be fleshed out in several directions. One option draws inspiration from views labelled *pragmatic encroachment* in epistemology, and corresponds (roughly) to the idea that the degree of *practical importance* of the purposes of a conversation for the speech participants determines how *hard* it is to know what is said in the context. The other option (known as *epistemic contextualism*) is to say that what it means to "know" what is said again depends on the conversational context. (Both views can merge: it is open for the contextualist to say that the relevant contextual features are the practical stakes governing the speech participants' context). Of course, much more needs to be said in order to get a full-fledge contextualist theory of knowledge of what is said in referential communication. My goal was to motivate the idea that there exists an important notion of 'successful communication' that is sensitive to such pragmatic factors. I leave this conception as a coherent and interesting assumption, worth exploring in further work.[114]

Pragmalignment construed as the conjonction of principles (i) and (ii) might look like a disparate package of principles. But it is not. In particular, principle (ii) implies of form of contextualism about knowledge of what is said, which is what principle (i) expresses. I conclude this chapter by pointing out important and obvious problems for the pragmalignment criterion.

[113] As already mentioned, I am following Cumming in assuming the merging model of mental symbols. Postulating *indexed* mental symbols is a way to explain how one could represent the misaligned perspective of one's interlocutor without maintining superfluous *regular* files.

[114] For a recent and comprehensive survey on pragmatic encroachment in epistemology, see Kim & McGrath 2019.

10.1 Some obvious defects of pragmalignment

Pragmalignment is a synchronic, context-bound notion. Whether two symbols are pragmaligned at time t depends on the activation of mental symbols at t in a particular context. But this is a serious limitation of the proposal. We need a broader picture which does not have this defect in order to account for representation sharing across different epochs of time. Likewise, *pragmalignment* requires that mental symbols be related by communicative dispositions in both directions. This is a problem if one wants to explain how e.g. I may report beliefs Aristotle held. (Note that *alignment* has the same defects). I consider these issues in the next chapter.

4 Pragmalignment in action: Attitude and speech reports

Abstract

In the previous chapter, I have advocated a replacement of the alignment constraint with pragmalignment, having argued that misaligned agents can successfully coordinate in a discourse context (when they have their activated symbols aligned). In fact, representing the misaligned perspective of one's interlocutor is also a way to successfully coordinate in a misaligned configuration. Accordingly, I proposed to include *indexed* mental symbols in the domain of the pragmalignment relation. I ended the chapter by noticing obvious flaws in saying that the pragmalignment relation is samethinking *simpliciter*. For one thing, pragmalignment is a synchronic notion.

This chapter addresses this issue. Capitalizing on the concepts just introduced, my goal is to account for samethinking as it occurs outside of communication, namely, in attitude and speech reports, and in agreement and disagreement in non-interacting pairs of agents. To progress on these issues, I examine networks of mental files associated with the use of names in causal-historical chains. Specifically, I examine Perry's description of them (Perry 2012). Perry defines a same-saying relation without alignment constraint in terms of file-networks. His solution involves a further partitioning of the network, tracking which file of an agent is involved in a particular discourse or thought context, and how that file is used or updated in that context. I propose to understand this file-network structure, which Perry calls *thread*, as what underlies the relation of pragmalignment introduced in the previous chapter (with some rearrangement).

1 Introduction

In the previous chapter, I have advocated a replacement of the alignment constraint for *pragmalignment*, having argued that misaligned agents can successfully coordinate in a discourse context, when they have their activated symbols aligned. In fact, representing the misaligned perspective of one's interlocutor is also a way to successfully coordinate in a misaligned configuration. Accordingly, I proposed to include *indexed* mental symbols in the domain of the pragmalignment relation. But there is some vagueness and straightforward problems attached to this suggestion.

For one thing, pragmalignment is a synchronic notion. Whether two symbols are pragmaligned

at time t depends on the activation of mental symbols at t in a particular context. This is a serious limitation of the proposal.[1] There is no doubt that representations are transmitted between the past and the present. Diachronic samethinking is just part of our explanandum. For example, there is no doubt that I can have thoughts in agreement with what Aristotle thought long ago. As Frege (whose passage I have already quoted in the introduction) says:

> [A thought] may well be common property of many and is therefore not a part or mode of the single person's mind: for it cannot well be denied that mankind possesses a common treasure of thoughts which is transmitted from generation to generation. (Frege 1892: 188 in Martinich (ed) 1996)

Since Aristotle does not have any activated mental symbols any more, pragmalignment as defined previously does not account for diachronic samethinking.

Another, related defect of pragmalignment is that it requires mental symbols to be related in both directions via communicative dispositions. But one wants to explain how e.g. I can make a true report about beliefs Aristotle held long ago, even though Aristotle has no communicative dispositions now.

This chapter addresses these issues. I appeal to intersubjective representation networks that are both similar to, but importantly different from, the *Cummingian* networks of competence-level communicative *dispositions* we saw in the previous chapter. The networks I consider here are networks of mental files associated with the *use* of names in causal-historical chains. These are communication networks, with links along which information flows that maintain (in favorable cases) the reference or subject matter.[2] Specifically, I will use Perry's description of them. Perry calls them *intersubjective file networks* (Perry 2012). We can see them as another, more descriptive viewpoint on the sea of linguistic and mental intersubjective coordination.

The two sorts of networks are related. Two agents whose symbols are related by communicative path in the Cummingian sense, also share a causal-historical network. A mental symbol which is associated with the use of a public name N, is thereby related by communicative path to all linguistic users that have the public name N in their strategy of expression and interpretation (I mean the 'common currency' name, not the generic name).[3] [4] The converse is not true. Two agents who share a causal-historical network may or may not have their symbols connected by

[1] *Alignment* has this defect too.

[2] They are in this respect similar to Onofri's chains of linking relations also examined in the previous chapter.

[3] The terminology comes from Kaplan 1990. A generic name, for present purposes, may be thought of as the class of names whith the same shape. Thus, a generic name does not refer, it has bearers. By contrast, a 'common currency' name refers to a particular individual, and it has only one bearer (the referent). As Kaplan says:

> There is the generic name "David", and then there is my [common currency] name "David", there is David Lewis' [common currency] name "David", and so on. (Kaplan 1990: 111)

[4] One trivial difference is that an interaction-free pair of agents can be *related* in the networks of communicative dispositions, but not in causal-historical networks (at most they would be connected).

communicative path in the Cummingian sense. For instance, you may share a causal-historical network of deference and use for the word "meat" with a 15th-century agent. Yet your joint strategy of expression and interpretation with respect to this word does not constitute an equilibrium, due to the reference shift.[5] Actually, since the 15th-century agent does not have a communicative policy now, he is not part of the network of communicative dispositions in the first place. Things are different if the 15th-century agent produced written signals that still exist today. In that case, we may deploy equilibrium-yielding strategies of interpretation of the past mental symbols that were expressed with written signals.

1.1 Chapter plan

The plan for the chapter is as follows. In the second section, I present a brief sketch of Perry's reflexive-referential theory of content in terms of the networks supporting naming conventions. If the reader is already familiar with Perry's intersubjective file-networks, skipping directly to section 3 will induce no sense of discontinuity. In section 3, I present how the file-networks described by Perry can help us to explain the sensitivity of speech and attitude reports to the status of particular mental symbols. The goal of this presentation is to get hold of Perry's account of *samesaying* in terms of a thread, which I will understand (roughly) as the file-network structure underlying the pragmalignment relation I have introduced in the previous chapter. In section 4, I explore how the account of samesaying proposed earlier might serve in an analysis of interaction-free agreement and disagreement.

2 Intersubjective file-networks

On Perry's proposal (therein inspired by Kripke 1980 and Donnellan 1974), the uses of names are supported by causal networks. Two particularities of Perry's networks is that (a) they consist of utterances involving singular terms of all sorts, and not just of those involving names and (b) they have a cognitive/informational layer, corresponding to how the naming conventions are actually *exploited* by speakers in order to exchange information on shared subject matters.[6] Let me describe both kinds of networks in more details.

2.1 Networks of conditionally-coreferring utterances

Coco-networks connect linguistic acts that are about the same thing. They are networks of *utterances*. Utterances are concrete objects, namely linguistic representations we produce by speaking or writing; they are perceived by others; they have causes and various direct and indirect effects. What ties any two adjacent utterances in the coco-network is the relation of conditional coreference, defined as follows (Perry 2012: 172):

[5] The word "meat" used in the fifteenth century meant anything edible; but as we use it now it refers to edible flesh (Sainsbury & Tye 2012).

[6] Evans 1982 is an early theorist of information flow through networked mental 'dossiers'.

> **(Coco-reference)** For any two utterances u and u' such that u' is later than u, u' conditionally corefers (*coco-refers*) with u just in case, if u refers to some unique object x, then u' will refer to x. [7]

Observe that coco-reference is consistent with u failing to refer. Hence two utterances of an empty name can be said to coco-refer, for example. Moreover, note that this definition is an idealization. In real life, a speaker will *intend* to coco-refer with an earlier use. When things go well, this intending results in objective coco-reference. But intending to coco-refer with an earlier use is compatible with failing to coco-refer with that use. The *subjective* notion of coco-reference is the one Perry uses to defines 'coco-reference' in his glossary:

> A case of purported reference — that is a use of a name, indexical, demonstrative or demonstrative phrase in the normal way — coco-refers with an earlier use if the speaker intends to co-refer with the earlier one *if* it refers. (Perry 2012: 295)

This latter definition strikes me as inadequate as a definition of (objective) conditional-coreference: as just mentioned, a speaker may be confused and fail to corefer despite intending to do so. But it is a good definition of *deference*, which may be thought of as the link in the file-networks, as we shall see.[8]

Perry thinks of the relation of coco-reference as reflexive but non-symmetric and non-transitive. The reason is that, in his framework, for an utterance to coco-refer with another one, the former has to immediately follow the latter in the chain. In other words, coco-referring utterances are 'proximal': we may thus represent two coco-referring utterances as two adjacent nodes in a directed graph. Nevertheless, every pair of the set of utterances in a coco-network *are* comparable thanks to the relations of *coco-ancestry* and of *coco-descendancy*, defined as follows:

> **(Coco-descendancy)** For any two utterances u and u' such that u' is later than u, u' is a *coco-descendant* of u if there is a chain of coco-referring utterances from u to u'.
>
> **(Coco-ancestry)** For any two utterances u and u' such that u' is later than u, u is a *coco-ancestor* of u' if there is a chain of coco-referring utterances from u to u'.

[7] Slightly more formally, the relation of conditional coreference between the utterances u and u' may be defined as:

$$\forall x(\text{Ref(u, } x) \equiv \text{Ref (u', } x))$$

We use a material equivalence and not a material conditional, if we think that u' cannot be said to coco-refer with u in a situation in which u' refers but not u.

[8] We need deference, that is, intention to coco-refer as opposed to objective coco-reference, in order to explain cases of reference and semantic change. The 'subjective' notion supports at most what Recanati 2016 calls *weak coreference de jure*, because the presupposition of coreference may be false. On Recanati's proposed analysis, when an agent S utters u' intending to coco-refers with an earlier use u, S knows the following proposition: u and u' corefer if <u>both</u> u and u' refer (Recanati 2016: 26). Slightly more formally, what such an agent knows in such cases is that:

$$\forall x \forall y \, ((\text{Ref}(u, x) \wedge \text{Ref } (u', y)) \rightarrow x = y)$$

This analysis of the state a speaker is in when she intends to coco-refer with an earlier use assumes that all cases of confusion are cases where one of the utterances fails to refer, a non-trivial assumption. Perry seems to make this assumption as well, as we will see in due course (section 2.4).

The relation of *being a coco-descendant of* (resp. *ancestor*) is reflexive, transitive and of course non-symmetric (that's why we distinguish between being an ancestor and being a descendant). Here is an illustration (Figure 4.1):

$$u \longleftarrow u' \longleftarrow u'' \longleftarrow u''' \longleftarrow \cdots$$

Figure 4.1 – u' coco-refers with u; u''' is a coco-descendant of u; u' is a coco-ancestor or u'''

Note that the given definitions of 'coco-descendant/ancestor' feature sufficient, but not necessary conditions. The reason is that it's *sufficient* there to be *one* chain of coco-referring utterances in between the *relata*, but it's not necessary that there be exactly one: there may be *more than one*. That is, both relations are many-one and one-many, as depicted below:

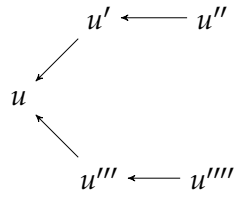


Figure 4.2 – *being a coco-descendant* is many-one / *being a coco-ancestor* is one-many

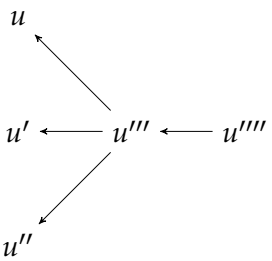


Figure 4.3 – *being a coco-descendant* is one-many / *being a coco-ancestor* is many-one

The relation of *coco-connectedness* is roughly the *union* of the relation of *being coco-descendant* and the relation of *being a coco-ancestor*. Slightly more precisely,

> **(Coco-connectedness)** Any two utterances u and u' are coco-connected just in case either u is a coco-ancestor of u', or u is a coco-descendant of u', or there is an utterance u'' such that both u and u' are either coco-descendant or coco-ancestor of u''.

A coco-network thus assembles linguistic acts of reference by the relation of coco-reference. Each linguistic act of reference on a coco-network can be reached from any other by a series of steps going from a given utterance to a later one that coco-refers with it or an earlier one to which it coco-refers. In doing so, a coco-network *supports* a naming convention, and provides the link between the use of a name and the object to which the use refers, which is the *origin*

of the network. Let me explain this. A coco-network starts with a reference to a real object (or a kind, perhaps).[9] The initial reference to the object is special, as it *assigns* a name to the object and thus *establishes a naming convention*. By the relation of coco-reference, all subsequent utterances will thus refer to the same real object, if there is one.[10] Therefore, once we have an *initial utterance* (a baptism) referring to the origin, by the relation of conditional coreference, we get a network going, as the chains of coco-reference spread out in time and space. Here is Perry's gloss on the notion of a naming convention:

> A convention for name *N* is *supported* by a network, if there is a practice along the network to use *N* to coco-refer. A use of a name that exploits a convention refers to the origin of the network that supports the convention, if it has one; otherwise the convention and the use are empty. (Perry 2012: 179; my emphasis)

> When a person or thing is assigned a name, a *permissive convention* is established: that name may be used to designate that person. When David Israel's parents named him 'David', they established a convention that made it possible for people to designate their son with the name 'David'. However, it did not preclude people from using 'David' to designate other people, or using other means of designating David—that's what I mean to emphasize by calling it "permissive." (Perry 2012: 117; my emphasis)

On the definition of *coco-connectedness* I gave, it is an equivalence relation. However, it should be noted that this is an idealization. To understand why coco-connectedness in fact fails, strictly speaking, to be an equivalence relation, we need to consider *the mental states* of the producers and the interpreters of utterances of the coco-networks. To anticipate: if a language user has a confused mental symbol i.e. for what are in fact two different things, then the utterances she produces motivated by such a mental symbol will belong to two different networks, leading to the mess. In order to be able to theorize about messes of this kind, we will need (following Perry) to supplement the coco-networks with 'cognitive nodes' of various sorts.

Before I do this, I will introduce the notion of *network content*, namely, the level of content that we get when we consider the role of the networks in the truth-conditions of utterances involving

[9] A coco-network may start without reference to a real object, but e.g. with a simulated reference to an object, i.e. an act of invention. After Donnellan, Perry calls this kind of origin a **block**. Here is a relevant quote from Donnellan:

> When the historical explanation of the use of a name ends (. . .) with events that preclude any referent being identified, I will call it a 'block' in the history. (Donnellan 1974: 23)

For an application of Perry's file-networks to the analysis of fiction and fictional reference, see Friend 2011, 2014.

[10] To simplify the discussion, I leave out cases of *reference change*, examplified by the history of use of words such as "Madagascar", "meat" or "fish". Theorizing about reference change is an interesting task, but it is a slightly different project from the one undertaken here, which must be left for another occasion. At any rate, we need a richer model, including attitude states, in order to understand reference change, because confusion is a mental phenomenon, with repercussion at the linguistic level. Following Perry 2012 (chap.8-10), I will add cognitive structure to the linguistic networks as we go.

names. But first I need to say a word on the general system of Perry's reflexive-referential theory of content, and its semantic pluralism.

2.2 A reflexive-referential theory of content

Perry is a *pluralist* about content. In his view, the content of a single utterance consists in a complex system of multiple related propositions, not in just one proposition. The rationale for content pluralism is that it's useful to have different levels of content, given the variety of our interests and purposes with the notion of content. For instance, sometimes we are interested in explaining cognitive significance and speakers' behavior; but sometimes we are more interested in specifying referential content. Within the array of propositions that may be attached to the utterance of a well-formed sentence, Perry distinguishes between *reflexive* and *incremental* content.

Reflexive content provides conditions on the utterance itself. A relation is reflexive just in case it relates any object of its domain to itself. Now, a remarkable class of linguistic expression is such that their meaning features a reflexive element. These are the *token-reflexive* expressions. To determine the reference of a token-reflexive expression (that is, what a speaker is referring to by uttering it), the audience must appeal to context. Without varying their meaning, token-reflexive expressions can change their content from context to context. Paradigmatic token-reflexive expressions are personal pronouns ('I,' 'you,' 'she'...), demonstratives ('this,' 'that'), time and place adverbs ('here,' 'there,' 'now,' 'yesterday,' 'tomorrow'...). The characteristic feature of the meaning of such expressions is that, for any token of an expression of this class, what the token refers to depends on specific facts about the token *itself*. For example, what determines the reference of an utterance of "I love butterflies" is a fact about the utterance itself, to wit, who uttered it. One innovation of Perry is to generalize this reflexivity feature of meaning to expressions of all sort, and to make explicit the role reflexive contents play in linguistic communication.[11]

Perry is willing to talk of reflexive *contents*, because according to him, the various reflexive elements of utterances give rise to full-fledge reflexive *propositions*.[12] To illustrate, consider an utterance u of "yo te amo". Given the linguistic meanings of the words composing the sentence uttered, any competent speaker will know that u is true iff the person who uttered u loves the person s/he is addressing with u. Call this latter proposition, q: it is a true proposition about u, and a proposition u conveys about itself; q is a reflexive content of u.

Incremental content is the content we get from reflexive contents by determining facts about the utterance. The label 'incremental' evokes the fact that we may take more and more things as given (in an 'incremental' fashion) to specify the truth-condition of the utterance, from a

[11] Recanati is also an artisan of this project e.g. Recanati 1993, 2012, 2016.
[12] I rely on Bochner's 2010 useful exegesis here.

minimally determined level to a fully determined level where all the relevant facts are taken as given. That latter level is the level of fully referential content. By way of illustration, consider an utterance **u** of the following sentence:

(1) Romain Gary is a writer

Here is an array of truth-conditions we may associate with **u**, starting from the more reflexive to the more referential (equivalently, from the less specified to the more specified):

Reflexive/utterance-bound truth-conditions of u $\exists x, \exists y, \exists z$ such that x is the speaker of **u**, y is the network x exploits with 'Romain Gary', z is the origin of y and z is a writer.

Now, let's take as given the identity of the speaker of **u**. Let's assume the speaker is me (=RB); accordingly:

Speaker-bound truth-conditions of u $\exists y, \exists z$ such that y is the network RB exploits with 'Romain Gary', z is the origin of N and z is a writer.

Observe that the *speaker-bound* content of **u** still does not specify *which* network is being exploited (instead, there is an existential quantification on the domain of networks). As a result, speaker-bound content is still fairly reflexive, although it's not *fully* reflexive, since in particular the identity of the speaker is fixed.

Network-content is the truth-conditions that we get when we determine which network is being exploited by the speaker. In this instance, the network being exploited is the network that supports the naming convention consisting in using the word form "Romain Gary" as a name for Romain Gary/Emile Ajar. Accordingly, the truth-conditions we get once the network is provided are as follows:

Network-bound truth-conditions of u $\exists z$ such that z is the origin of $\mathbf{N}_{RomainGary}$, and z is a writer.

Note that by specifying the identity of the network being exploited, we do not thereby specify the *origin* of that network, i.e. the referent. Hence, network-content is an instance of incremental content that is strictly less informative than referential content, in which, by contrast, all the relevant facts determining the truth-conditions of the utterance are fixed. Assuming that the origin of $\mathbf{N}_{RomainGary}$ *is* Romain Gary/ Emile Ajar, we are now in a position to provide the fully referential content of **u**:

Referential truth-conditions of u Romain Gary is a writer.

Here you can see that referential content is fully specified because there is no variable left in the truth-conditions. Referential content is coarser-grained than network-content. In order to make this salient, consider an utterance **u'** of the following sentence:

(2) Emile Ajar is a writer

Here are the network-bound truth-conditions of **u'** (assuming, as before, that **u'** is in English, has the syntax it does, 'is a writer' means what it does in English, RB is the speaker):

Network-content of u' $\exists z$ such that z is the origin of $\mathbf{N}_{EmileAjar}$, and z is a writer.

The network-bound truth-conditions of **u** and **u'** are different. However, the referential content of **u'** is the same as the referential content of **u**, despite the fact that the naming conventions exploited in **u'** and in **u** vary. Here we have two distinct naming conventions, two networks, but only one origin, shared by the two networks.[13] What distinguishes $\mathbf{N}_{RomainGary}$ and $\mathbf{N}_{EmileAjar}$ is thus *not* their origin, because they have the same origin; it is the distributions of utterances composing the respective networks. The names "Romain Gary" and "Emile Ajar" were introduced on different occasions, and so the networks that support each naming conventions are distinct.

In the last paragraph, I have been assuming that $\mathbf{N}_{RomainGary}$ and $\mathbf{N}_{EmileAjar}$ were two distincts networks with a common origin, and from this I argued that network-content was sometimes finer-grained than referential content. But on reflection, such a construal is not obviously correct. One potential problem with my construal is the following. There is nothing which prevents a speaker from using the word form 'Emile Ajar' to corefer with an earlier subutterance involving the word form 'Romain Gary' (provided that e.g. it's common ground among the speech participants that Emile Ajar is Romain Gary), since both naming conventions serve as a name for the very same individual (i.e. the two names corefer).

If that's the case, then the network $\mathbf{N}_{RomainGary}$ will include subutterances of " Emile Ajar", or the network $\mathbf{N}_{EmileAjar}$ will include subutterances of "Romain Gary". And this would seem to imply in turn that we do not really have one network $\mathbf{N}_{EmileAjar}$ being disjoint from another network $\mathbf{N}_{RomainGary}$. Instead, we have a bigger referential network composed of referential utterances of both sorts. Accordingly, it would appear that the network-bound truth-conditions of **u** and **u'** are in fact *not* distinct, since the two networks are not distinct. More precisely, virtually any subutterance of 'Emile Ajar' will be in fact coco-connected with some subutterance of 'Romain Gary' and conversely.[14] Compare with what happens when we consider the Babylonians and the names 'Hesperus' and 'Phosphorus' as used by them: here, given the naming conventions in use in the Babylonians, we clearly have two distinct networks with a common origin, as opposed to one network, since no Babylonyan knew the names were in fact coreferring (the relevant coco-networks are of course no longer disjoint).

The problem lies in the fact that Perry's coco-networks are allowed to include referential utterances of all sorts. Because of this, Perry's coco-networks are coarser-grained than the

[13] I address the question whether the two networks are really distinct in a minute.

[14] People can be wrong about the identity of the networks they are participating in. But network-content involves networks, not networks under MOPs, on Perry's analysis.

linguistic continuants we would get by individuating the coco-network with chains of *explicit* conditional-coreference, where 'explicit' means something like: involving the *same* naming-convention, and explicitly anaphoric devices.

Be that as it may, an adequate theory of samethinking should capture the contrast between (i) the cases where "Ajar" and "Gary" are used interchangeably and constitute a single network vs. (ii) the cases where "Ajar" and "Gary" *are not* used interchangeably and thus constitute *two* distinct networks (as it is presupposed by the speech participants). The notion of a network is thus an hybrid of *subjective* discourse context and *external* causal anaphoric relations. Speakers not only participate in the networks, they represent them as they intend to extend them when they use singular terms. Call the latter type of discourse context, *hyperintensional*, and the former type, *non-hyperintensional*. Can Perry's framework represent this contrast?

Yes, it can. As we shall see, hyperintensional discourse contexts are modelled by *partitioning* the coarser-grained, non-hyperintensional networks. To do this, however, we will need a way to represent the mental states of the speech participants in discourse context (and the *cognitive status* of the relevant subutterances). I now turn to Perry's richer picture we get when we supplement the coco-networks by considering how utterances populating the coco-networks connect with the *states of mind* of their producers and interpreters.

2.3 Networks of inter-coordinated mental files

File-networks are the networks we obtain when we consider the cognitive acts and states surrounding the production and interpretation of utterrances composing the coco-networks. Roughly, just as coco-networks collect all the utterances about the same subject-matter, file-networks collect all the cognitive states and acts about the same subject matter. They involve the cognition that different people of a population have of the same object.

A variety of nodes Before I describe the file-networks, I need to briefly introduce Perry's ontology of the cognitive domain, and relate it to the terminology I have been using. Perry's cognitive ontology is a particular version of mental file theory, which is itself a particular version of *Language-of-Thought Hypothesis*. I already introduced (and committed to) the mental file theory in this thesis. In order to avoid dissipating the terminology, I will use the vocabulary already in place, which slighlty differs from Perry's. So, I will call a *mental file*, the complex entity composed of a mental symbol and the repository of predicative entries it is associated with. (Perry uses "notions" for *mental symbols*, "ideas" for *predicative entries*, and he uses "file" for what I call a *mental file* here).[15] Thus, a mental symbol is roughly a mental representation about a particular object; predicatives entries are mental representation of properties or relations. Mental symbols and predicative entries combine together under some attitudinal mode

[15] In Perry's terms, therefore, a file is the association of a notion and the ideas attached to it.

(belief, desire, hope, intentions, etc) to form the propositional attitudes. I will restrict to beliefs in what follows. In the kind of framework I am assuming, a belief is thus functionally realized by a mental file. Again, a mental file is the combination of a mental symbol together with a set of co-filed predicative entries. Co-filed predicative entries represent properties taken to be co-instantiated by the object the mental symbol is of. The co-instantiation bit is encoded by co-filing. Hence, a file is like a belief-box regarding a putative object.

In Perry's framework, mental symbols are of two types: they are either dependent on a perception of the object (in which case they are typically short-term and unstable; Perry calls such files 'buffers'), or they are not dependent on a perception of the object. When they are not, mental symbols will be typically stable — Perry calls them 'standing notions'. Perception serves to anchor names in the objects they serve to talk about. We may call this anchoring, *perceptual grounding*. In perceptual grounding, a name is anchored in the object in virtue of the causal link between a person and that object when it is the focus of *that person's perception*. Groundings do not only occur during baptisms: they occur whenever a name is used concomitantly with a perception of the referent of the name (see Devitt 1981, 2015, Recanati 2020)—as depicted in Figure 4.4:

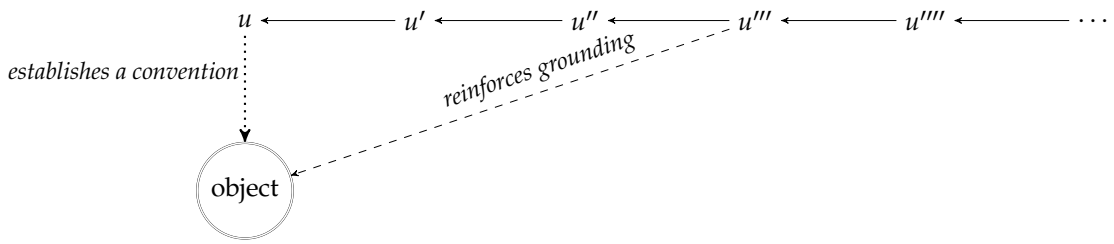


Figure 4.4 – Multiply grounded chain of reference
(Adapted from fig 5.2 in Recanati 2020)

The other kinds of nodes we need to recognize as integral parts of the file-networks have to do with utterance production and interpretation. As previously mentioned, coco-networks are a proper part of the file-networks. Hence *referential utterances* – by this I mean utterances involving singular terms – are another kind of node to be found on the networks. When singular terms are used to refer, they are what Perry calls *references*. In principle, file-network include all kinds of referential utterances, not just assertions but also e.g. questions and directives. Here I follow Perry in focussing on beliefs and assertions, in particular, on simple indicative sentences containing singular terms with the intention of communicating content (Perry calls them *statements*). In addition to *references* and *statements*, we need to include *perceptions of reference*. (Language users also perceive *predications*, i.e. the predicative parts of statements). Although Perry is not explicit about this, perceptions of reference and predications involve the mental lexicon of language users, which can also be modeled as a system of files, namely *lexical*

files (Gasparri 2015). I will think of lexical files as *pointers* to the relevant mental files; and I will think of the interface between the lexical system of agents and their mental encyclopedia as being (very roughly) realized by metalinguistic entries in the files. So for instance, when a mental file is pointer-related to a lexical entry for a word *M*, I say it includes a metalinguistic entry such as 'is named *M*' .

To recap, file-networks are comprised of the following kinds of elements (represented by as many 'nodes' in the graph models of these networks): object-perceptions; files (i.e., co-filed mental predicates combined with either buffers or standing mental symbols); references and statements; perceptions of references, of predications and of statements. I now turn to the kinds of *links* that relate such elements.

A variety of relations Much of the complexity of file-networks has to do with their dynamics, in particular the dynamics of the networked mental symbols, and of their association with predicative entries. Spelling out the whole machinery of the file-networks would required a whole dissertation in itself. In what follow, I provide only as much detail as is necessary for dealing with the issues of samethinking I am concerned with. Since the intersubjective file networks are *causal-historical*, to ask what kinds of link there are in the file-networks is to ask what is the *etiological* structure of the various nodes.

Intersubjective file-networks develop over time: they grow as further nodes are created; they may fork or pool.[16] For instance, references are made, and perceived. Mental symbols are created, linked, merged. Mental symbols may be 'deleted' from agents' minds, but even when agents 'delete' mental symbols from their minds, the past nodes do not disappear from the file-networks. File-networks are 'append-only', meaning that one can only add nodes to the network, but not erase nodes. (Likewise, one cannot erase the past by forgetting).[17] Predicative entries (i.e. *ideas*) are also exchanged along the communicative chains.

When thinkers exchange ideas, they *coordinate* their mental symbols for the time of the exchange. (Perry uses 'linking'. Linking is, roughly, one of Perry's names for the relation of intersubjective coordination). Coordination is a necessary condition for information to flow between files; we may think of coordination in terms of the relation I have presented in the previous chapters, namely as *mutual presupposition of (conditional) coreference*.[18] Although the *coordination* relation is very important, it is not general enough to be our graph-relation, because coordination occurs between mental symbols only (either buffer or standing mental symbols). Instead, we want to understand how coordination occurs; for instance, coordination occurs in communication in part on the basis of reference perception.

[16] Perry rather talks of 'branching' and 'merging', respectively.

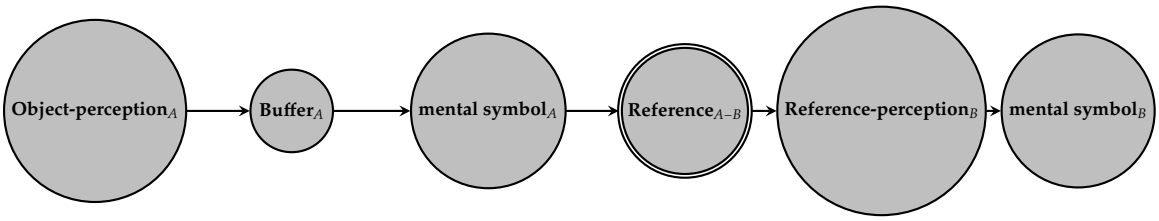
[17] For the fascinating suggestion that even human episodic memory is 'append-only' i.e. such that new data can be appended to memory, but where existing data is somehow immutable, see Cho and al. 2018.

[18] Recanati 2016 calls this relation 'weak coreference de jure'.

There are mainly two relations Perry uses to describe the etiological structure of the nodes in the file-networks: the *parent-relation*, and the *origin-relation*. The parent-relation holds between pairs of representational nodes — files, references or statements, and perceptions thereof. It is (roughly) the relation that obtains between a first representational node and a second just in case the first caused the second, and both coco-refer. More precisely, here is a typical etiological template (depicted in Figure 4.5):

> Object-perceptions *give rise* to buffers; buffers *give rise* to mental symbols; mental symbols *give rise* to references; references *give rise* to perceptions of them; reference-perceptions *give rise* to mental symbols in turn.[19]

Figure 4.5 – The *parent-of* relation



The *parent-of* relation is the relation that obtains between one node and another just in case the former gives rise to the latter. It is our graph-relation.[20] A causal chain of parent-relations is characterized by an alternation of mental representations (object-perception, reference perception, files, either stable or buffers), and utterances. The *parent-of* relation holds (roughly) between pairs of representational nodes — files, references or statements, and perceptions thereof – in the manner described in the typical etiological template described just above. As Perry puts it:

> When a [mental symbol] gives rise to and governs a reference, the [mental symbol] is the parent of the reference. (. . .) The reference is the parent of the perception of the reference. (. . .) The reference perception is the parent of the [mental symbol whose creation it triggers]. (Perry 2012: 204-205, modifications mine)

Perry seems to think of the *parent-of* relation as a non-symmetric, intransitive relation, just like coco-reference in his framework. Consequently, we may use the symmetric transitive closure of the *parent-of* relation to compare non-adjacent nodes — i.e. the relation of *having a common ancestor or descendent*. In effect, we may think of the *parent-of* relation as the inverse, and a generalization of, the *coco-reference* relation, generalized to all types of nodes and not just the

[19] Perry 2012 articulates each step in section 9.3 of the book.

[20] Below I qualify this claim and say that file-networks are better undersood as multi-graphs, involving more than one types of link.

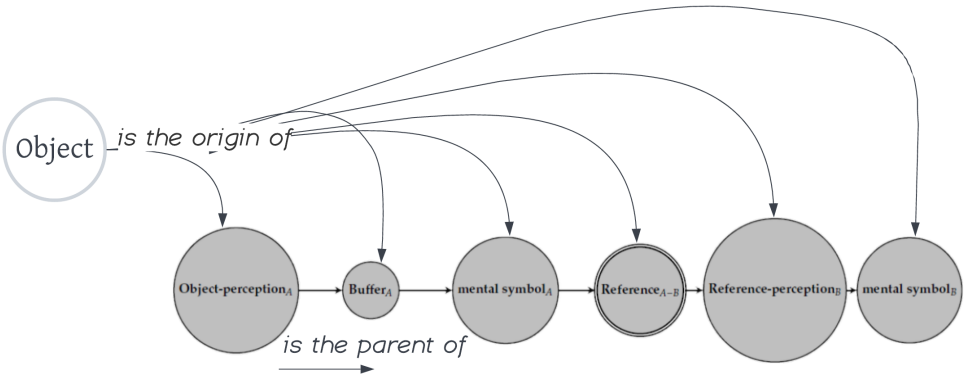
linguistic ones as in the coco-networks. (I note that there is something potentially odd in saying that e.g. a reference-perception coco-refers with, say, a (non-lexical) mental symbol. For reference-perceptions are about references, and references are utterances, hence the mental symbols involved in the reference-perceptions arguably pertain to the lexical system of the agent, not to her mental encyclopedia. My perception of the word "Napoleon" does not corefer with my mental symbol of Napoleon; it corefers with my lexical concept for 'Napoleon' assuming I have this lexical item in my lexicon. I won't explore this issue further here, because it is not crucial for getting what I want from Perry's framework.)

By contrast, the *origin-of* relation holds between an object (typically at the origin of a network) and the representational nodes. In Perry's terms:

> The object a perception is of is the *origin* of the perceptual buffer to which it gives rise. (. . .) *This property of being the origin of a node is preserved by the parent relation*; that is, when one node gives rise to a second, the origin of the first is the origin of the second. (...) The origin of the reference is the origin of the perception of the reference. (Perry 2012: 204; my italics)

Coco-reference is comprised in the *parent-relation*; that is why the origin of a representational node is preserved under the parent-relation (as depicted in Figure 4.6 below). But the parent relation is also a causal-historical link. A parent node *gives rise* to its child node; and to give rise is to cause. Therefore, the parent relation is both a semantic and a causal-historical relation. (The causal history of a node, Perry calls 'pedigree').

Figure 4.6 – The origin of a node is preserved under the *parent-of* relation



Etiology also concerns predicative entries.[21] Ideas exist, and flow, within the same file-networks than references. Perry metaphorically speaks of '*the route of the flow of ideas*'. The equivalent of

[21] I find that mental representations for properties and relations are under-studied in the direct reference tradition. Perry 2012 9.4 is a nice exception, but even Perry is much less explicit about the network-structures underlying the flow of ideas than about the ones underlying the transmission of mental symbols. I feel like this is an area where more work needs to be done. Indeed, many direct-reference theorists proceed as if there was no Frege's puzzle for properties.

the *origin-of* relation as applied to predicative entries is the relation: *is the source of information of*.[22] Not only *mental symbols* are transmitted, also *ideas* are transmitted. Paths by which ideas are transmitted need not be the same as the paths by which mental symbols are transmitted – in fact, they will often differ. Hence, the file-networks should be arguably modeled as a *multi-graph*, that is, as a graph which is permitted to have multiple links that have the same end nodes. Two nodes may be connected by more than one edge, of different types. In particular, we need to have edges for representing the *pedigree of ideas*, that is, the flow of information across mental symbols, in addition to the edges representing the *pedigree of the mental symbols* (that is, roughly, reference transmission).

Things are even more complicated, since we should distinguish the etiology of the various conventions for naming properties (their *origins*), from the networks that provide the etiology of the associations of ideas in particular files (their *pedigrees*). Since such details – although very important to understand the whole picture – are not crucial for my discussion, I won't articulate them here.[23]

Regarding reference perceptions, we might want to distinguish among the following types of links:

- Reference perceptions when they lead to the creation of a new file – call this link-type, $\mathcal{P}$;
- Reference perceptions when they lead to the addition of entries of type 'IS NAMED N' within a file (whether the file be newly created or old) – call this link-type, $\mathcal{Q}$;
- Reference perceptions when they lead to the recruitment of an existing file in which the metalinguistic information'IS NAMED N' is already present – call this link-type, $\mathcal{R}$.

Link-type $\mathcal{P}$ and link-type $\mathcal{Q}$ make up the causal-historical chains of reference *transmission* proper, while link-type $\mathcal{R}$ makes up communicative chains of use. The first two types of paths ($\mathcal{P}$ and $\mathcal{Q}$) constitute *reference borrowing* (e.g. Devitt 2015). The third sort of paths ($\mathcal{R}$) does not constitute reference borrowing; but in all sorts inter-coordination is present.[24] We can think of the backwards chains whose links are all of type $\mathcal{Q}$ as (roughly) the causal-historical chains Kripke (1980) is talking about. They collect all the links by which a name is transmitted from one speaker to another.

[22] Originally due to Evans 1973

[23] See the diagrams in Crimmins 1992 which could serve to flesh out the details of Perry's networks, in particular the intersubjective dynamics of the association of ideas with mental symbols.

[24] I find a similar suggestion in Kamp 2019 and Cumming 2013 about the distinction between causal-historical chains and communicative chains. As Kamp also remarks, studying the graph properties of file-networks (he calls them *Entity Representation Networks*) is an interesting avenue for further research. I add that the same is true of their empirical counterparts, the *cultural cognitive causal chains* (CCCCs), or social CCCs that stabilize mental representations and public productions in a population and its environment see e.g. Sperber 1996, 2000, 2001, Morin 2015, Schönpflug (ed) 2008. Nowadays, many communication researchers use blogspace and programmatic access to social media such as Twitter to scrutinize the propagation of cultural artefacts across social networks (Sloan 2016 for an introduction).

Having described the main components of a file-network, I will now present two main ways in which coreference relations on networks can go wrong, and how the file-network framework may deal with them.

2.4 Messes in the networks

Coreference relations may be misrepresented in essentially two ways: either the agent mistakenly represents coreference where there is actually no coreference (confusion), or the agent fails to represent coreference where coreference actually obtains (so-called Frege-cases). I present them in turn.[25]

Confusion cases Sometimes language-users are confused about the object(s) of their thoughts. In particular, what Perry calls 'mess' is the mental configuration whereby a thinker has *one* mental symbol for thinking about *two* different objects.

Messes are a serious impairment. As Kamp (2019) not uncontroversially remarks:[26]

> Being linked to more than one referent isn't much better than being linked to none at all; being divided over what entity you represent is a way of not representing either or any of them. (Kamp 2019: 17)

When there is a mess as defined, the subject's mental symbol belongs to *two* different referential networks. Hence the utterances governed by such a messy mental symbol is doomed to belong to two different networks in turn. But this is a problem, for that would imply that any utterance that is in the coco-descendent or coco-ancestor relation to a mess-motivated utterance coco-referred to two different objects in turn. It seems that *all* the network would then be impaired by the presence of a single mess-motivated utterance.

Instead, it is stipulated that any messy-motivated utterance triggers a *block*. Why? This is because a singular term cannot refer to two different objects, hence *no* referent can be assigned to a mess-motivated utterance: this is a case of reference failure, hence a block.[27] The notion of a 'block', Perry borrows from Donnellan. Donnellan defines it as follows:

[25] I use 'messes' to mean either confusion cases or Frege cases. But Perry rather uses 'messes' to denote confusion cases only. Nothing hinges on this.

[26] I take Kamp's project and Perry's framework to have a certain remarkable affinity (besides, Kamp is explicit about the affinity of his framework and Recanati's mental file theory). Kamp's approach is formal. Perry's is rather informal and descriptive. I think Perry's framework would benefit greatly from a formal regimentation; on the other hand, I think Kamp's framework would benefit from incorporating some of Perry's descriptions and insights, as his framework is extremely idealized as of now. More generally, Perry shares a theoretical mindset with theorists who model the discouse context by appealing to mental representations — such as in Discourse Representation Theory.

[27] For a different treatment of messes in terms of *partial* reference and 'degrees of designation', see Devitt 1981 borrowing from Field 1973. For an overview of different treatments of confusion cases, see the monograph of J.L. Camp 2002.

> When the historical explanation of the use of a name (with the intention to refer) ends (. . .) with events that preclude any referent being identified, I will call it a 'block' in the history. (1974: 23)

(The idea that *any* mess precludes any referent being identified can be disputed. In the next chapter, I argue that by incorporating an interpretationist layer in the model, one is able to ignore at least some of the messes). Accordingly, whenever an utterance is governed by a messy mental symbol, then the particular subnetwork terminates. That is, being coco-connected to an utterance whose parent mental symbol is messy (a very common situation) does not make the coco-connected utterance part of the two different networks that lead to the mess. Instead, we have a block whenever a mess occur on the chain (i.e. reference failure).

As a result, because messes trigger *blocks* in the chain whenever they occur, messes as defined prevents the reachability relation (i.e. coco-connectedness) from being an equivalence relation, because *we loose (locally) transitivity*. As Perry remarks:

> A reference cannot be the coco-descendant of two references that belong to networks that are different up until the time of the reference, unless both networks lead back to the same referent. *When messes occur, the particular subnetwork terminates*. That is, being coco-connected to a mess-motivated referential utterance, does not make an utterance part of both of the networks that lead to the mess. (Perry 2012: 298; my italics)

Frege cases The other kind of mess we may distinguish should be familiar by now. It is the *dual* of the configuration we just saw.[28] That is, when a thinker has *two unlinked mental symbols* for the *same* object. When two mental symbols are unlinked for an agent, ideas cannot flow between them, because the thinker is not disposed to trade on identity of reference. (Remember Pierre, associating IS PRETTY with his LONDRES–symbol, and IS UGLY with his unlinked LONDON–symbol, without irrationality). This is where the *pragmalignment* relation finds its home in the file-networks, and interfaces with Perry's framework.

3 Samethinking along threads

In this section, I explain what the relation of samethinking is with respect to attitude and speech reports. I introduce the notion of a thread, and I characterize samethinking in terms of pragmalignment understood as thread sharing. As I shall explain, even though we can reinstate shared content relativized to the threads along which agents are pragmaligned, sharing content along a thread does *not* amount to thought identity. This is because the agents whose mental files are in the samethinking relation need not be aligned along the network — they only need

[28] Again, I must warn the reader that Perry is using 'mess' for cases of confusion only, whereas I am using it to cover both kinds of network disruption : confusion cases and Frege cases.

to be pragmaligned.[29]

In hyperintensional discourse/mental contexts, we want to be able to analyze the sensitivity of attitude reports and indirect discourse to the cognitive status of particular mental symbols.[30] A recurring theme of the previous chapter was that, when one considers the states of mind of a thinker in a Frege case, one wants to consider the *relevant* mental symbols of the agent in particular discourse/mental contexts, to the exclusion of her unactivated, unlinked albeit co-referential other mental symbols. This section introduces more fine-grained file-network structures, which can serve as fine-graining on network-content. In particular, I present the notion of a *thread*, a further partitioning of the network that enables to track which file of an agent is active in a particular discourse or mental context, and how the file is used or changed in that context.

The goal of this section is to convince my reader that the notion of pragmalignment finds its home in Perry's framework. More specifically, I will construe pragmalignment in terms of the file-network structures presented by Perry in order to provide a *samethinking* relation which is responsive to the cognitive statuses of mental symbols, as desired. Before I turn to attitude reports and indirect discourse however, I introduce the more fine-grained file-network structures we need.

3.1 Local Networks

A *local network* of a given file-network is a part of the file-network that is involved in the production and interpretation of a particular discourse. It encompassess all and only the mental symbols (together with the references, and reference perceptions) involved in a particular conversation. Take a file-network, scale it down to a particular conversation: you get a local network. In Perry's words:

> A local network is a subnetwork that is involved in a particular conversation. A local network goes through the minds of the sequence of individuals involved in the conversation, $a_1,\ldots,a_n$ along which content is passed, plus all of the nodes from which content flows to a_1's [mental symbols], and all the nodes into which information flows from a_n's [mental symbols]. (Perry 2012: 244)

I suggest that we don't have to tie the notion of a local network to *conversations*: we may usefuly extend the notion to any kind of context involving files and the coordination of files of different

[29] As I will remark, in attitude reporting, the relation of communicative path need not obtain in both directions: one direction is enough. I use the pragmalignment relation of the previous chapter rather flexibly.

[30] See Chastain 1975 for the conception of an agent's mental state as a type of context. For examples, he says:

> My visual field is a context consisting of elements commonly called "visual sensations." In general anything which has content is a context, as I use the term. Anything that has meaning or sense is a context. Anything which expresses something or represents something is a context. (Chastain 1975: 195)

agents (Geach 1967, Chastain 1975, Burge 1983). For example, consider the following scenario, provided by Sandgren 2017, which is a variant on a case originally presented by Geach 1967:

> Hob and Nob live in the same village and read the same newspaper. The newspaper reports that a witch has been terrorizing the village. Hob believes that she has blighted Bob's mare and Nob believes that she killed Cob's sow. (Sansgren 2017: 2)

We may construe this scenario as involving a local network, that comprises the witch-related cognitions of Bob and Nob, and in which the newspaper is a link between these cognitions. Likewise, I suggest we can be liberal as to the scope of a local network. Why couldn't we allow that a local network consists in a *distributed* context of utterance, when it is useful to to so? Consider this scenario, provided by Sainsbury 2002:

> Suppose that one group of speakers [living in a village] sees a mountain from the north side and calls it "Everest" and another group sees it from the south side and calls it by a name which, coincidentally, sounds and is spelled the same. At first, the groups do not meet. But then the route through the range is discovered. North-siders and south-siders talk to each other freely using "Everest". (Sainsbury 2002: 215)

We may see the scenario as involving a local network consisting in the fusion of two local networks, one starting from the north side of the mountain, and another starting from the south side of the mountain.

Perry's description of a local network suggest that we may represent them in terms of an induced subgraph of the main directed graph that represents the file-network whose local network is a subnetwork.[31] In addition, Perry's use of expressions such as '*goes through* the minds of the *sequence* of individuals', or 'the nodes *into which* information *flows*' (my italics), suggest that we may represent a local network in terms of a *path* in the relevant subgraph.[32]

[31] In particular:

> Let $G = (N, L)$ be the directed graph representing a file-network, and let $S \subset N$ be a subset of nodes of G such that they are exactly the nodes involved in a given communicative episode (or, following my suggestion above, in some other type of context involving files, or in a sequence of such contexts). That is, S will include the nodes representing "the minds of the sequence of individuals involved in the conversation, $a_1, \ldots, a_n$ along which content is passed, plus all of the nodes from which content flows to a_1's [mental symbols], and all the nodes into which information flows from a_n's [mental symbols]". Then the graph whose node set is S and whose link set consists of all of the links in L that have both endpoints in S, represents the local network associated with the conversation. Formally, it is the induced subgraph $G[S]$.

See Kamp 2019, 2022 for an interesting formalization of the causal-historical chains in terms of directed graphs.

[32] In graph theory, a path in a graph is a (here finite) sequence of links which joins a *sequence* of nodes (here, the minds of the sequence of individuals involved in the conversation, $a_1, \ldots, a_n$ along which content is passed). As I have already pointed out in the previous chapter, a sequence is an enumerated collection of objects in which repetitions are allowed and order matters.

3.2 Threads

A *thread* is a partition of a local file-network, determined by the cognitive status of a given mental symbol of an agent involved in a particular conversation. A thread thus tracks which file of an agent operates in a particular discourse or mental context, and how the file is used or updated in that context. Take a local network, select a mental symbol whose status you want to inquire about with respect to a communicative episode, select all the nodes of that local network involved in the causal transfer of information from and to that mental symbol: you get a thread. In Perry's words:

> A thread is a part of local network defined by a particular [mental symbol] n of a particular person. A thread only includes nodes from which ideas flows to n, and to which ideas flows from n. (Perry 2012: 244)

Just as I have proposed to liberalize the notion of local network to include non-communicative contexts (such as mental contexts), and sequences of contexts, one may also want to extend the notion of a thread to mental (e.g. perceptual, recognitional, imaginative, mnesic) contexts, which need not involve communication. Likewise, I suggest that we may usefully extend the notion of a thread to cover distributed contexts. A thread that spans across contexts might enable us to do interesting generalization about the nodes of different agents included in the context-spanning thread.[33] On this more extensive notion of a thread,

> any context (synchronic or distributed) where an information flow determines a directed path in a local network, may be considered as a thread.

For example, using this broader notion, we may now see the Babylonians as participating in two threads embedded in the big file-network for Venus. One thread is determined by the distributed Babylonian mental file associated with the name *Hesperus*. Another thread is defined by the distributed Babylonian mental file associated with the name *Phosphorus*.

This broader notion of a thread (extended to distributed contexts) might enable us to do interesting generalization about the Babylonians' communal use of these two names, and of their cognition related to these uses. As I argue below, threads are essentially tools to interpret others, and represent contexts: we can use them as we think fit. They are interpretive entities we use to represent contexts and make sense of others' cognitive and linguistic behavior. Understood in this way, threads are not unlike Fregean *senses*. This analogy will be reinforced later on, when I deal with the role of threads in attitude and speech reports. Note however that, because threads are ingredients of the metasemantics, they are perfectly compatible with a referentialist view of content. In fact, I will argue that they *imply* a referentialist view of content, because samethinking along a thread does not require thought identity.

[33] See Cumming (2007: 55) and Cumming (2013a: 7) where a notion of distributed context is introduced.

Perry's description suggests that we may represent a thread as a *graph partition* of the induced subgraph that represents the relevant local network. A thread determined by a mental symbol n, partitions the set of mental symbols of a local network into two disjoint subsets: the subset of all and only the nodes from which and to which ideas flow from n in the conversation, and the subset of all the other nodes.[34]

In particular, assuming conversations are individuated by their conversational topics, if the mental symbol n is *unlinked* from some other mental symbol referring to the object under discussion in the same agent, then the partition will have two pairwise disjoint elements: the thread T_n relative to n, and the set of all the nodes of the local network not belonging to T_n.[35] But if there is no other mental symbol for the object under discussion belonging to the agent such that it is unlinked from n (i.e. if the discourse context is non-hyperintensional), then n determines a *trivial* partition with only one element (i.e. the thread is identical to the local network itself).[36]

3.3 Threads & pragmalignment

As we shall see, threads serve as a *fine-graining* on network-content, but without alignment. This is reminiscent of the notion of *pragmalignment* introduced in the previous chapter. A thread in a hyper-intensional discourse context, collects all the mental symbols that are pragmaligned in the communicative episode, and leaves out the other unactive coreferential symbols of the agent. A thread targets the mental symbols that are interpersonally co-activated in discourse context with respect to a given conversational topic, and related by *actualized* communicative dispositions.

As I want to understand them for our purposes, threads are thus well-suited to model *misaligned* communication. Frege cases relative to the object under discussion are the *raison d'être* of the notion of a thread. The notion of a thread enables one to discriminate discourse contexts in which a pair of coreferential singular terms are used *interchangeably* for an agent or a pair of agents, and ones in which they are not. On this construal, threads are trivial when no agent involved in the conversation has unlinked mental symbols for the object(s) under discussion.[37]

[34] In other words, a thread relative to a mental symbol n of a local network **LN** is the equivalence class T_n of all the nodes involved in the flow of ideas from and to n, and that set is disjoint from the set of nodes that are not involved in the flow of ideas to and from n in the local network.

[35] I am using the notions of 'conversational topic' and 'object under discussion' informally in their intuitive meaning here.

[36] I will rely on such a construal of the threads for getting the transparent readings of attitude reports. I believe this is what Perry has in mind. For example, Perry says things like:

> Since Smith has two different notions of Dot, he also has two different threads with Dot as their origin, just as Ivan had two threads for San Sebastian. (Perry 2012: 259)

We can stipulate that when one selects a notion n that is not involved in the conversation at issue, then the thread is the singleton $\{n\}$.

[37] Again, if we individuate conversations by conversational topics. If a conversation can switch its subject matters, my point is of course no longer valid, for we would have different threads in any case. But my understanding of a

As a way to recap the points just made, let me put forward the following principles (Perry may or may not be committed to them):

(A) Hyperintensional discourse context A discourse/mental context is *hyperintensional* if an agent has unlinked *labelled*-mental symbols for some object under discussion.

(B) Content and threads When the discourse/mental context is hyperintensional, *same-content* tracks coreference or sameness of network-content relative to the thread along which the mental symbols of the speech participants are pragmaligned. Otherwise, *same-content* tracks same *network-content* or else coreference.

. . . Where a mental symbol is *labelled* just in case it is associated with a metalinguistic predicate 'is named N' for some proper name N.[38] A mental symbol m which contains 'is named N in its associated repository of information, represents its referent as named N. Here I construe the *is-labelled* relation as a relation between a mental symbol and a public, 'common currency' name (Kaplan 1990).[39] (We may have to construe it as a relation between a mental symbol and an idiolectal name, in order to deal with the Paderewski type of cases in which a thinker associates *two* mental symbols with just one public name.[40]). In Cummingian terms, we may say that a N – *labelled*-mental symbol is connected by communicative path to all linguistic users that share the name N.

Why principle **(A)**? The rationale is straighforward, and should come as no surprise given my discussion of the need for alignment in the previous chapter. A conversational situation in which an agent has two *unlinked* – *labelled*–mental symbols for the same referent, induces a

thread is oriented towards the analysis of samethinking without alignment.

[38] I borrow the *labelled* characterization to Kamp 2015 who speaks of "labelled" or "named Entity Representations":

> An entity representation ER whose distinguished discourse referent is x and whose [entries] contains the condition 'Named(x,N)' is called N – *labelled* and N will be referred to as *a name of* the represented entity *according to* ER (or as *the name* in case 'Named(x;N)' is the only naming condition in ER). Entity representations that are N-labelled for some N will be referred to as *named*. (Kamp 2015: 26)

See also Recanati 2012:

> In the case of proper names the mode of presentation contributed by the expression type is arguably metalinguistic. The referent of a name NN is presented as *bearing the name NN*. (...) The utterance of a name NN therefore triggers the search for a mental file containing the information 'called NN'. The referent of a file containing that information may not actually be called by that name (improper uses). (Recanati 2012: 234).

[39] This is how Recanati 2012 seems to conceive of name-based files:

> One uses a public word in thought whenever one bases a mental file on the word through some kind of 'deferential' mechanism. (...) For the thought to inherit the reference of the name, the name itself has to be disambiguated; it must be a 'common currency name' in the terminology of Kaplan (1990). This is what happens in ordinary deference. (Recanati 2012: 142)

[40] This is addressed in the next chapter.

discourse context in which there is *substitution-failure of co-intensional singular terms*. When we want to report *de dicto* the thoughts or the speech of such agents, the reports are typically sensitive to the status of particular mental symbols. Hence the characterization *hyperintensional* for such type of discourse situation.[41] I have not yet defined the notion of *same-content* relativized to a thread. Accordingly, the principle **(B)** is still only a schema. Having introduced local networks and threads, I will now say how they may be used to analyze indirect discourse, attitude reports, and more generally how they may serve in an analysis of samethinking outside of communication.

3.4 A file-network account of speech reports

Indirect speech reports are of the form '*S* said that *p*', by which a speaker aims at reporting what another speaker – which *S* stands for in the schema – has said. To spell out Perry's proposed analysis of indirect discourse in terms of the file-network structures we saw, I will reuse Perry's *Ivan*-Donostia/San Sebastian example (Perry 2012: 239-246).

Ivan has two mental symbols referring to the city of San Sebastian, also called 'Donostia' in the Basque country. He has one symbol labelled 'Donostia'. And he has another, distinct symbol labelled 'San Sebastian'. Using the way of talking I introduced earlier, we may say that his *San Sebastian* labelled symbol is pointer-related to his lexical file for the word "San Sebastian". Let S_I^{enc} be the encyclopedic San Sebastian file of Ivan ("*enc*" standing for "encyclopedic file"). And let S_I^{lex} be the corresponding mental word.

An important detail of the case is that S_I^{enc} is not pointer-related to Ivan's lexical file for the word 'Donostia' (let's call it D_I^{lex}). Besides, Ivan has another mental symbol (let's call it D_I^{enc}) referring to San Sebastian, that he expresses with the word 'Donostia' but not with the word 'San Sebastian', and that he activates to interpret tokens of 'Donostia' but not to interpret tokens of 'San Sebastian'. In short, D_I^{enc} is pointer-related to D_I^{lex}, but not to S_I^{lex}.

The context is that Ivan is at the airport and adresses a group of people wanting to go to San Sebastian. Ivan sees a bus equipped with a sign that reads 'Donostia' and he says "Not this one!". Let's assume that the audience is not confused about Donostia/San Sebastian. Asked about his reaction, Ivan says:

(3) That bus was not headed to San Sebastian. It was headed to Donostia.

Intuitively, the following are true reports about what Ivan said:

(4) Ivan said that the bus was not headed to San Sebastian.

(5) Ivan did not say that the bus was not going to Donostia.

[41] I acknowledge that there are hyperintensional contexts that have nothing to do with Frege cases, e.g. claims about what grounds what—see section 2.2 of Bliss & Trogdon 2021, and section 1.2 of Berto & Nolan 2021.

(6) Ivan said that the bus was going to Donostia.

(7) Ivan did not say that the bus was going to San Sebastian.

The problem is to explain how (4)–(7) can all be true given that the proposition that *the bus Ivan saw was going to San Sebatian* = the proposition that *the bus Ivan saw was going to Donostia* (Perry 2012: 239).

Samesaying, first pass Perry thinks that an utterance **u** is in the samesaying relation to an utterance **u'** only if **u** and **u'** have the same content. While this is a *prima facie* plausible principle, it might not be true of all indirect speech reports. Definite reports of indefinites are a case in point. For example, imagine Anna tells Bob "I asked someone to teach my classes next semester". Anna might report her utterance to Carl with "I told Bob that I asked you to teach my classes next semester", if Anna had Carl in mind when she made her utterance to Bob (Cumming 2020, Kamp & Bende-Farkas 2019). I will ignore this complication, and follow Perry in assuming his simple picture. Consider the following analysis of samesaying (I am using Perry's formulation):

(Samesaying, first pass) $SS(u,u',N)$ iff [$\text{Content}_N(u) = \text{Content}_N(u')$]

'*SS*' stands for the *samesaying* relation; *N* stands for the *network* shared by **u** and **u'**; '$\text{Content}_N(\mathbf{u})$' stands for the *network-content* of **u**, '$SS(u,u',N)$' reads '**u** and **u'** are in the samesaying relation relative to network **N**'.

This analysis does not explain why (4)–(7) are all true reports of Ivan's utterance of (3). This is because, for reasons already mentioned, the network-content of 'Donostia' and the network-content of 'San Sebastian' are the same. Let me repeat briefly why: since many competent language users associate 'Donostia' and 'San Sebastian' with a single file, virtually any subutterance of 'Donostia' will be coco-connected with some subutterance of 'San Sebastian' and conversely. Therefore, $\mathbf{N}_{Donostia} = \mathbf{N}_{SanSebastian}$, despite the different naming-conventions associated with 'Donostia' and 'San Sebastian' respectively.[42]

Samesaying, second pass This is where local networks and threads usefully enter the analysis. Let us assume that the speech reporter is among the audience at the time of Ivan's utterance of (3). Like all the rest of the audience, and unlike Ivan, the reporter has one symbol for Donostia/San Sebastian. The reporter and Ivan are therefore misaligned. Let us consider the local network involved in the conversation in which Ivan and his speech reporter participate. Then, let us ask: What is the thread involved with Ivan's utterance of (3)?

Actually, there are two. When Ivan said (3), ideas flowed to his reference to San Sebastian from his San Sebastian symbol (S_I^{enc}), and to his (distinct) reference to Donostia from his Donostia

[42] Perry seems to grant that much in Perry 2012 p. 244.

symbol (D_I^{enc}). Prior to his uttering (3), when Ivan saw the bus with the sign 'Donostia', he formed a buffer for the city to which the bus was going, and (let us assume) merged that to his D_I^{enc} symbol (see Figure 4.7). Since Ivan takes San Sebastian to be a different city from Donostia, Ivan inferred *that the bus is not going to San Sebastian*, information which he fed into his S_I^{enc} symbol (see Figure 4.7).[43] In the piece of discourse Ivan uttered, the reference to San Sebastian is governed by S_I and not by D_I. Here, the relevant thread (depicted in blue below) does not go through D_I. But when Ivan said "It was headed to Donostia", the relevant thread (depicted in red below) goes through his D_I mental symbol and not through his S_I mental symbol.

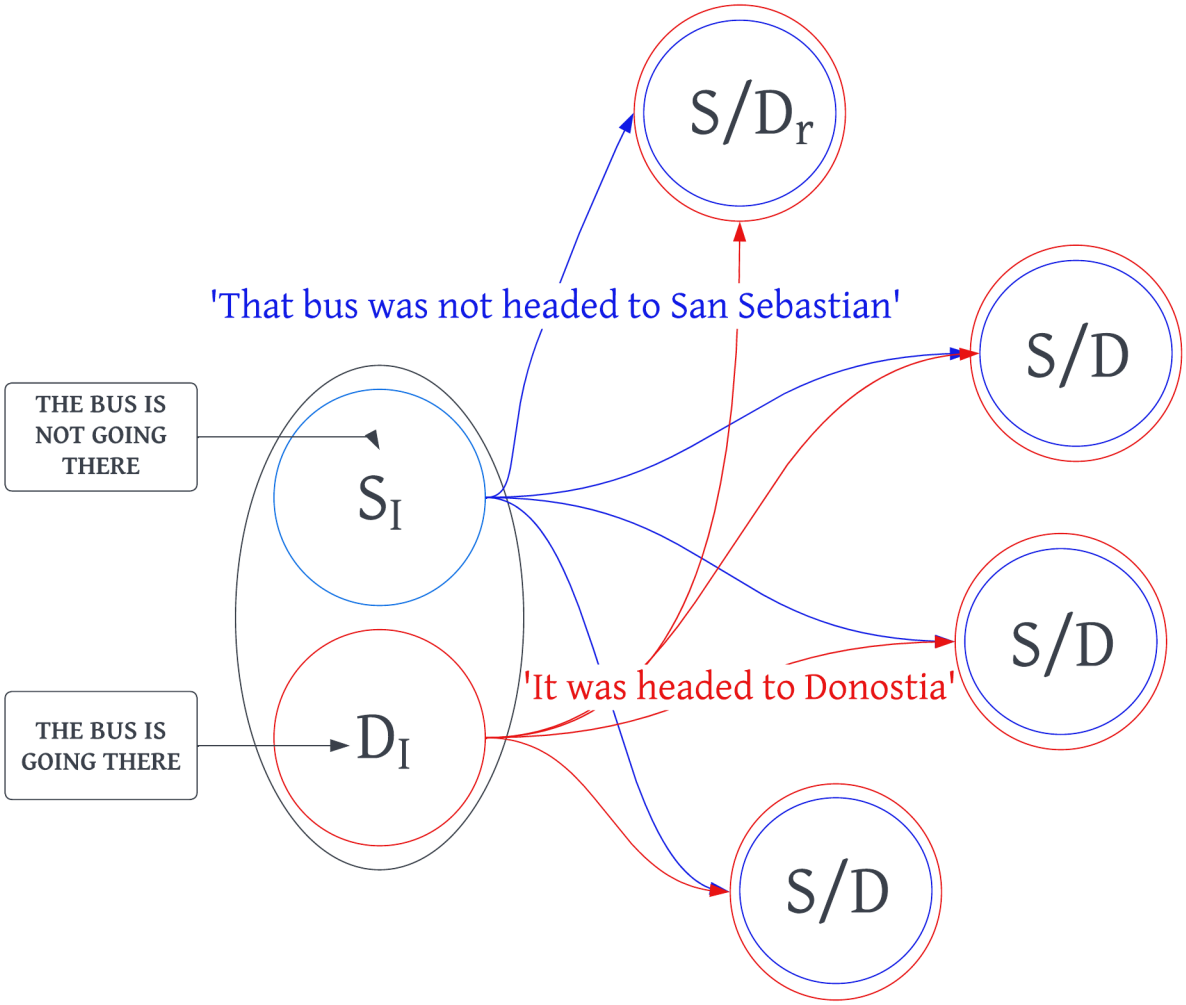


Figure 4.7 – Two threads in the local network of Ivan's utterances

Given the two different threads involved in Ivan's utterance, and given which names are used

[43] In what follows, all mentioned files are *encyclopedic* when the notation is not explicit. I ignore the distinction between lexical and encyclopedic in the notation, because it makes the text cumbersome to read and nothing hinges on this distinction here. Likewise, I do not depict them in the relevant diagram.

in the various reports, we can now explain why (4)–(7) are all true of what Ivan said. I am assuming that, in general, the use of a particular name in a report pragmatically conveys which mental symbol of the reportee is being used to represent the associated part of the attributed content. This is in line with Perry's proposal:

- The utterance of (4) (*Ivan said that the bus was not headed to San Sebastian*) and Ivan's utterance (first bit) are in the samesaying relation relative to the thread induced by S_I (in blue on the graph), which is the relevant thread given the name used in both Ivan's statement, and (4).
- The utterance of (5) (*Ivan did not say that the bus was not going to Donostia*) is true because the subordinate clause of (5) and Ivan's utterance are not in the same-saying relation relative to the thread induced by D_I (in red on the graph), which is the relevant thread given the name used: *the bus is not going there* is not an information to be found in D_I, hence there is not samesaying relative to the relevant thread.
- The utterance of (6) (*Ivan said that the bus was going to Donostia*) is in the samesaying relation to Ivan's utterance relative to the thread induced by D_I (in red on the graph);
- The utterance of (7) (*Ivan did not say that the bus was going to San Sebastian*) is true, because the content-sentence of (7) is not in the same-saying relation to Ivan's utterance relative to the thread induced by S_I. The subordinate clause of (7) *is* in the samesaying relation to Ivan's second statement relative to the thread determined by D_I, but that is not the relevant thread, given the name used.

Let's recap. What we did (following Perry) was to relativize the *samesaying* relation to the different threads involved in the different discourse segments Ivan uttered, and reported by a member of the audience. The sensitivity of indirect reports to the status of particular mental symbols is thus accounted for by relativizing reports to the threads along which participants are pragmaligned.

Although Perry does not theorize in these terms, we may bring *indexed files* into the picture (following Recanati 2012, 2016). Plausibly, the reporter (who is not in a relevant Frege case, hence misaligned with Ivan), deployed indexed symbols, one for each of Ivan's, in order to target the relevant threads, as depicted below (where the letter *r* denotes the mental symbols belonging to the reporter):

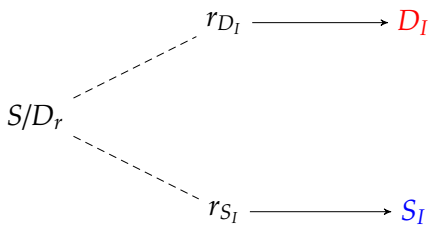


Figure 4.8 – Indexed symbols targeting different threads

A speech-report, on this analysis, can be true relative to one thread, but false relative to another. For instance

(6) Ivan said the bus was going to Donostia.

is true relative to the thread defined by D_I but false relative to the thread defined S_I. When we go back to the mental symbol that is the parent of Ivan's statement 'It was headed to Donostia', we find that S_I was *not* involved — as shown in Figure 4.7 above.

One upshot is that speech-reports are in fact about *threads* — according to Perry, threads are *unarticulated constituents* of speech-reports:

> Let's introduce as a technical locution *believes that F(a) via notion n*. Ivan believes the proposition that San Sebastian is beautiful via his San Sebastian notion, but not via his Donostia notion. (And he also believes the proposition that Donostia is beautiful—the same proposition—via his San Sebastian notion, but not via his Donostia notion.)
>
> A report using this technical locution is *notionally explicit*. According to Crimmins and me, ordinary reports are typically notionally *im*plicit. The notions via which belief are held are not explicitly referred to; they are *unarticulated constituents* of the content of the report. (Perry 2012: 240; emphasis mine)

In our *Ivan* example, the reports were such that 'San Sebastian' picked up the thread defined by Ivan's San Sebastian-symbol; but 'Donostia' picked up the thread defined by Ivan's Donostia-symbol. Perry's strategy consists in keeping the idea that a speech-report should match in content with the reported utterance for it to be true, but does so with a notion of content obtained by relativization of network-content to threads. Accordingly,

(Samesaying, second pass) $SS(u, u', \tau)$ iff $[\text{Content}_\tau(u) = \text{Content}_\tau(u')]$

Where 'τ' stands for the thread the speech report is (covertly) about. With this definition of samesaying, we are now in a position to state Perry's account of indirect speech report. Let's note 'u_C' the subordinate clause of the report that identifies the content of the reported speech event. Then:

(Indirect Speech Report) Given that an utterance **u** of 'A says that p' is about threads $\tau_1,\ldots,\tau_k$,
u is true iff $\exists$ **u'** [A is the speaker of **u'** & [$\text{Content}_{\tau_{1,\ldots,k}}(u_C) = \text{Content}_{\tau_{1,\ldots,k}}(u')$]

Although Perry is not explicit about this, it should be noted that, in misaligned configurations, the level of 'shared content' thus relativized to threads does *not* amount to content identity.[44] In effect, for reasons I have provided in the previous chapter, agents with unlinked coreferential mental symbols introduce spurious information, not match by misaligned agents.[45] Therefore, in a misaligned configuration, speech participants may samethink along a thread, but this does not amount to content or thought identity. Accordingly, the solution in terms of shared content relativized to threads does not validate Shareability (for reasons I have articulated at length in the previous chapter).

3.5 The dual nature of a thread

As I understand them, a thread has a dual nature: it is an external causal-historical path along which information flows (i.e. a part of a file-network), and it is partly internalized by speech participants when they represent a context.[46] Speech participants make implicit assumptions about which thread is operative in a given discourse context, either when planning, or interpreting an utterance. I find this versatiliy attached to the notion of a thread interesting. At the same time, an explicit unpacking of this metaphorical and hybrid notion proves difficult. A thread is supposed to stand for a combination of informational/causal and psychological factors (not unlike the Fregean senses of the sophisticated Fregeans that we reviewed in the previous chapters). A thread really is an hybrid of an objective discourse context, and a subjective one. Let me highlight these two dimensions (external, subjective) in turn.

Threads are represented

One consequence of Perry's analysis is that reports are in fact about threads.[47] I want to highlight that threads are not mere theoretical posits (although, in an obvious sense, they are). Rather, they are meant to have some psychological reality. What is claimed is that people implicitly think and talk about threads when reporting attitudes or speech. More generally, it is claimed that conversational partners make implicit assumptions about threads when producing or interpreting utterances. Linguistic users represent a discourse context in part by representing one or various threads.[48]

[44] Perry is more explicit about this aspect of his proposal in other texts, see e.g. Perry 2003, Perry 2020.

[45] I am following Cumming here.

[46] *Causal descriptivists* only retain this internalized dimension of a thread/network, and believe we can explain all what needs to be explained in terms of this dimension. See e.g. Kroon 1987, Sandgren 2016.

[47] I am skipping over the *Meaning-Intention Problem*, see Schiffer 1992, Braun 1998, and e.g. Rappaport 2017. I limit myself to the following remarks: (1) When reporting a subject in a Frege case, it is plausible that speakers often have in mind particular ways that the subject whose attitude they are reporting has of thinking about the object (i.e. they are able to index files to the perspective of the reportee). (2) But these ways of thinking about objects correspond to mental symbols on the network, which define threads.

[48] Here the notion of a thread interestingly converges with the literature on information packaging (Vallduvi 1992, 1996), and file-cards model of the common ground (Heim 1982, Asher 1986, 1987; other references in Murez & Recanati 2016).

The idea that users make implicit assumptions about threads is reminiscent of the notion of a referential Givenness feature. When a speaker uses a singular term, she plans her utterance by making implicit assumptions about the cognitive statuses their referents have in the minds of their interlocutors. Likewise, a hearer must be sensitive to the (assumed) status of the intended referent as indicated by the chosen expressions. Understanding in singular communication is thus an activity which essentially consists in handling implicit instructions to sort information carried by the discourse. Understanding requires keeping track of the various objects being referred to in a discourse – determining, for each token uses of a singular term, whether a new reference is being made, or a previous reference is being recalled. We may call this, 'thread-management' (adapting a phrase from Murez & Recanati 2016: 270-271; see also Cumming 2014).

Threads are external But threads are also external, objective entities. A thread is a directed path of information flow on a network. A network is an external entity: it is the causal-historical infrastructure underlying the uses of a name. Users exploit such networks when using a name, as they intend to refer to the putative origin of a network, or intend to conditionally-corefer with their linguistic peers along a network. Just as causal-historical chains fail to be transparent (recall my PIERRE example), threads *qua* proper parts of a causal-historical chain, fail to be transparent as well. The Loar cases reviewed in the two first chapters are an illustration of the non-transparency of threads. I will briefly itemize these two aspects in turn, as you should be familiar with them by now.

Name users can be wrong about the identity of the networks they exploit or are in. They can be wrong at the level of the whole network. Remember Peter, who had two mental symbols that refer through the same causal-historical network to the same individual (without his knowledge). But users can be wrong at the level of a *local* network, like Jones in the original Loar case, who assumes that the parent of Smith's utterance is the man on the train, whereas in the actual thread, it is the man on television (see Figure 4.9). Note that Smith is also wrong in the original Loar case about the cognitive status of the discourse referent in the mind of his interlocutor Jones. As a result, there is no joint attention on the relevant ib-feature.

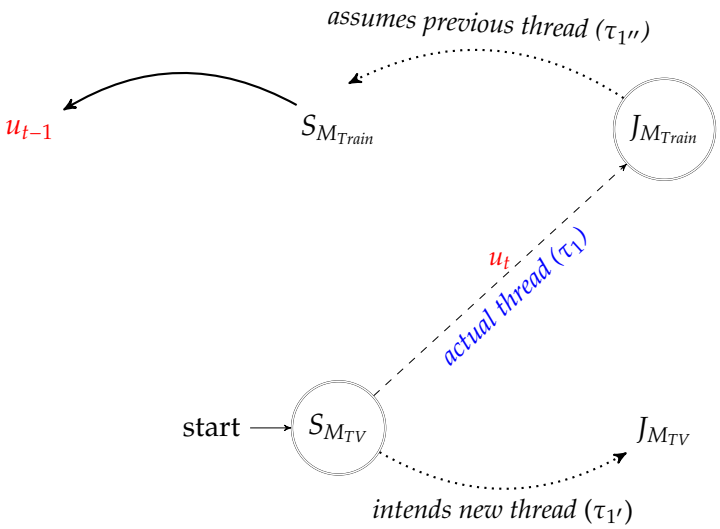


Figure 4.9 – In the original Loar case, Jones is wrong about the 'parent' of Smith's utterance

I now turn to the use of threads in Perry's analysis of beliefs reports.

3.6 A file-network account of attitude reports

As already mentioned, Perry is a pluralist about content. Attitude content is no exception. As Perry remarks (Perry 2012: 241), attitude reports have two main purposes. We may report attitudes to tell about the mental states of the agent, and explain their behavior. The focus is then on cognitive content. Perry says that when reports are used in this way, the use is 'explanatory':

> To the extent that the purpose of the report is to help us understand the agent, it is the *cognitive face* that is important. I call this use '**attitudes as explanation**'. (Perry 2012: 241; my emphasis)

On the other hand, we may report attitudes to focus on the subject matter the attitudes are about, and the worldly information they provide. Such uses, Perry proposes to call 'informational use' of the reports:

> If the agent's views of things are authoritative and accurate, knowing his attitudes provides us information about the subject matter of those attitudes. I'll call this '**attitudes as information**'. In this case, it is the *worldly face* that is important. (id.)

Perry's pluralism about content as applied to attitude reports materializes in Perry's claim that the truth-conditions of an attitude report vary depending on whether the attitude is reported 'as information' or 'as explanation'. In particular, a report which is meant to tell about the reportee's cognitive state should be faithful to the reportee's perspective, in a way that a report meant to be informational (in Perry's sense) is not always expected to be.

Consequently, as we shall see, which level of content is relevant may depend on which use of the attitude is made by the report. When the attitude is used as explanation, the relevant level of content tends to be thread-content. When the attitude is used as information, the relevant level of content may be coarser-grained than thread-content (when applicable).

In Perry's account of attitude reports, this flexibility seems to be achieved through a *uniform* analysis. That is, attitude reports are taken to be relative to *threads* in every case, and the coarser-grained levels of content are obtained by selecting the relevant threads. So just as for speech reports, attitude reports are taken to be implicitly about *threads*: they have threads as 'unarticulated constituents'. *Which* threads are relevant is, again, determined *pragmatically* and not by the semantic contribution of the subordinate clause of the report. Perry's account is thus in the contextualist family.

As before, let's note 'u_C' the subordinate clause of the report that identifies the content of the reported attitude. Here is the account:

(Attitude Report) Given that an utterance **u** of 'A believes that p' is about threads $\tau_1,\ldots,\tau_k$, **u** is true iff $\exists\ \sigma$ [σ is a belief of A & [$\text{Content}_{\tau_{1,\ldots,k}}(u_C) = \text{Content}_{\tau_{1,\ldots,k}}(\sigma)$]

So, when the reportee has two unlinked mental symbols for the object under discussion, and a report is intended *de dicto* ('as explanation'), it is not sufficient that the state of mind of the reporter (expressed in the subordinate clause), and the state of mind of the reportee, share their network-content. In addition, the states of mind must share their network-content along the *relevant* thread (the one defined by the mental symbol the report is implicitly about). Here, the aforementioned principles apply:

(A) Hyperintensional discourse context A discourse context is *hyperintensional* if a discourse participant has unlinked *labelled*-mental symbols for some object under discussion.

(B) Same-content and threads When the discourse context is hyperintensional, then the *same-content* relation is relativized to the relevant thread. Otherwise, *same-content* tracks *same-network-content* or else coreference.

To see how the account works, let's take Perry's other *Ivan* example (Perry 2012: 249). The reports I will consider about Ivan are the following:

> **(8)** Ivan thinks that man [where the complex demonstrative refers to Jesus-Marie] is an idiot.
>
> **(9)** Ivan thinks that Jesus-Marie is an idiot.

The context of the report is that Ivan has two files for Jesus-Marie: one is a stable file labelled 'Jesus-Marie' that Ivan formed in response to email from Jesus-Marie, and to gossip with colleagues about him. This file has been acquired 'on the strength of verbal inputs alone' (Kamp

2013). The other file is a perceptual file which is not associated with the name 'Jesus-Marie'.

The reporter uttering (8) bases his report on the fact that Ivan *said* 'That man is a rude idiot'. The reporter knows that Ivan formed the belief he expressed on the basis of his uncounter with Jesus-Marie (without recognizing him), not on the basis of his standing file. Moreover, it's clear to the reporter that Ivan does not recognize Jesus-Marie and so has two unlinked files for Jesus-Marie. By contrast, the reporter does recognize Jesus-Marie on that occasion. Echoing my proposal of the previous chapter, we may say that the reporter deploys a file indexed to the perceptual file of Ivan when targeting the relevant thread. That indexed file is one of a pair, and the reporter represents the pair as internally unlinked (Recanati 2012).

Given the context then, and given that Ivan's attitude is used as explanation in both reports, (8) is a true report, but (9) is not.[49] The problem is to explain why (8) is a true report but not (9) in terms of the file-structures we have introduced.

The explanation goes as follows. As the use of the complex demonstrative indicates, the utterance of (8) is implicitly about the thread determined by Ivan's *perceptual* file of Jesus-Marie. This thread is depicted in blue in Figure 4.10. As just mentioned, the reporter is justified in believing that the piece of information *is an idiot* is to be found in that buffer, because of the utterance Ivan made, and the conversational situation.

By contrast, assuming that the utterance of (9) is meant to report the attitude as an explanation, and given the name used in the report, the relevant thread for evaluating (9) is the one determined by Ivan's *stable* file of Jesus-Marie. In that file, one does not find the information 'is an idiot'. Hence (9) is false relative to that thread. The situation is depicted in Figure 4.10:

[49] There is a transparent reading of (9), according to which (9) is true, which would be relevant e.g. if the attitude was used in the report as information and not as expanation. The distinction between transparent and opaque readings is due to Quine 1960: 145, drawing inspiration from Russell. I will show how to account for such a reading in Perry's framework in due course.

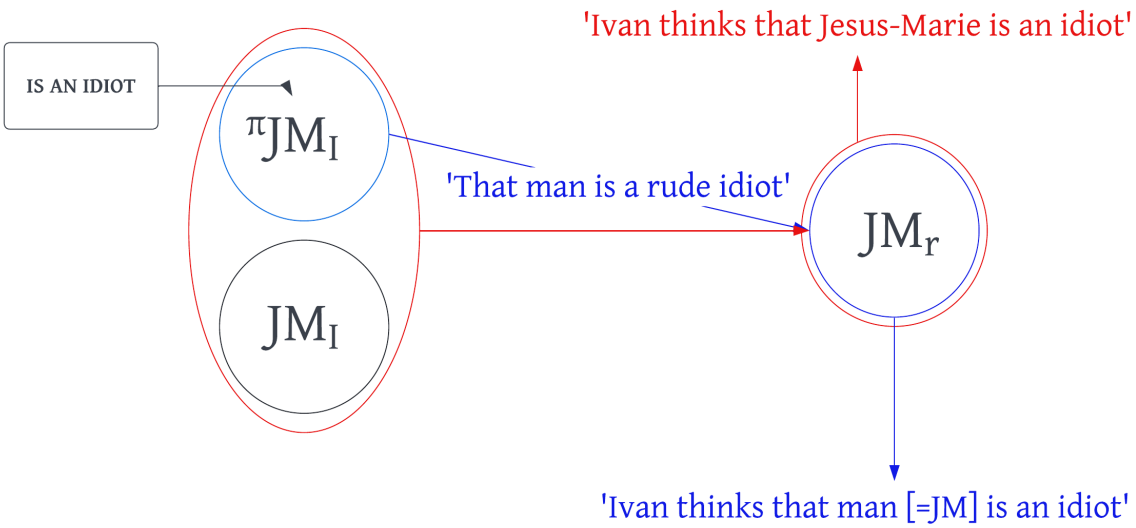


Figure 4.10 – Reports are relative to threads
('π' indicates that the file is a perceptual, unlabelled file)

If the reporter had used the attitude *as information* however, the relevant subnetwork for the utterance of (9) would have been the entire local network, and not the one defined by Ivan's buffer (see Figure 4.10). Let τ_1 and τ_2 be the threads defined by Ivan's buffer and Ivan's stable file, respectively. I take it that we can recover the local network in terms of the conjunction τ_1 and τ_2, because the two threads *cover* the local network (thread in red in Figure 4.10). Why? Because Ivan's stable file is the one he would have used if he had recognized the guy he was seeing; and the audience members, as for them, are not confused. Hence we get the transparent reading by relativizing the content of the subordinate clause of (9) to the local network (on my understanding, τ_1 and τ_2).[50]

In what follows, I explain that the thread-theoretic analysis of reports does not straighforwardly extend to two classes of cases. I examine how we might extend the analysis to these cases. In doing so, I generalize my notion of pragmalignment (with some rearrangement), which was a main goal of the chapter.

3.7 Diachronic reports

The proposed analysis of the truth-conditions for attitude reports requires that the content of a reported attitude, and that of the complement clause of the report, be comparable relative to a *same* thread. However, if a thread is defined as a partition of a local network, and sharing a local network involves interacting in a conversation, then (trivially) it seems that not all attitude reports involve threads. To illustrate, consider the following report made by me:

[50] I am extrapolating somewhat; in fact to my knowledge Perry 2012 is not explicit as to how we get the transparent readings of reports given that reports are always relative to threads. I'm not confident that what I say is what Perry has in mind.

(10) Aristotle believes that Plato is a philosopher

There is no time at which I share a thread with Aristotle, if a thread is defined as a partition of a local network, and sharing a local network involves interacting in a communicative exchange. With an interactive notion of a thread, thread-relativization makes sense with respect to communicative episodes, but seems inadequate when we ascribe content to agents from different epochs of time.

One anwer to this worry is to remark that attitude reports do involve a communicative path in at least one direction. To use my example, I can know about Aristotle's past states of mind by reading Aristotle's writings in translation. The writings mediate a link from Aristotle's mental symbols to mine. In particular, the use of a name by an agent is governed by a mental file labelled with that name in the mind of the agent. Aristotle used written signals to express his PLATO-mental symbol. The signals Aristote used, still exist today (at least some of them). The token of my use of the name *Plato* in the report is 'coco-connected' with the signals Aristotle used to refer to Plato. Hence, I may target the thread that leads from Aristotle's PLATO-mental symbol. Therefore I share a thread with Aristotle even though we never interacted.

It might be objected that sharing a network is not sharing a thread. Clearly, my mental symbol for Plato cannot be pragmaligned (as I have defined this notion in the previous chapter) with Aristotle's symbol, which no longer exist. So the presence of a causal-historical link between agents does not entail that the agents share a thread, when no conversation takes place between them.

My answer to this worry is that the very practice of ascribing content defines a thread between me and Aristotle, because it establishes a communicative path (at least in one direction). When, relying on Aristotle's writing, I report Aristotle's state of mind with an utterance of (10), I thereby establish a thread involving Aristotle's particular notion that is the parent of the references to Plato he made, references that I can still perceive today. My report in (10) is implicitly about Aristotle's PLATO-mental symbol. What I convey with my report is that if you go back to the relevant thread, you will find the information *is a philosopher* associated with Aristotle's PLATO-mental symbol.

This extended notion of a thread is what Perry seems to have in mind with the definition he provides in his glossary:[51]

> A thread consists of the nodes that lead to or lead from a specific [notion] in the mind of an agent. If a an agent has two [notions] of the same object without knowing it, there will be two threads passing through that agent's mind, each determined by one of the [notions]. (Perry 2012: 299)

[51] This definition is non-equivalent with the definition he gives in the relevant chapter (Perry 2012: 244).

Perry's formulation suggests that a thread can be construed in terms of a path in a file-network that connects the informational state of a certain agent (such as the belief of Aristotle that Plato is a philosopher), and the state of a reporter (such as the state of mind I express in the subordinate clause of the report (10)). In other words, the state of mind I express with my report, and the state of mind of Aristotle I thereby report (if my report is true) are *reachable* along a thread. My report is about that very thread.[52]

3.8 Two notions of pragmalignment

I suggest we may generalize the *pragmalignment* relation in terms of thread sharing understood as above. So there are two notions of pragmalignment. One notion is about the alignment of activated mental symbols in a given context. This is the notion I have presented in the previous chapter. The other notion of pragmalignment is about thread sharing, where the relevant thread is the one that is targeted in a report, or otherwise supplied by context. In an attitude report, the relevant thread is the one defined by the mental symbol that is associated (as assumed by the reporter) with the information featuring in the attributed content. Both notions of pragmalignment have to do with the cognitive statuses of particular mental symbols from and to which information flows. We may subsume both of them under the more general notion of pragmalignment as (roughly) alignment *of the relevant mental symbols* in a context, where *relevant* means either *activated* when the context is synchronous (e.g. in communication), or *reachable along the distinguished thread* when the context is distributed/diachronic (e.g. in reports).

According to the extension of pragmalignment I advocate, my Plato–mental symbol can be said to be pragmaligned with Aristotle's, even though Aristotle is no longer an active thinker (and no longer has communicative dispositions at the time of the report). On this construal, pragmalignment does *not* require communicative path in both directions: a communicative path in one direction may be enough. The communicative path from Aristotle's mental symbol to mine is mediated by Aristotle's writing. When I track references to Plato in Aristotle's texts, I deploy a strategy of interpretation which is congruent to the strategy of expression Aristotle deployed at the time of writing. Thus we share a distributed context.

3.9 Alignment through indexed mental symbols

Lastly, we can bring the idea of alignment through indexed files into this picture[53] The basic picture tells us that, in order to report the attitudes of an agent which is identity-confused about an object, the speaker must be able to target the thread defined by the mental symbol which is associated with the information featuring in the attributed content. The basic picture enriched with the indexing idea tells us that targeting the relevant thread requires indexing files to the reportee in such a way that alignment is metarepresentationally restored between the reporter

[52] Thus a report can be said to have a reflexive element.

[53] I borrow indexed files from Recanati 2012 chapters 14-15, Recanati 2016. Recanati suggests we may think of an indexed file as subfile of the regular file of the reporter about the reportee.

and reportee. More specifically, the speaker must (i) index files to the identity-confused agent so as to restore alignment and (ii) deploys the 'good' indexed file for each targeted thread, namely, the indexed file that represents the mental symbol of the reportee that is associated with the information that features in the attributed content. In other words:

> ***De dicto* attitude reports:**
> An attitude report in a context $\mathcal{C}$ is true *de dicto* only if:
>
> > **(a)** The reporter indexes to the reportee as many symbols as required to restore alignment with the reportee in $\mathcal{C}$ [the reporter indexes n mental symbols to the subject of the report ($n \geq 2$) iff the subject of the report has n mental symbols to think about the object];
> >
> > **(b)** The content of the reported attitude relative to the thread defined by the relevant mental symbol of the reportee is the same as the content of the reporter's state expressed by the subordinate clause in $\mathcal{C}$. In particular, for each *de dicto* occurrence of an expression in the scope of the report, and for every i ($0 < i \leq n$), if the subject S of the report represents the associated part of the attributed content with the mental symbol s_i, then the symbol the reporter expresses as that very occurrence in $\mathcal{C}$ is $< s_i, S >$ namely, the mental symbol that represents the symbol s_i of the reportee in the mind of the reporter.[54]

Condition **(a)** is the alignment condition (i.e. through indexed files), and ensures that the reporter makes as many distinctions as required to correctly represent the perspective of the subject whose attitude is reported. For example, if I report an attitude of Peter about Paderewski and I index exactly *one* symbol to Peter, then my report cannot be true *de dicto*, at best it can be true *de re*. The reason is that the reporter is not making enough distinctions in order to correctly represent Peter's perspective on Paderewski, which involves two symbols.

Even if the reporter makes enough distinctions (i.e. even if condition (a) is met), the report can fail if the mental symbol indexed to the reportee is not the 'good' one. Hence condition **(b)**, which makes a bridge between Perry's notion of sameness of content relative to the relevant thread, and the indexing idea.

Let me illustrate the account with the two reports below:

> **(11a)** Peter believes that Paderewski is a musician.
> **(11b)** Peter believes that Paderewski is a politician.

On the picture I am suggesting, the two reports can be true on a *de dicto* reading only if the misaligned non-identity-confused speaker has *two* files indexed to Peter, one for each mental symbols Peter has to think about Paderewski. Otherwise, the reports will be true on a *de re*

[54] I am relying on the notation proposed in Recanati (2012: 183); I am also drawing on Cumming (2013a: 9).

reading only (which only requires a match in reference). *In addition*, the reporter must deploys the relevant indexed file: the one that distinguishes the mental symbol of the reportee that is actually associated with the information featuring in the attributed content. Here is a schema to illustrate the conditions required to think a true *de dicto* reading of **(11a)** (Figure 4.11):

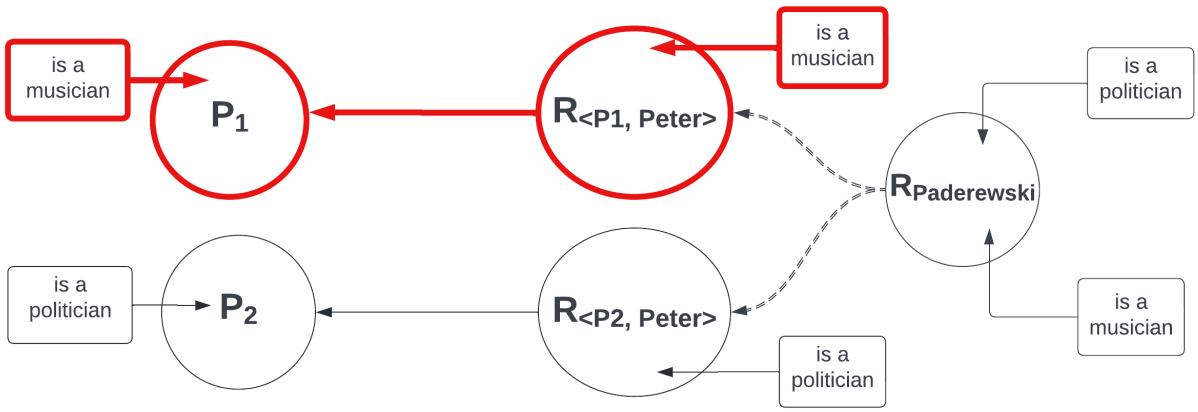


Figure 4.11 – Necessary conditions for a *de dicto* reading of (11a) by an 'enlighted' speaker. Pragmalignment through indexed files along the thread depicted in red.

Here is the symmetrical schema to illustrate the conditions required to think a true *de dicto* reading of **(11b)** (Figure 4.12):

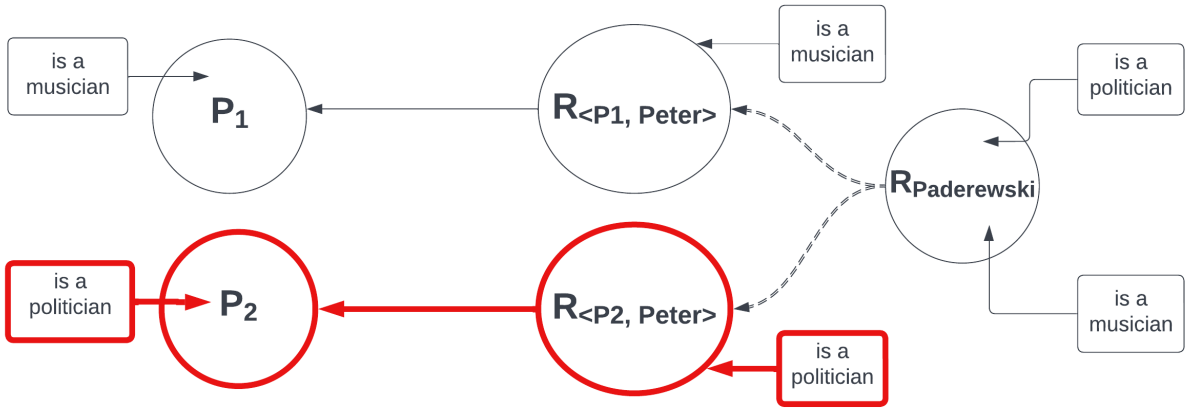


Figure 4.12 – Necessary conditions for a *de dicto* reading of (11b) by an 'enlighted' speaker. Pragmalignment through indexed files along the thread depicted in red.

Finally, the context can be such that, although the reporter thinks a true *de dicto* report, she cannot felicitously communicate the *de dicto* reading to the audience (e.g. if they don't satisfy the conditions (a) and (b) themselves in the context). For example, imagine that Anna thinks a *de dicto* reading of (11a) and satisfies conditions (a) and (b) of the definition (her silent attitude report instantiates the configuration depicted in Figure 4.11), in a context where Paderewski is

giving a political meeting. Were Anna to make an utterance of (11a) in that context, she would fail to communicate the intended *de dicto* report, given that the salient thread supplied by the context is the one defined by $\mathbf{P_2}$. At best, Anna would communicate a true *de re* report.

The remark of the last paragraph is meant to highlight that conditions (a) and (b) are not intended as a theory of the truth-conditions for *de dicto* reports. For example, I am leaving out the issue of spelling out the conditions of the felicitous reception of a *de dicto* report by the audience (dimension which makes the issue incredibly more difficult). Rather, I am providing an account of the samethinking relation that must hold between reporter and reportee for the reporter to be able to think a true *de dicto* report about an identity-confused misaligned agent.

The pragmalignment picture enriched with indexed files is more flexible than Cumming's version, who is forced to say that speakers cannot report *de dicto* Peter's beliefs *even when they know that Peter has two symbols for Paderewski* (a counterintuitive result).[55] On the picture I suggest, correctly metarepresenting the perspective of a misaligned agent is a way to restore alignment in a misaligned configuration (intersubjective or across time). Correctly metarepresenting the reportee's perspective about the object she is identity-confused about, enables one to think a true *de dicto* reports about the misaligned agent.

3.10 Threads and counterfactual reports

The proposed analysis of attitude reports in terms of file-networks does not seem to be able to explain the comparison of content *across* different possible scenarios. But we need to compare content across different possible scenarios to explain the truth-conditions of utterances such as:[56]

> **(12)** Even if my father had tried to convince me otherwise, I would still have wanted to do what I want to do now: explore the cloud forest of Ecuador.

Assessing the truth of (12) involves comparing the content of my actual desire (in the actual world in which my father does not try to convince me) with the content of my counterfactual desire (in a counterfactual world in which he does). The counterfactual (12) is true if and only if the content of my counterfactual desire matches the one of my actual desire. But there is no causal link between my actual attitude and my counterfactual attitude, because causal links do not span different possible scenarios. There is no (actual or counterfactual) file-network that my counterfactual desire and my actual desire share. How, then, could they share or fail to

[55] Cumming is categorical:

> Peter has two symbols that refer to the polymath Paderewski, while neither has the same content as any other agent's symbol for Paderewski. Hence, the reports [(11a) and (11b)] are false on a *de dicto* reading, but true on a *de re* reading. (Cumming 2013b: 395).

But Cumming 2013a looks more in line with what I am suggesting here, see pp. 9-13.

[56] See Cumming 2013b: 393 for a similar example and a different treatment.

share network content? Perry's account seems incomplete in that it leaves out counterfactual attitude reports.

First, observe that the problem is *not* about making sense of the content of a counterfactual attitude in terms of a file-network. Since a possible scenario is a complete possible history of the universe, a pair of counterfactual attitudes may share a counterfactual file-network in a possible scenario. But we at least understand how to compare counterfactual attitudes when they share a counterfactual file-network.[57] What we apparently cannot make sense of is the idea that a causal link connect an actual attitude with a counterfactual one.

One reaction is to propose that counterfactual attitudes derive their representational repertoire (i.e. their mental symbol constituents) from actual ones. This is a very plausible claim. But this does not address our problem. What we want to do is to compare the content of full-fledge attitudes which are not 'worldmate' (Lewis 1986): we want to determine how the distribution of predicative entries in mental files vary across possible scenarios.

I admit that the problem seems a bit odd. We could try to make sense of an *ersatz* of a causal-historical chain bridging different scenarios as follows. We do not need a trans-world link to compare the content of an actual attitude with the content of a counterfactual attitude. Rather, all what we need is to compare the content of an actual attitude with the content of a *simulated* attitude, given a simulated thread. "Simulation" refers to the ability we have to mentally project ourselves into a situation different from the one we are in, and in doing so see or think about the world from another perspective. The situations or perspectives to which simulation gives us cognitive access can be the actual perspective of another person, the fictional perspective of a fictional character, or the perpective we would have, given a particular counterfactual scenario.[58] In the model of interworld content comparison I am suggesting, the relata of the comparison are contents of *actual* attitudes. One relatum is the content of the actual attitude to be compared. The other relatum is the content of the simulated attitude, which is another actual attitude (namely, an imagining). It seems that we could thus make sense of the comparison of content across possible scenarios in terms of a file-network. Let me explain by way of the aforementioned example.

My counterfactual attitude takes place in a world (on a file-network) in which I have had a conversation with my father about my wish to explore the cloud forest in Ecuador, and he tries to convince me not to do that. That counterfactual file-network determines whether there is an association, under the attitudinal mode of desire, of the idea *that I explore the cloud forest there* with my mental symbol for Ecuador. The difference between the *actual* file-network rooted

[57] I don't mean that counterfactual attitudes do not pose problems for theories of content, see section 2.3 of Ninan 2008, 2012, Maier 2016b, and Blumberg 2018.

[58] On simulation, the *loci classici* include Stich & Nichols 2003, Nichols 2006. See also e.g. Recanati 2000, 2021, 2022.

in Ecuador, and the *counterfactual* file-network rooted in Ecuador, lies in the pedigree of the putative association of the idea *that I explore the cloud forest there* with my Ecuador file.[59] The counterfactual pedigree includes a set of cognitive and linguistic nodes, and a set of links relating them, which we do not find in the actual file-network rooted in Ecuador (everything else being invariant). These additional nodes constitute the counterfactual local network involved in the (counterfactual) conversation with my father.

How does the idea of a *simulated thread* help with the issue? Consider the actual thread of the file-network rooted in Ecuador which leads to my actual desire. I may update *this* thread in imagination, given the scenario in which my father tries to convince me not to visit Ecuador.[60] The counterfactual (12) is true if, along the simulated thread leading to my simulated Ecuador–file, I find the idea *that I explore the cloud forest there* (under the attitudinal mode of desire), otherwise it is false. Now, the simulated update of my Ecuador–file itself occurs on the *actual* file-network (see Figure 4.13). Hence, no need for a trans-world link. The contents compared are those of two actual attitudes.

We may think of the simulated file as a file *indexed* to the relevant simulated agent-slice (here, my counterfactual self). Just as we token files indexed to other agents in order to represent their perspective, we may token files indexed to other agents in order to represent the perspective they would have in counterfactual scenarios. Here is a diagram of the proposed model of content comparison applied to the example (Figure 4.13):

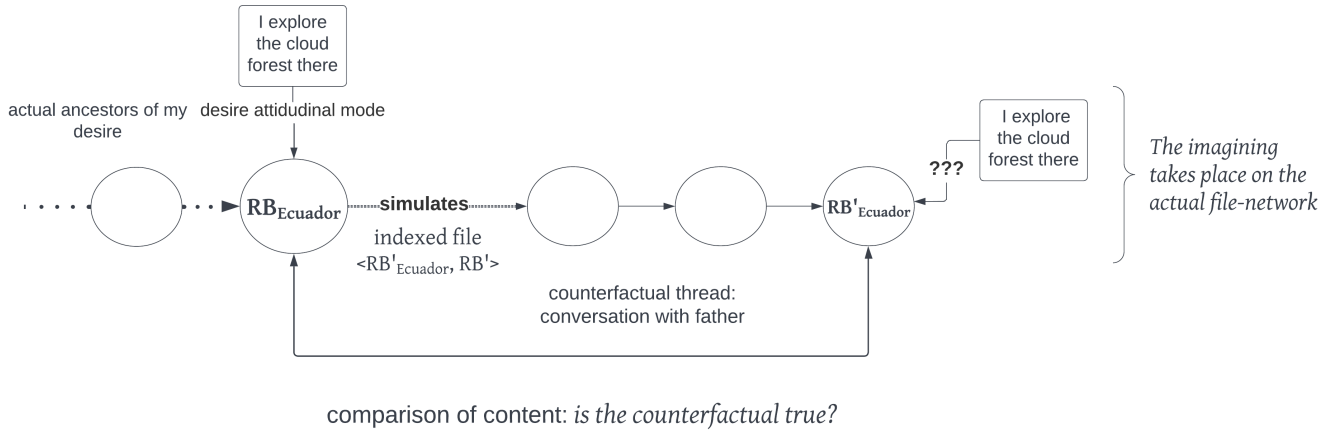


Figure 4.13 – A file-network model of 'interworld' content comparison

To recap: comparing content across possible scenarios is not a problem for a causal/informational network account, if we can explain how this is done in terms of the comparison between the

[59] I take the term *root* from Friend 2011 who takes it from Perry's unpublished "Saying Nothing?" (1997). Our respective token uses are coco-connected, so to speak.

[60] See the proposed cognitive architecture underlying simulation in terms of the Possible Worlds Box and the UpDater in Stich & Nichols (2003: 87).

contents of an actual attitude and of the (actual) simulation of a counterfactual attitude, given a simulated thread. Of course, this hardly constitutes a semantics for counterfactual attitude reports (see Maier 2016b, Ninan 2008, 2012). But this will do for present purposes. I now turn to the problem of accounting for agreement and disagreement between non-interacting agents.

4 Agreement & disagreement without interaction

In a passage already quoted in the previous chapter, Cappelen and Hawthorne (2009) usefully distinguish between two senses of 'agree'. They say:

> The verb 'agree' has a use according to which it picks out a state of some plurality of individuals—where some individuals agree that p if they all believe the proposition that p. There is also a different use according to which it denotes an activity, where agreeing that p is the endpoint of a debate, argument, discussion, or negotiation. On this use, 'agreeing that p' marks an event. The latter use is interactive: it requires that the agents who agree or disagree interact in some way. However, the former use is perfectly applicable to interaction-free pairs of individuals so long as there is some view about the world that they share. (Cappelen and Hawthorne 2009: 60, quoted in Ninan 2016: 100)

The topic of this section is the state sense of *agree*. This section continues the discussion between the conception of samethinking in terms of alignment, and the one in terms of pragmalignment. The aim of this section is to identity further central commitments of these two approaches. Cumming criticizes views which relativize content to communicative paths. Interestingly, Perry can be construed as an instance of such views: threads are paths, and content is relativized to threads. Cumming writes:

> One [alternative to alignment] would be to relativize the standard of correctness to the communicative path taken. In adopting this measure, we effectively relativize content to a path: [a symbol may have] the same content as [another] relative to a path (...), but not relative to [another]. While this makes sense for judging sequences of communicative exchanges, path-relativization seems out of place in other intersubjective content attributions. Suppose you come to believe what I never doubted—relative to path π but not relative to path π'. Is the report "You have come to believe what I never doubted" true? It is not clear how to choose a path and so decide this question, since it is consistent with the attribution that your belief did not originate with me. You might have arrived at it on your own, or from testimony originating elsewhere. (Cumming 2013b: 384)

I take Cumming's objection to path-relativization to be this. When one needs to decide whether two attitude states agree in their content 'in abstracto', there is no context to determine which path is the relevant one. Absent any constraints on the path parameter, path-relativization

is not predictive, because it is unconstrained. This objection can be directly phrased against Perry's account (and mine, to the extent that I have made Perry's theory integral to my criterion of pragmalignment). When intersubjective content attributions are between attitude states of thinkers that do not interact, no particular thread is supplied by context. As a result, the pragmalignment account is too unconstrained to be predictive.

Note that the objection is making an assumption one could reject, namely, that we can compare attitude states 'in abstracto'. If this assumption is false, then it is not an objection to Perry and pragmalignment that the account is too unconstrained in thoses cases. What is at stakes here seems to be two competing views about the nature of questions about agreement and disagreement. *Are questions about agreement and disagreement in the state sense fully decided by agents' communicative dispositions?* Cumming says yes. Perry says no. I will call the first option, 'moderate contextualism' about agreement without interaction; the second I call 'radical contextualism'.[61] *If* radical contextualism is false, then agreement in the sense at issue is not straightforwardly explainable within Perry's framework. If moderate contextualism is false, then Cumming's objection stems from a unwarranted assumption: it is not an argument in favor of alignment.

Before I present the radical contextualist view of the matter in more detail, let me give a more vivid sense of what Cumming's objection is, as I understand it. I will illustrate the point of issue on the PIERRE example of the previous chapter. All the relevant threads are depicted in Figure 4.14:

[61] I label Cumming's view 'contextualist' because his alignment relation is relative to sets of agent. Which set of agents is set as a parameter is left to the discretion of the interpreter, depending on their explanatory interests. So the view counts as contextualist.

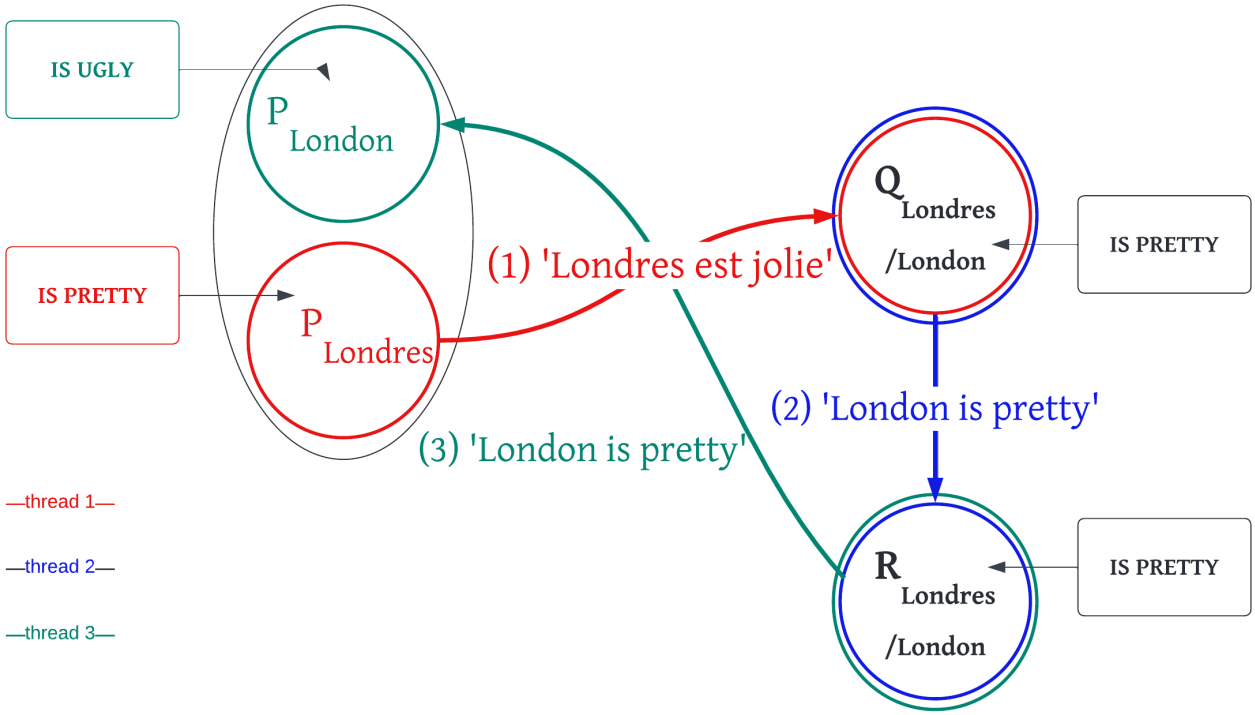


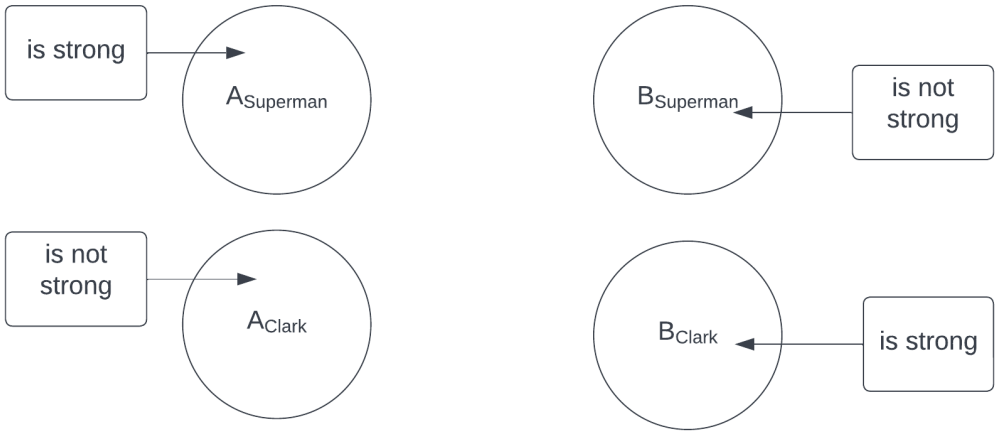
Figure 4.14 – The local network involved in PIERRE

So for example, let us ask, Do **P** and **R** agree or disagree? **P** agrees with **R** along the thread in red, and disagrees with **R** along the thread in green (where **R** deploys the same thought relative to both threads). When we go back to the thread defined by **P**'s 'Londres'-mental symbol, we find the information IS PRETTY associated with that symbol. This belief shares network-content with a belief of **R**, so they agree. And, there is another thread along which **P** and **R** disagree, namely, the one leading to **P**'s 'London'-mental symbol where the information IS NOT PRETTY is associated with that symbol. So **P** disagrees with **R** because **P**'s belief along that thread is the negation of the network-content of **R**'s belief. By the same token, **P** both agrees and disagrees with himself along these two threads.

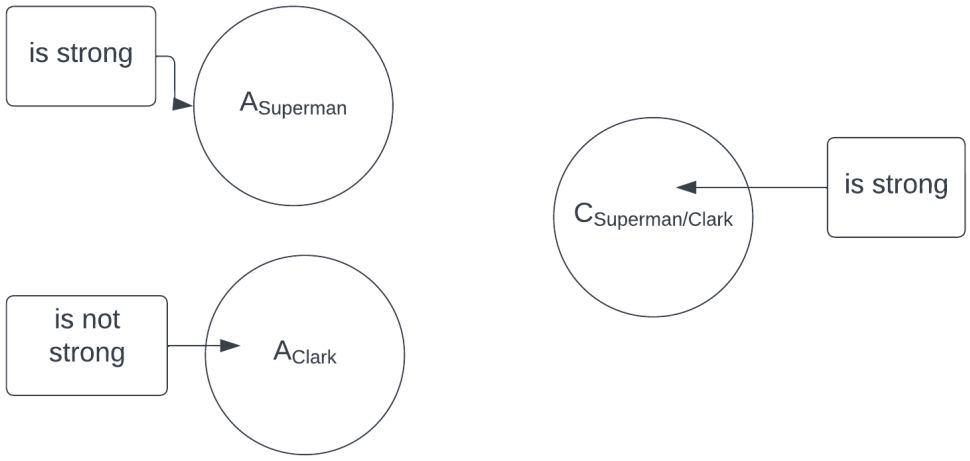
As you can see, sharing content along a thread does not amount to content or thought identity. For if shared-content relativized to thread *was* a matter of content identity, we could not find that a single belief of **R** is in the agreement relation to one belief of **P**, and in the disagreement relation to another belief of **P** all at the same time, while **P** respects norms of minimal logical consistency. Agents which are insensitive to the coreference of two of their mental symbols introduce thereby 'spurious information', not match by misaligned agents (Cumming 2013b).[62] Unless one is willing to embrace radical contextualism about agreement without interaction, the thread-treatment of agreement and disagreement without interaction does not look unproblematic.

[62]See my discussion of alignment in the previous chapter.

What the PIERRE example suggests is that thread-relativization, when no conversational cues and pragmatic factors are there to determine which thread is relevant, actually undermines a systematic theory of the propositional attitudes. For example, once we relativize agreement without interaction to threads, it seems to become trivial that a person can intend to do something without thinking that they will do it. Imagine that Tarik (who does not speak French) intends to go to London, and **P** intends to go to London via his Londres-symbol; then Tarik and **P** agree in their intentions to go to London relative to some thread (they have intentions that share their network content relative to some thread). I'll make the point even more vivid. Constrast these two intersubjective configurations of propositional attitudes (Figure 4.15):



(a) Aligned



(b) Misaligned

Figure 4.15 – Agreement/Disagreement relative to threads

Let's review the misaligned case first. Agent A agrees with agent C relative to the thread de-

fined by A's Superman-mental symbol. And A disagrees with C relative to the thread defined by A's Clark mental symbol. Now, consider the aligned case. Are we to say that A agrees with B relative to different threads, namely the thread defined by A's Superman-mental symbol, and the thread defined by A's Clark symbol? But it seems to me that A and B disagree! Of course, we can *also* explain that A and B disagree in terms of threads: if we go back to the thread that leads to A's Superman mental symbol, we find the predicative entry IS STRONG. And, we find a contradictory predicative entry associated with B's Superman symbol along the thread defined by that symbol. So they disagree with respect to this pair of threads. Likewise with respect to the pair of threads defined by A's and B's Clark symbols. At this point, we may rephrase Cumming's objection, as follows. The problem with relativizing the relation of agreement without interaction to threads is that there are no contextual constraints on which thread is supposed to parametrize the content attributions. As a result, the account is too unconstrained to be predictive, in part due to the fact that we don't have a recipe to *type* threads and their constituents.

How can we individuate threads, and the mental symbols defining them? Perry does not really say, but gives some hints. I take it that comparing threads is a contextual process through and through in Perry's framework. To compare two threads, you need to know about how the agents think about the origin of the network, in particular, what type of file they have to think about the origin of the network, and to compare the causal-historical pedigree of the respective files. In other words, you need to know about the respective threads that lead to the mental files to be compared. Other than that, I can't think of a more systematic way to compare threads. For example, in the case of name-involving attitudes, we might often be able to type threads in terms of public names when they are conventionally associated with particular ways of thinking about the object — (roughly) what Crimmins 1992 calls *normal notions* (e.g. 'Clark Kent' way of thinking vs 'Superman' way of thinking).[63] But, someone who is identity-confused and associates one name with her symbol but not the other, cannot share a type with someone who knows the relevant identity and associates both names. Moreover, in cases where a subject associates two different mental symbols with a *single* public name (which the subject wrongly takes to be two names; these are the Paderewski type of cases), we cannot type symbols in terms of their association with a public name.[64]

However, while alignment makes more categorical predictions absent any constraints on the thread parameter, it should be noted that alignment does not provide a recipe to type mental symbols in terms of their functional role. Rather, communicative dispositions are only a part of a mental symbol's functional role. It might just not be incumbent on a theory of communication and reporting (which is what samethinking is all about) to provide a recipe to determine how

[63] Also, this is what Perry does in nearly all of his examples in Perry 2012, chapter 10.

[64] But, when there are two files associated with one public name, there will (arguably) be information associated with one file that is not associated with the other (Gray 2016), and so we might be able to type the files through the relevant information.

similar two thinkers are in their way of thinking with respect to a subject matter. But the question *how similar do two thinkers are in their way of thinking about an object* is perhaps the real question we ask when we want to compare attitudes 'in abstracto'.[65]

What if the moderate contextualist assumption about agreement without interaction (*questions about agreement and disagreement in the state sense are fully decided by agents' communicative dispositions*) is misguided? The rationale behind radical contextualism is that one cannot compare attitude states in the abstract (Taschek 1998). On this construal, threads are tools we use to make sense of others' attitudes in context. Questions about agreement in abstracto are perhaps an instance of philosophical problems that arise when language 'goes on holiday' (Wittgenstein 1953, 38). Questions about agreement and disagreement between attitude states are essentially contextual. When there is no context of utterance, and no joint mental context, the only residual notion of context has to do with an interpreter. This is radical contexualism.[66] [67]

It should be noted that alignment might be no less interpreter-relative than pragmalignment. Alignment is not relativized to path. But it is relativized to sets of agents. Therefore, one thing we need to do before comparing two thoughts is to define the *scope* of the distributed context shared by the thinkers whose thoughts are being compared. And *that* also seems to depend on the explanatory interests of an interpreter. So we do not get rid of all interpreter-relative context-sensitivity. We have what I call a 'moderate contextualist' view about agreement without interaction. The view makes more categorical predictions about sameness of content than content relativized to thread, because it is more principled than thread-relativization when there is no constraint on the thread parameter. Once the scope of the shared distributed context is fixed, alignment decides whether two thoughts agree or disagree for any pair of attitude states on a given file-network. (I refer the reader to the end of the footnote below for remarks that qualify this last claim).[68]

[65] Stich 1983 (in particular chapter 7) contains a lot of relevant discussion.

[66] See Stalnaker 2008 for a view of this kind. Schroeter's (2013) review of Stalnaker's (2008) book ascribes a similar view to Perry.

[67] When agents do not have unlinked notions for an object, we can capture agreement and disagreement relations between such agents simply in terms of sameness of network-content. Let A and B be two non-identity-confused agents. Then, A and B agree on something in virtue of having the token beliefs β_A and β_B (respectively), just in case $\text{Content}_N(\beta_A) = \text{Content}_N(\beta_B)$, where (restricting to beliefs of the form $\ulcorner a \text{ is } F \urcorner$) network content is defined as

> **Network-Content:**
> $\text{Content}_N(\beta_\alpha) = \exists x\,(O(x, N^\alpha) \,\&\, Fx)$
> In natural language: the network-content of a token belief β_α is that there is an object which is the origin of the network along which the mental symbol in subject position in β_α lies, and that object is F.

[68] How might we define a *same-thought* relation in terms of alignment? As Cumming defines it, alignment is a relation between mental symbols, not between thoughts. The Cummingian networks of communicative policies only include mental symbols. Hence we cannot use them to compare attitude states such as beliefs. File-networks are much richer. They include full-fledge attitude states (an attitude is typically realized by a file on some file-network). Perry does countenance an array of network contents for mental predicates, for he defines network content of the attitudes. We may use Perry's proposed network apparatus, and add the alignment constraint on top of it. The file-network would enable us to locate any pair of attitude states to be compared according to the relation

Let us conclude this section. Is content attribution always contextual? The pragmalignment view is committed to an affirmative answer. This is compatible with the *metaphysics* of agreement and disagreement being non-contextual. Alignment does not bear on the metaphysics of agreement and disagreement: alignment-based 'content' does *not* determines functional role (except incidentally for the bits of computational roles that are involved in communicative dispositions attached to mental symbols). So the comparison of alignment-based contents does not decide questions having to do with the metaphysics of agreement/disagreement. However, even dimensions such as *similarity of functional roles* do not look a-contextual: similarity dimensions can be weighted in more than one ways; the interpreter might still be here.

5 Taking stock

The main goal of this chapter was to provide an account of the relation of samethinking as it occurs in attitude and speech reports. Pragmalignment defined in terms of activation of mental

of agreement (equivalently, *sameness of thought*); alignment would determine whether the thoughts are the same. I will restrict to thoughts of the form $\ulcorner a$ is $F\urcorner$. Thoughts of this form are realized by mental files. They include (i) a mental symbol (ii) a mental predicate (the copula is encoded by the association of the mental predicate with the mental symbol). Roughly, we may say that two token thoughts of the form $\ulcorner a$ is $F\urcorner$ are the same, just in case (i) the mental files which realize the thoughts are reachable along a file-network; (ii) the mental symbols in subject position in each token thought are aligned, and (iii) the respective mental predicates in each thought are aligned. (When dealing with more complex thoughts, involving e.g. a relation or more than one predicate, a notion of *isomorphism* is required. As long as there is only one predicate, we can keep the isomorphism requirement implicit.) Here is a tentative definition:

> **Agreement:**
> Let β_A and β_B be token beliefs belonging to A and B respectively. Let a and b be the mental symbols in subject position in β_A and β_B respectively. Let Ψ and Ω the mental predicates in β_A and β_B respectively. (So, β_A is the belief realized by the association (under the attitudinal mode of belief) of mental predicate Ψ with mental symbol a, and β_B is realized by the association (under the attitudinal mode of belief) of mental predicate Ω with mental symbol b).
> Then, A and B agree on something in virtue of having the token beliefs β_A and β_B (respectively), just in case:
>
> > **(i)** there is a file-network **N** such that β_A and β_B are reachable along **N** [Equivalently: $\text{Content}_N(\beta_A) = \text{Content}_N(\beta_B)$];
> >
> > **(ii)** $a \rightleftharpoons b \wedge \Psi \rightleftharpoons \Omega$

The alignment relation in condition (ii) can be parametrized according to a distributed context of varying scope. At the minimum, we must have

$$a \rightleftharpoons_{\{A,B\}} b \wedge \Psi \rightleftharpoons_{\{A,B\}} \Omega$$

But we may require a wider distributed context when needed. The more agents there is in the set that parametrizes *alignment*, the more difficult it is for thinkers to be in the relation, because the possibility of indirect forking or pooling incereases in proportion with the cardinality of the set of agents that parametrizes alignment. The proposed criterion is not good as it stands, because it is synchronic. But we may define a more general relation which subsumes alignment, in terms of information (Cumming op.cit., Dretske 1981). Two thoughts are the same (i.e. agree) if they carry the same information. But see my sceptical remark at the end of this section: Cumming's notion of information attached to mental symbols does not capture their computational roles, except incidentally for the bits of computational roles that are involved in communicative dispositions. But, if information does not determine computational roles and does not ground reference (it only determines it in the mathematical sense, see Cumming 2013b: 394-395), then it is not clear what sameness of content buys us. One may legitimately wonder: In which sense is alignment-based 'content' *content*?

symbols is a synchronic notion. But we want to make sense of samethinking across different epochs of time. Using Perry's description of the intersubjective file-networks (Perry 2012), I have proposed that we reconstrue the pragmalignment relation of the previous chapter in terms of thread sharing. A thread may be thought of as a path in a causal-historical file-network that connects the informational state of a certain agent (such as Aristotle's belief that Plato is a philosopher), to the informational state of an other agent (such as the state of mind I express when I report Arisotle's belief that Plato is a philosopher). The very practice of ascribing content distinguishes a communicative path, no matter the time span between the attribution and the reported attitude. Threads thus help us to make sense of the sensitivity of reports (when used 'as explanation' i.e. *de dicto*) to the cognitive statuses of particular mental symbols. When reporting the attitudes of a subject in a Frege case, speakers have in mind particular ways the subject has of thinking about the object (which I have proposed to understand in terms of the fact that they are able to index files to the subject whose attitude is reported). In so doing, they implicitly distinguish threads in the network.

A file-network *supports* naming conventions. Users exploit such conventions when using a name *in accordance* with a naming convention. What is it to use a name in accordance with a convention? Which parts of a network support the convention? Do agents confused about the identity of an object (whether they have a single symbol for several entities, or several unlinked symbols for a single entity) support naming-conventions in the same measure as non-confused agents? More generally, what determines the meaning of a an expression which is used along a network? That is, what is the set of facts about a file-network which determines the meaning of a word that users associate with their file in the network? In describing the intersubjective file-networks, Perry helped us to analyze samethinking and information flow along the networks, but did not answer these questions. These questions are the topic of the next chapter, which is also the final chapter of this thesis.

5 Participating in representational traditions

Abstract

If concepts are not shared, how is it that a speaker can *learn*, *be wrong* about, or have merely *partial knowledge* of—what a word means? This chapter addresses this issue. It begins by proposing a typology of the distribution of concepts (i.e. the extent and manner in which they are spread in a population, in a theory-neutral sense of "spread"), and places *word meanings* within this typology. The rest of the chapter provides a metasemantic story to account for *learnability*, *being wrong* and *partial grasp* without shared meanings other than extensions (i.e. referent, class, property, etc).

Drawing on a critical discussion of Fiengo & May 2006 and Schroeter 2012, the proposed metasemantic story relies on two claims. The first claim says that the use of 'common currency' words trigger appearances of semantic sameness in language users. The second claim says that these semantic appearances make it the case that things happen as if meanings were shared, and give rise to representational traditions.

I revisit the notion of the division of linguistic labor in light of the representational traditions to which semantic appearances give rise. Drawing on O'Madagain 2018, I put forward that people defer in the semantic sense in order to defer in the epistemic sense. The goal of the linguistic practices (talking, writing, thinking with 'common currency' words) is to accumulate information on encyclopedic entries of general interest through testimony. Correctness conditions for the use of words has to do (perhaps essentially) with this collective epistemic goal. Following Recanati 2016, I propose that what underlies the division of linguistic labor is a peer-to-peer distributed file managed at the community level.

How are these distributed files managed? I suggest that the way *Wikipedia* encyclopedic entries are managed is a good reflection of the social mechanisms by which the community manages a distributed file. In closing, I say how we can make sense of samethinking without causal link in the present framework.

1 Introduction

1.1 Various patterns of concepts distribution

Call *representational practice*, the practice of using a concept to represent a given entity or kind. When the concept is lexicalized or "labelled", the representational practice is linguistic. Call

a practice's *distribution*, the set of points in space and time where the practice can be found (Morin 2015).[1] We can distinguish representational practices' distributions in terms of their *scope*, and in terms of their *mode*. Representational practices are distributed more or less widely (on a scale from undistributed to community-wide). Moreover, representational practices can be distributed with or without transmission, and the transmission may be linguistic or not. Let me explain.

Many of our concepts (perhaps most of them) are distributed through *linguistic transmission*, that is, via communication with others. Among the concepts distributed via linguistic transmission, we may distinguish between the concepts that are *locally* distributed (i.e. that are context-bound, or at any rate not widely distributed), and the concepts that are distributed throughout a population. Locally distributed concepts include demonstrative context-bound concepts. For example, the representation THIS BUTTERFLY tokened by a pair of hikers jointly attending to a particular butterfly in a portion of forest, is a locally distributed concept. Other concepts are more stable and widely distributed in the community of language users, in contrast with the concepts associated with indexicals.[2] For example, the concept BLOCKCHAIN is a stable and widely distributed concept. Concepts associated with 'popular' names (such as 'London'), are in this latter class.

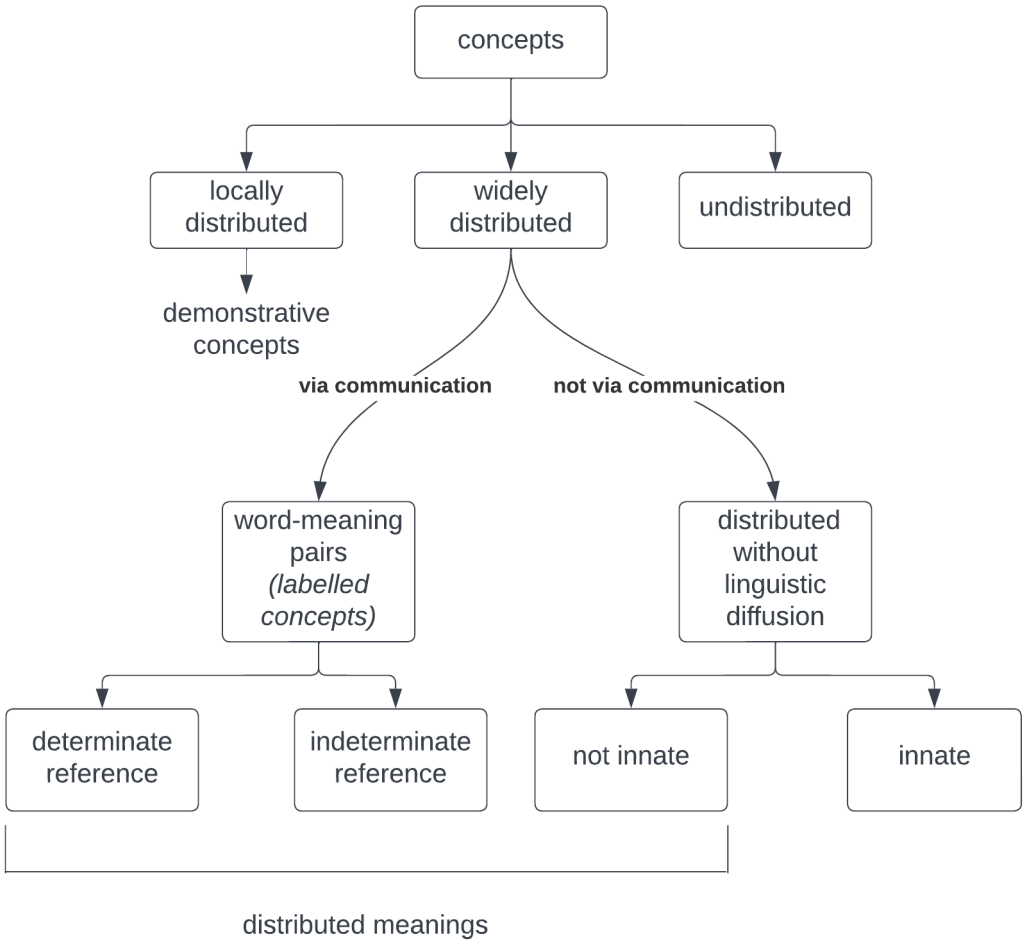
But concepts may be widely distributed *without transmission*, for example as a result of causally isolated thinkers sharing a similar environment and biological makeup, or as a result of causally isolated thinkers using concepts with a shared epistemic goal (Brigandt 2010). For example, Leibniz and Newton, quite independently of one another, both come up with the idea of an integration function (though they used different words to refer to it) that plays the same role in their respective mathematical reasoning. Transmission requires social learning (i.e. learning from others). But another way for concepts to be widely distributed without transmission (in the sense at issue) is through innateness. Some of our concepts arguably have an innate basis e.g. (perhaps) OBJECT or CAUSE (Carey 2009). Non-human animals presumably share concepts without transmission, although it is pretty uncontroversial that at least some non-human cultural transmission occurs. Hence some animal concepts might be spread via cultural transmission. Finally, some concepts are *un*distributed. Examples of these include concepts consigned to episodic memory traces, or proprioceptive phenomenal concepts.

Here is a working taxonomy of various patterns of concepts' distribution, which summarizes the distinctions just made (Figure 5.1):

[1] I intend the expression "distribution" to be theoretically neutral. In particular, the expression does not imply Shareability.

[2] As mentioned in the general introduction, indexical concepts have "roles" or "Kaplanian characters" (roughly, functions from utterance context to content). Hence there is a sense in which indexical concepts can be said to be widely distributed: insofar as these characters are.

Figure 5.1 – A taxonomy of concepts' distribution



This chapter is about the widely distributed meanings. The chapter mostly focuses on the meanings distributed via linguistic transmission, but I will say a word on how to account for the distribution of meanings without transmission in the last section of the chapter. In so doing, I outline an answer to Sandgren's (2019) objection that causal-historical models cannot account for intentional identity without causal link. In this thesis, I have so far mainly focused on singular concepts with a determinate reference, such as NAPOLEON, LONDON or THIS BUTTERFLY. The model proposed here is meant to apply also to concepts whose reference may be more indeterminate, such as BLUE, ICE CREAM, TABLE.[3]

1.2 The need for objective meanings

> Suppose you are like me and cannot tell an elm from a beech tree. We still say that the extension of "elm" in my idiolect is the same as the extension of "elm" in anyone

[3]For an approach that, to some extent, "divorces the notions of meaning and reference", see Richard 2019.

> else's, viz., the set of all elm trees, and that the set of all beech trees is the extension of "beech" in both of our idiolects. Thus "elm" in my idiolect has a different extension from "beech" in your idiolect (as it should).

I am not more knowledgeable than Putnam (Putnam 1975: 143) with respect to elms and beeches: I cannot tell an elm from a beech. Nevertheless, my token uses of 'beech' and 'elm' refer to different kinds of tree and express different concepts. For example, when I ask "Is that an elm?" when facing a tree I have never seen before (let us suppose the tree is not an elm), I still mean elm.

This case suggests that there is what Putnam called *a division of linguistic labor* for these terms (*op cit.*). Linguistic labor determines a term's reference. *I* am not in a position to do this linguistic labor for the words "elm" and "beech". But *other speakers* in my community can distinguish between the two kinds of tree (e.g. the botanists). In virtue of *their* expertise, my token uses of "beech" and "elm" have different referents. So individual idiolects do not always (and typically do not) determine the referent of a term. But the distributed idiolect of a linguistic community somehow does. It is conceivable that a whole community could be radically wrong about the referent of one of their own terms. Even in such cases, the distributed idiolect in the community could help identify the referent of the term, and constitute the supervenience base for reference determination.

Exactly which class of terms in the language is concerned by the division of linguistic labor is an open question. At least the natural kind terms are, and very likely other kinds of words. What is important is that we need an objective notion of meaning to make sense of it. The meaning of 'elm' I mentally represent fails to match this objective meaning, but is deferentially related to it. How can we borrow meanings in this way if Shareability is false? That is the issue of the chapter.

1.3 Linguistic continuants

To get the causal chains that stabilize meanings in human populations, we need to consider networks that include linguistic expressions. We have already examined such networks in the previous chapter, namely, the coco-networks described by Perry 2012. Perry's coco-networks include referential utterances of all sorts. But we want to ascribe meanings to linguistic expressions, one for each. Perry's coco-networks are coarser-grained than the linguistic continuants we would get by individuating the coco-network with chains of *explicit* conditional-coreference, where 'explicit' means something like: involving the *same* naming- or term- convention, and explicitly anaphoric devices (Taylor 2021). For the purposes of this chapter, I will thus adopt the conception laid out by Kaplan (1990); he writes:

> I propose a quite different model [from the type/token morphophonemic conception on which words are individuated by spelling] according to which utterances and

> inscriptions are *stages* of words, which are the *continuants* made up of these interpersonal stages along with some more mysterious *intra*personal stages. I want us to give up the token/type model in favor of a stage/continuant model. (. . .) I think of my conception of a word as a naturalistic conception. Because the interpersonal transmission of words is so central to my conception, I adopt a phrase of Kripke's, and I call my notion the *Common Currency* conception of a word. (Kaplan 1990: 98)

I will consider labelled-files continuants whose links correspond to (roughly) explicit coreference. Every such network corresponds in principle to a *common currency* linguistic expression. (I don't mean that these expressions cannot undergo a gradual shift in their semantic or syntactic properties). These networks are the relevant inputs for semantic interpretation as far as 'community-wide meanings' are concerned. That being said, as will be clear in due course, I will be flexible and oscillate between a (non-orthodox, non-morphophonemic) type/token conception, and a Kaplan-like conception, for the sake of discussion.

1.4 Chapter plan

In section 2, I start investigating the coordinating effect of words by discussing Fiengo and May 2006. I extrapolate the following claim from their view: thinkers samethink by sharing metalinguistic beliefs. I disagree with F&M that we can reduce MOPs to linguistic expressions—except, I suggest, for purely deferential concepts. Importantly, the Paderewski cases suggest that linguistic MOPs (and the *de lingua* beliefs that individuate them) are no more shareable than non-linguistic MOPs. Borrowing from Fine (2007), I provide a 'relationist' twist on F&M's *same-expression* relation across idiolects, in terms of an intransitive *same-use* relation.

The fact remains that speakers *take themselves* to use the same words. There is a grain of truth in the claim that speakers samethink in virtue of sharing *de lingua* beliefs. Following up, in section 3, I continue to explore the idea of a boostrapping effect, this time at the level of *semantic appearances*. The view I examine says (roughly) that mutual *appearances* of semantic sameness make it the case that things happen as if meanings were shared (Schroeter 2012, Schroeter & Schroeter 2014, 2016). It is hypothesized that networks of such mutual appearances constitute representational traditions, which we can take as *inputs* at the metasemantic level, so as to project a "community-wide" meaning onto them, in a way that I explain. Here I will depart from Schroeter, and take an instrumentalist stance with respect to such 'post-hoc' holistic constructs. I essentially endorse this bootstrapping claim, but emphasize that we don't have to construe undefeated mutual appearances of semantic sameness in terms of meaning sameness (i.e. in a Fregean way).

In section 4, I then revisit the notion of the division of linguistic labor in light of the representational traditions that mutual semantic appearances give rise. I emphasize, after Putnam 1975, the epistemic dimension of the norms of word use. Following Recanati 2016, I propose

that what underlies the division of linguistic labor at the meta-semantic level is a peer-to-peer distributed file managed at the community level. I give a sense of the kind of social mechanisms by which a community manages a distributed file. In closing, I say how we can make sense of samethinking without causal link in the present framework.

2 Sharing words (Fiengo & May 2006)

In this section, I examine the view that words play the role of MOPs, and that we samethink by sharing (beliefs about) words. The view that linguistic expressions play the role of MOPs has actually been endorsed (and then rejected) by Frege. I briefly present the path that goes from Frege's endorsement of the metalinguistic view, to its rejection for a conception in terms of *descriptive senses*.[4]

2.1 Metalinguistic semantics of identity statements

In the *Begriffsschrift*, Frege originally endorsed a metalinguistic account of identity statements, according to which what plays the role of a MOP in this kind of linguistic context is a *linguistic expression*. When you assert (resp. interpret) something of the form "*a* is *b*", what you really express (resp. interpret), is the judgement that the symbol *a* and the symbol *b* have the same content. In the *Begriffsschrift* notation:

$$\vdash (a \equiv b)$$

Here is a relevant quote from the *Begriffsschrift*:

> Whilst elsewhere symbols simply represent their contents, so that each combination into which they enter merely express a relation between their contents, they at once stand for themselves as soon as they are combined by the symbol for identity of content [(≡)]; for this signifies the circumstances[5] that two names have the same content. (Frege 1879/1997, p.64 of the Beaney reader)

Frege explains the rationale for this metalinguistic view in the following passage:

> Equality gives rise to challenging questions which are not altogether easy to answer. Is it a relation? A relation between objects, or between names or signs of objects? In my *Begriffsschrift* I assumed the latter. The reasons which seem to favour this are the following: $a = a$ and $a = b$ are obviously statements of differing cognitive value; $a = a$ holds *a priori* and, according to Kant, is to be labelled analytic, while statements of the form $a = b$ often contain very valuable extensions of our knowledge and cannot always be established *a priori*. (. . .) Now if we were to regard equality as a relation

[4] I set aside a number of exegetical problems, which are not mine here. See e.g. Perry 2019 chap3-4 for a more detailed historical and conceptual overview.

[5] At the time, Frege thought that the semantic value of a sentence was a situation (a structured entity) as opposed to a truth-value.

> between that which the names '*a*' and '*b*' designate, it would seem that *a* = *b* could not differ from *a* = *a*, i.e provided *a* = *b* is true. A relation would thereby be expressed of a thing to itself, and indeed one in which each thing stands to itself and to no other thing. What we apparently want to state by *a* = *b* is that the signs or names '*a*' and '*b*' designate the same thing; a relation between them would be asserted. (Frege 1892/1997; p. 151 of the Beaney reader)

In this passage, Frege argues that construing identity statements as being about objects cannot explain that some identity statements are informative; whereas the metalinguistic account explains this. As the quote indicates, Frege abandoned this metalinguistic view of identity statements at the time he wrote the text displayed in the quote (from *On sense and reference* (1892)). On the new conception Frege (1892) puts forward, what plays the role of MOPs are not linguistic expressions (i.e. syntactic entities), but rather, semantic entities finer-grained than reference (or *senses*—Frege thought of them as descriptions). Here is the passage where Frege explains his change of mind:

> But this relation [(≡)] would hold between the names or signs only in so far as they named or designated something. It would be mediated by the connection of each of the two signs with the same designated thing. But this is arbitrary. Nobody can be forbidden to use any arbitrarily producible event or object as a sign for something. In that case the sentence *a* = *b* would no longer be concerned with the subject matter, but only with its mode of designation; we would express no real knowledge by its means. But in many cases [expressing real knowledge] is just what we want to do. If the sign '*a*' is distinguished from the sign '*b*' only as an object (here, by means of its shape), not as a sign (i.e. not by the *manner* in which it designates something), the cognitive value of *a* = *a* becomes essentially equal to that of *a* = *b*, provided *a* = *b* is true. A difference can arise only if the difference between the signs corresponds to a difference in the mode of presentation of the thing designated. (Frege 1892, p. 151-2 of the Beaney reader, trad. slightly modified; emphasis mine).

Frege puts forward two worries for the metalinguistic view of identity statements. First, if we construe identity statements in a metalinguistic way, then we make the relation between linguistic expressions and their content non-arbitrary, because a linguistic expression would be required to have the same content as the other linguistic expression featuring in the identity statement. But, as Frege points out, we are free to stipulate that an arbitrarily chosen expression be used to refer to something. The arbitrariness of linguistic conventions is negated by the metalinguistic view of identity statements. Second, if we construe identity statements in a metalinguistic way, then we cannot account for the extension of real knowledge that such statements often are meant to provide. According to Frege, identity statements construed as statements about signs become trivial, because they do not express non-linguistic knowledge. Given these two worries, the metalinguistic account must be wrong. MOPs should not be construed as linguistic expressions, but rather as *ways of thinking* about something (which *may*

be conventionally associated with linguistic expressions).

Frege's point about the arbitrariness of the relation between a sign and the thing designated by it, is arguably ambiguous. On one reading it is true, on another reading (the relevant one, as I argue), it is false. It is true that signs are abirtrary in the sense that there are many possible worlds in which we have different words, or in which our actual words are paired with different semantic values. But, *given* a natural language used by a population at a certain time, it is not true that "nobody can be forbidden to use any arbitrarily producible event or object as a sign for something."[6] On the contrary, any language user is embedded in representational traditions, and the way she uses words is interconnected with, and not independent from, the way other language users of her community use these words. As a result, there is an important sense in which the pairing of a sign with a semantic value is *not* arbitrary (given a shared language). Given a shared language, it is not the case that one can use any sign to refer to anything one wants. So the claim that identity statements cannot be about signs because this would make signs non-arbitrary is not a good argument against the metalinguistic view.[7]

I turn to Frege's point about triviality, which is not fully clear to me.[8] This, at any rate, is clear: even if we accept that a purely metalinguistic interpretation of identity statements makes them unable to impart non-linguistic knowledge, this does not imply that identity statements are not metalinguistic *at all*: they could still be *partly* metalinguistic (an option I explore in the next section).

There are other problems with the view that equates MOPs with words. Schiffer 1990 mentions a few of them. For example, the view fails to apply to non-linguistic thinkers (such as infants). Moreover, the view is forced to appeal to propositional attitudes such as implicit beliefs about words, hence there is a threat of circularity in the account (aren't words grasped under MOPs? If not, why not?).[9] Thirdly, we have the Paderewski cases, in which a thinker associates *two* MOPs with only one public word. So MOPs cannot be reduced to public words. But if we take the relevant notion of word to be *idiolectal* instead, then words are not shared, and it seems that Fregean senses would turn out to be unshareable.

The first part of this chapter aims at critically reviving something like the old metalinguistic view

[6]The notion of a natural language as used by a population at a time may be extremely vague. I don't think this is a reason to deny its existence (*pace* Chomsky — see chapter 2 of Chomsky 1986; 1987, 2000). This chapter tries to give a sense to the notion of a public language in non-mysterious terms. I think we should understand idiolects also in terms of how they are explicitly connected to each other by deferential relations, something I try to make sense of here.

[7]To be fair with Frege, he certainly had in mind *formal* languages, with respect to which his point sounds more reasonable.

[8]Does my reader share the impression that Frege seems to contradict himself from one quote to the other concerning the capacity of the metalinguistic account/the objectual account (resp.) to account for informativeness?

[9]I echo this objection against Fiengo & May when I remark that, even if idiolectal words are transparent, words *qua* common currency words are not.

abandoned by Frege, however in a different theoretical context. Crucially, the metalinguistic view I will examine does not individuate syntactic expressions by means of their shape, but works with a richer conception of syntax, such that if two occurrences are of the same syntactic type, they must corefer.

Why bother discussing the metalinguistic view, if it looks so desperate? The first reason is that even if the metalinguistic view is false, it could still be true that speakers coordinate in part in virtue of the fact that they presuppose that they are using the same words. How much of the metalinguistic theory is needed to make this claim is an interesting question. The second reason is that the metalinguistic view, even if false as a general theory of MOPs, could still be true for a certain kind of MOPs. In fact, I will suggest that words do play the role of *purely deferential concepts*.

2.2 Coordination as recurrence of expression-type

Fiengo & May claim that coordination is a matter of recurrence of the same expression-type. An expression-type, for them, is an item in the syntactic representation of a sentence. They intend this characterization to be fully general, that is, even with respect to *anaphora*. Accordingly, they have a somewhat unorthodox conception of expression-type individuation. They use *indices* (=numerical subscripts) to represent type-identity and type-difference of linguistic expressions. Two tokens are *co-indexed* iff they share an expression-type. The notion of expression-type involved is to some extent independent from morphophonemic identity. For example, a name and a pronoun may count as the same expression, in their framework. Consider:

Bob$_1$ reads. He$_1$ is very focused.

'Bob' and 'He' as they feature in the discourse above are two distinct realizations of the same expression, on F&M's framework (a counter-intuitive claim). These occurrences are grammatically required to corefer, if 'Bob' refers. F&M's notion of an expression-type is thus very rich: it has to do with governance relations, anaphoric dependency relations, and coreference profiles (Fiengo & May 1994). Relatedly, they distinguish *names*, and *expression-types*. A name-type is a lexical item individuated by spelling and pronunciation. As such, it does not refer. For example, take the word forms "Michaël" and "Michael" individuated by spelling alone. Is Michael the same person as Michaël? This question is empty. A name individuated by spelling alone is only a word form, without any reference. Only a lexical item used in discourse may refer. This is what F&M call *an expression-type*, which can have multiple occurrences in a discourse.[10] As they say, an expression-type may *contain* a name. What matters for coordination is recurrence of the same *expression*, not recurrence of the same *name*. For example, distinct 'Aristotle'-expressions are not coreferential (one may refer to the philosopher, the other to the ship magnate), even though the same name-type is repeated

[10] F&M's construal of "expression type" is closely related to Strawson's (1950) concept of "use", as Ostertag 2007 notes.

twice. In many cases, reapeating the *same* name actually indicates that *different* expressions are involved. Compare (1a) and (1b):

(1a) Bob_1 tells Bob_2's sister that it is raining.
(1b) Bob_1 tells his_1 sister that it is raining.

In (1a), the repeated use of the name-type 'Bob' signals that two different Bobs are introduced in the discourse: the occurrences anti-corefer (Pinillos 2020). If a speaker understands (1a), she knows that the two occurrences refer to distinct objects if they refer at all. By constrast, unless the context prescribes otherwise, the correct interpretation of (1b) makes 'his' coindexed with 'Bob'. An agent who understands the discourse knows that the two occurrences refer to the same thing if they refer at all.

2.3 *De lingua* beliefs

An assignment states the pairing of an expression-type with a semantic value. We may think of assignment as a *function* whose domain is the set of expression-types in a given lexicon, and that pairs each expression-type with its unique semantic value.[11] Assignments can be represented with sentences of the following form:

⌜"$[_{\text{NP}}X]$" has the semantic value NP⌝

Whenever a speaker uses an expression referentially, they believe an assignment about that expression. In an assignment, the expression inside quotes is mentioned, and the same expression featuring to the right end side is used. Assignments are thus metalinguistic statements in which the truth-conditions of elements of the speaker's idiolect are laid out (Higginbotham 2006). As F&M put it,

> **The Assignment principle:**
> To be sincere, if a speaker uses a sentence containing an occurrence of the expression NP, the speaker believes an NP-assignment.

Such beliefs are typically tacit. In particular, agents need not have the *personal-level* concepts ASSIGNMENT or SEMANTIC VALUE in order to use referential expressions. These concepts only need to be operative in agents' *grammar*. However, such beliefs guide speakers' referential behaviour: they characterize stable dispositions of agents to use certain expressions to think or speak about certain objects.[12] Two assignments are the same just in case they ascribe the same semantic value to the same expression-type. So again, on F&M, strictly speaking, there is no *nambiguity* (to use Perry's phrase)—each expression-type has its own assignment, and assignment is a *function*.[13] It is not the case that a name is ambiguous in the sense that it has

[11] Note that this characterization makes assignments unshareable, because the set of departure of the assignment function is limited to an idiolect. More on this shortly.

[12] You can think of Assignments roughly as resembling to the subpersonal algorithms Cumming was talking about in terms of communicative strategies (except that Cummingian strategies are not defined in semantic terms).

[13] On the *nambiguity* view, the identity of a name is merely a matter of morphophonemic form, and names are massively ambiguous. Perry 2012, Kamp 2022 are two proponents of the nambiguity view. It is not clear to me that the debate is not verbal.

more than one referent: a name does not refer, because it is merely a linguistic type—a name has bearers. Still, I observe that whatever one's conception of name individuation, we cannot take lack of ambiguity in *name interpretation* as a normal condition. Names *are* ambiguous in discourse, and hearers often need to disambiguate.

On F&M's framework, the identity and difference of expressions tokened by a speaker are transparent for that speaker (F&M: 53).[14] The idea is that an agent is an authority of their idiolect:[15]

> **Expression identity is transparent:**
> Two occurrences are occurrences of the same word for a speaker iff that speaker believes that they are occurrences of the same word.

Moreover, F&M endorse the following principle:

> **Singularity principle:**
> If cospelled expressions are covalued, then they are coindexed.
> Given transparency, equivalently: speakers believe that cospelled expressions corefer if coindexed, and that they do not corefer if not coindexed.

Singularity implies that a speaker can have e.g. "[Paderewski$_1$]" and "[Paderewski$_2$]" in her idiolect only if she believes that the respective expressions are not co-valued. The Singularity principle ensures that syntax and semantics work in tandem.[16] The Singularity principle looks *prima facie* too stringent. Consider Kripke's Peter. An enlighted speaker (i.e. aware that Peter is identity-confused about Paderewski) might report Peter's beliefs about Paderewski in a *de dicto* manner, thus deploying more than one 'Paderewski'-expressions in order to mimic Peter's idiolect. For the speaker, these various expressions will be covalued (the speaker is not identity-confused about Paderewski). However, they won't be co-indexed in a *de dicto* report, if they are meant to align with Peter's idiolect. So, the use of expressions of this sort seems to go against the Singularity principle. I explain how F&M accommodate this kind of language use below.

Beliefs about assignments are only one sort of *de lingua* beliefs. Another fundamental sort of *de lingua* beliefs are beliefs about translation between non-coindexed expressions. Translation

[14] When I do not specify the reference, I refer to Fiengo & May 2006.

[15] For a similar view, see Fine 2007, who writes:

> Syntax is transparent, even if semantics is not; and one's take on the expressions of the language should always be presumed to be the same, even if one's take on their referents is not. (Fine 2007: 108-109)

Another view is to say that co-indexing is not transparent even in the intrapersonal domain: words are objects one can be identity-confused about, just like with ordinary object. Such a view would be at odd with a traditional understanding of linguistic knowledge (e.g. Chomsky 1995). But it is open to a radical externalist. See e.g. Richard (1990: 181-182). See Pinillos (*ms)* for discussion on this issue.

[16] The model seems to be standard predicate logic.

statements between noncoindexed expressions amounts to stating that two assignments are *equivalent*, in that they assign the same semantic value to different expressions (F&M:63).

Distinct expressions are thus never co-indexed (hence never corefer *de jure*), but they can be assumed to *translate* each other.[17] (If two expressions are coindexed, they are grammatically guaranteed to be translations of each other). When people learn *identities*, they form beliefs about translation between noncoindexed expressions, such as:

⌜[$_1$ Cicero]⌝ translates ⌜[$_2$ Tully]⌝

This is, you will recall, roughly the metalinguistic view that Frege endorsed and then rejected.

2.4 Sharing assignments

2.4.1 Co-indexing across idiolects

As I have pointed out above, the expression-types that constitute the set of departure of the assignment function are idiolectal. On the face of it, this makes assignments *unshareable* if idiolects are pairwise disjoint. However, Fiengo & May want assignments not to be consigned to the intrapersonal domain:

> Recognizing syntactic identity, and thus distinguishing Assignments, is a capacity that speakers pervasively deploy throughout the back-and-forth of conversations. So, in a model of conversation, we would assume that if a hearer properly understands what a speaker says, he or she will come to represent the sentences the speaker utters as the speaker does. If there is such a formal match in the representations of production and perception, the speaker and hearer will be on the same conversational wavelength, *since across their representations the expressions of names will be coindexed*, and hence coreferential. In turn, the hearer may use the expressions in question in other sentences that he or she utters, *and by doing so a chain of coreference will be carried on*. (Fiengo & May 2006: 23, emphasis mine)

Let me unpack several ideas expressed in this passage, which are relevant to co-indexing across idiolects and across discourses. One idea is that speakers can share assignments:

> **(1) Shareability of assignments:**
> Speakers can share assignments. Two speakers share an assignment iff they assign the same semantic value to the *same* expression-type. So in particular, speakers can share expression-types.

Another idea is that *understanding* involves identifying the same syntactic expressions. They say: "if a hearer properly understands what a speaker says, he or she will come to represent the sentences the speaker utters as the speaker does." Accordingly:

[17] See the general introduction for an informal definition.

> **(2) Understanding requires co-identification of expressions:**
> Utterance understanding requires that the speaker and hearer represent the same syntactic expressions. In other words, a hearer understands an utterance only if the expressions are co-indexed across the representations of the speaker and hearer.

Finally, another idea I want to stress is that co-indexing of expressions can happen across distinct discourses, giving rise to chains of *de jure* coreference:

> **(3) Co-indexing across distinct discourses:**
> Expressions can be co-indexed across participants and across distinct discourses. Inter-discourse co-indexing gives rise to chains of *de jure* coreference.

I take it that the principles (1)-(2)-(3) are all very intuitive.

Regarding (1), a paradigmatic case of a shared assignment seems to be when a speaker learns a common-currency name from someone else. I use 'common currency name' to designate a distributed expression-type, as opposed to what F&M call 'name', that is, a *generic* name that does not refer but has bearers.[18] Let us call situations of this sort, *direct derivation*:[19]

> **(4) Direct derivation and shared assignment:**
> When a speaker A learns through an utterance u a common currency name "[N]" from a speaker B, then A and B come to share the expression-type "[N]" and the assignment: ⌜"[N]" has the semantic value N⌝

For example, a competent speaker who does not possess the common currency name "[Magnus Carlsen]" asks a friend who the current world chess champion is. The friend says: "His name is *Magnus Carlsen*. His father taught him to play at age 5. He drew Kasparov at 13." In such circumstances, the recipient will learn the name in question. It is natural to think that the participants come to share the same name "[Magnus Carlsen]".

The idea expressed with principle (2) looks very natural as well. In communication, we as speakers expect the hearer to identify the words we utter. We judge that the hearer has misunderstood if they failed to do so. For example, in the case of a name-involving utterance, it seems that the hearer will understand only if she correctly identifies which name is used. In the parlance of F&M, understanding requires that the expression in the interpretation of the hearer and in the thought of the speaker must share a type, that is, be co-indexed. This is principle (2).

But principle (3) looks natural as well. If an expression-type can be transmitted from one idiolect to another, and can be used by speakers in multiple contexts, then it seems that the relation of co-indexing can *span* distinct discourses. But given this idea, there is a question.

[18] The label 'common currency name' comes from Kaplan 1990, who argues against the type/token distinction, and wants to replace it with a stage/continuant model. Following him, we may construe 'shareable expression-types' as *continuants*. This will be clearer below, as I argue against the shareability of such expressions.

[19] The label 'direct derivation' is due to Fine 2007: 107.

Does the requirement that speaker and hearer represent the same expressions apply across distinct discourse contexts? Or is it an intra-discourse principle? I come back to this question below.

Metalinguistic view of samethinking To recap, the outlined view of samethinking is as follows. Speakers share MOPs by sharing assignments. Speakers successfully communicate only if the hearer and speaker represent the same sentences, and share every assignments believed in the discourse context. Moreover, as we shall see, speakers successfully report the attitudes of others in a *de dicto* manner only if they ascribe the assignments that are involved in the reported attitudes.

As we saw, Fiengo & May propose that *idiolectal* expressions are individuated in terms of the *de lingua* beliefs of a speaker. In the intrapersonal domain, expressions are coindexed if and only if the speaker believes that the expressions are coindexed. Coindexing —identity of idiolectal expression— is transparent. But how are *shared* expression-types individuated? It is not clear that the same story can be told in both cases. (As far as I can tell, unfortunately F&M do not really elaborate on this issue.)

One idea is to say that two expressions are co-indexed across idiolects in a discourse context just in case the participants believe that the expressions are coindexed. But, this is too weak, because participants can be *wrong* as to whether they represent the same expressions in their representation of the discourse (and coreference between the hearer's interpretation and the thought expressed by the speaker, may fail for this reason). Let us say that two expressions from different idiolects are *subjectively coindexed* just in case participants believe that the expressions in their respective representations of the discourse are co-indexed.

So we need an *objective* notion of indexing of expressions across idiolects, such that participants can be *wrong* as to whether they represent the same expressions. The notion that expressions can be co-indexed across the representations of different speakers is reminiscent of Prosser (2019)'s notion of *transparent communication*, which Prosser opposes to *interpretive* communication. In F&M's words, communication is *transparent* just in case speakers share an assignment with respect to the relevant expression-type *and they presuppose that they do*. In other terms, transparent communication occurs when participants trade on the coreference of their uses without using a translation statement between noncoindexed expressions. By contrast, communication is *interpretive* when a translation statement between noncoindexed expression is involved. The parallel with Prosser goes further: Prosser, just like F&M, holds that the presupposition of coreference is infaillible when it is linguistically determined. When speech participants do share an assignment, coreference is guaranteed as a matter of linguistic necessity. Of course, speakers can be wrong about whether they both represent the same word, and coreference may fail for this reason, *inter alia*. This implies that the putative relation of cross-idiolectal

"co-indexing" is at best *weak coreference de jure* (Recanati 2016, chap. 2 and 8). I shall explain.

I take it —as a defeasible hypothesis— that it is necessary for two expressions across distinct idiolects to be co-indexed, that the participants *believe* (tacit sense) that the expressions in their respective representations of the discourse are co-indexed. The idea is that just as expression-types are individuated by *de lingua* beliefs in the intrapersonal domain, they are (in part) individuated by mutual *de lingua* beliefs in the interpersonal domain. To be objectively coindexed, two expressions from different idiolects must be subjectively coindexed. (Again, "believe" should be construed in a tacit sense; Prosser would say "presuppose". Nothing important hinges on this here for present purposes.) Let us say that two expressions from different idiolects are *objectively coindexed* in a discourse context just in case (i) participants believe that the expressions in their respective representations of the discourse are co-indexed and (ii) the expressions are co-indexed. Note that, in virtue of component (i), the relation of coreference *de jure* which underlies objective co-indexing (as defined here) is not transitive.[20] For example, it is entirely conceivable that two persons A and B believe that the expression one uses is the expression the other uses (i.e. that the expressions are co-indexed across their representations), and likewise for B and C. It does not follow that A and C mutually believe that their expressions are co-indexed. Perhaps A and C would not recognize one another's tokens of the word in question as tokens of the same expression-type, because they pronounce the word differently. In fact, one can doubt that even component (ii) is a transitive relation. I will come back to this important result in due course, and revise the (alleged) notion of "coindexing' across idiolects accordingly.

Like I said, participants can be wrong as to whether they employ the same expressions in a discourse context. Whereas a word shape is easily recognizable across speakers, the identity and difference of expression-types *from one idiolect to another* may elude agent's awareness in interpersonal discourse context. This is because *indices* are not pronounced. To illustrate, consider this scenario provided by Pinillos 2011:

> Suppose that Pecos and Smith are at a party. Earlier in the evening Smith is found praising his friend John. Pecos listens and *understands everything* that Smith is saying. Later on in the evening, Smith is talking about John again but this time making slanderous remarks. Pecos is also in the audience and like before, *fully understands* what Smith is saying. However, Pecos is perplexed. He can't tell whether the person Smith was referring to with "John" earlier in the evening is the same person he is referring to with "John" now. (Pinillos 2011: 311, emphasis mine)

Pinillos' example is interesting in that Pecos can be said to understand even though it is not the case that he represents the discourse as Smith (the speaker) does, because for Peco it is indeterminate whether distinct occurrences of "John" are co-indexed (not so for Smith). That

[20] So it is at most *weak coreference de jure*. (See the beginning of chapter 4 for a definition). Something Fiengo & May do not countenance, because they of course assume that co-indexing is an equivalence relation. Below I say why I think the alleged "co-indexing" relation across idiolects is best modelled in terms of an intransitive relation.

means that Peco does not interpret all the occurrences of "John" with the same expression. What the example illustrates is that even though sameness of expression-type involves linguistically determined coreference, *identifying* which expression-type is involved in a discourse might be a *pragmatic* business involving not just the grammar but representations of intentions (that is, the theory of mind or social competence). As far as I can tell, Fiengo & May are silent about the pragmatic aspects of their story about intercoordination in communication.

Is "co-indexing" across idiolects transitive? In what follows, I cast doubt on the relation of coindexing across idiolects construed as type identity of expressions, and the related principles about co-indexing across idiolects. I will consider Kriple's Peter, and the common currency name "Paderewski". There are several things to be said with respect to Peter, which might constitute a counter-example to some of the principles above. I adapt the examples from Pinillos (*ms*).

> PADEREWSKI (AGAIN): Suppose Bob possesses a single expression spelled "Paderewski" and knows Paderewski as both the musician and the politician. In a first context, Peter meets Bob at a concert at Carnegie Hall. He forgot to look at the name of the musician he was going to see, and he has never heard of Paderewski before. Peter asks Bob: "Who is that guy playing the piano?". Bob says: "His name is *Paderewski*".
>
> At some later time (in a second context), Peter and Bob meet under the Washington Arch in Washington Square to hear a political speech. Peter asks Bob: "Who is the orator?" Bob answers: "That is Paderewski". It does not strike Peter that the two guys are one and the same person. In fact, due to their very different activities, Peter believes they are two people sharing a common name.

By the Singularity principle, Peter has *two* expressions cospelled "Paderewski" in his idiolect. Does he share some of them with Bob? Why? How does this bear on the status of the communicative exchanges between Peter and Bob?

One certainly does not want to deny that, at least until the second context takes place, Peter comes to share "[Paderewski]" with Bob. Moreover, in the second context, there is a sense in which Peter misunderstands what Bob said: Peter failed to recognize a word that the speaker used (a word that he shares with the speaker). In fact, Peter takes himself to be introduced to a *new* common currency name. In a *non-normative* sense of 'direct derivation', Peter does *directly derive* a new expression from Bob's utterance, even though it is not the case that he learns a new name from Bob (*normative* sense of 'direct derivation').[21]

[21] I say more on this normative sense attached to common currency words when discussing Schroeter 2012 and the division of the linguistic labour.

This option does not seem *fully* satisfying. If we restrict to the second context, there is no reason to say that Peter does not understand.

Problematically, the example is not purely linguistic. Both participants, in both contexts, deploy a composite MOP to think about the referent: they can see the guy under discussion. In the second context, Peter fails to recognize the guy he saw at the concert. This failure is not linguistic. Perhaps the non-linguistic component is a distraction here. To remedy this, consider this example, due to Fine 2007:

> PETER—PETER CASE: Peter asserts "Paderewski is musical"; and we, deriving our use of the name from him, may validly infer "Paderewski is musical." But Peter, deriving what he takes to be a new use of the name from us, may then infer "Paderewski is musical." (Fine 2007: 119)

The situation is depicted in Figure 5.2:

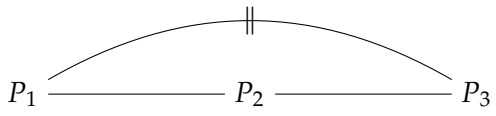


Figure 5.2 – PETER—PETER CASE

In this example, we learn the name "Paderewski" from Peter. But Peter, in turn, derives a new "Paderewski"-expression from us. Again, I believe it is useful to distinguish two senses of "direct derivation". A first sense of "direct derivation" is this. As a result of interpreting our utterance, Peter creates a new use for the lexical item "Paderewski". Consequently, he comes to have two cospelled "Paderewski"-expressions in is idiolect. One gets a second sense of "direct derivation" when one remarks that it is not the case that Peter *learns* a new name from us. Instead, Peter is confused and *believes* (falsely) that he is learning a new name. This is the normative sense of "direct derivation". Fiengo & May could say that Peter fails to understand our utterance, because he fails to recognize the expression that we used – as per principle (2). My take on this is that everything depends on how the discourse context is fleshed out (Fine does not provide any details). But, it is plausible that given some but not all ways to flesh out the context, the case is one in which Peter fails to understand our utterance, because he does not recognize the expression that we used.[22] Still, there is a question as to which (if any) of Peter's two "Padereski"-expressions is co-indexed with the "Paderewski"-expressions of non-confused speakers (on which more shortly).

Here is another example (again adapted from Pinillos (*ms*)) to put to test the principles above:

[22]This hypothesized contextual versatility is in line with the contextualist stance inherent to pragmalignment I have defended in the previous chapters.

BOB–PETER–ANNA CASE: Let us write "[Paderewski$_1$]" the expression that Peter associates with the musician, and "[Paderewski$_2$]" the expression that Peter associates with the politician. Imagine that, in a first context, Bob (who has never heard of Padrewski before) asks Peter: "Who is your favorite public personality?" Peter has in mind Paderewski the musician, and replies "That is [Paderewski$_1$]. He$_1$ is from Poland". Nothing in what Peter says conveys the idea that the guy from Poland is a musician. According to *Direct derivation and shared assignment*, Bob acquires the expression "[Paderewski$_1$]" from Peter.

At some later time (in a second context), Peter is with Anna. Anna has never heard of Paderewski before. She asks Peter: "Who is the most impressive person you saw recently?" Peter has in mind Paderewski the politician, and replies: "It is [Paderewski$_2$]. He$_2$ comes from Poland". Nothing in what Peter says conveys the idea that the guy from Poland is a politician. By *Direct derivation and shared assignment*, Anna acquires the expression "[Paderewski$_2$]" from Peter.

We may suppose that neither Anna nor Bob is confused about Paderewski: their conceptions of the man are very similar, and they each have only one "Paderewski"-expression in their idiolects. Crucially, however, by hypothesis they have *different* "Paderewski"-expression-types. Imagine that, in a third context, Anna and Bob meet for the first time, and have the following conversation:

Anna: Have you heard of Paderewski?

Bob: You mean the Polish guy? Paderewski is an important public figure.

Intuitively, Anna and Bob successfully communicate. However, if it is true that Anna and Bob each inherited a *different* "Paderewski"-expression, then this is a counter-example to the principle (2) above, i.e. *Understanding requires co-identification of expressions*. So either principle (2) or *Direct derivation and shared assignment* must be false.

The problem is that both ideas seem to stand or fall together. I don't have a principled way to reject one rather than the other. As Kamp 2022 says, " in those cases when referring by means of *N* has the effect of introducing the addressee to the given use of *N*, there is no room for misinterpretation and therefore also no risk of it" (Kamp 2022: 22). When an agent directly derives an expression from another speaker (at least in cases where the agent *does* learn a new name), there is no risk of misunderstanding. So I will explore the following line of thought: one option is that *both* ideas are false, because the putative relation of "co-indexing" across idiolects turns out to be *intransitive*.

2.4.2 Intransitive *same-use* relation (≈)

Of course, if two expressions are co-indexed iff they share an expression-type, then it is improper to say that "co-indexing across idiolects is not transitive", because the very notion of co-indexing involves identity. Borrowing from Fine, let us write "≈" the alleged relation of "co-indexing" across idiolects, of which I say it is intransitive. Intuitively, the relation amounts to a (possibly intransitive) *same-use* relation. Accordingly, here is a way to represent the situation in the BOB–PETER–ANNA CASE. I represent expressions and the idiolects they belong with the first letter of the name of the person whose idiolect it is:

- $a_{1,2} \approx b_{1,2}$
- $p_1 \approx b_{1,2}$
- $p_2 \approx a_{1,2}$
- $p_1 \not\approx p_2$

Here is a diagram to illustrate (Figure 5.3):

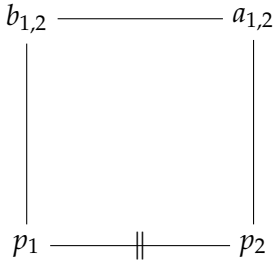


Figure 5.3 – Intransitive *same-use* relation (≈)

I have already suggested above that the alleged relation of "coindexing" across idiolects may in fact fail to be transitive. Following Fine 2007, I will now provide a partial characterization of the relevant *same-use* relation between idiolectal expression-types containing names (what Fine calls "individual use" or "individual name").[23]

Let S be the set of idiolectal expression-types containing names.
Let us call I the set of speakers (or "individuals").
The partition of S into agents' lexicons (the lexicon of a speaker being, for present purposes, the set of all expression-types containing names stored in the memory of that speaker) is made by a function N taking each member i of I into a set S_i (the agent i's lexicon).
Let us write >> the relation of direct derivation. >> (*"directly derives from"*) is a relation between idiolectal expression-types ("expression" for short) meant to capture the transmission of an expression from one individual to another, for example when one individual learns a name from another.

[23] I draw upon Fine 2007, specifically note 9 in Fine 2007: 138-139.

The various elements introduced so far are governed by the following conditions (Fine 2007: 138):

(i) The lexicons S_i are pairwise disjoint;
(ii) For each idiolectal expression $m_i \in S_i$, there is at most one m_j such that $m_i >> m_j$ (any idiolectal expression is directly derived from at most one idiolectal expression);[24]
(iii) It is never the case that $m_1 >> m_2$ for $m_1, m_2 \in S_i$;
(iv) For any chain of direct derivation, there is an expression such that it is not directly derived from any other expression.

Now we can use this characteriation of >> to define the *transmission paths* that link any speaker who has command of a common currency N, as follows:

> **Transmission path**
> Let us say that a sequence of length n ($n \geq 1$), $< m_1, m_2, \ldots, m_n >$ of idiolectal expressions is a transmission path from m_1 to m_n if, for every i ($0 < i < n$), either $m_i >> m_{i+1}$ or else $m_{i+1} >> m_i$.[25]

A *maximal* transmission path is thus a *name-continuant* (Kaplan 1990). Let us now define the property of coherence for transmission paths, as follows:

> **Coherent transmission path**
> A transmission path $< m_1, m_2, \ldots, m_n >$ is coherent if no two idiolectal expressions m_i and m_j on the path, for $1 \leq i < j \leq n$, belong to the same speaker.

We may then characterize the relation of *same-use* (the intransitive notion replacing the notion of co-indexing across idiolects) in the following terms:

> **Same use across idiolects** ($\approx$)**:**
> **(1)** Two idiolectal expressions m_i, m_j are in the *same-use* relation ($m_i \approx m_j$) if there is a *coherent* transmission path from m_i to m_j (possibly of length 1).
> **(2)** It is never the case that $m_1 \approx m_2$ for $m_1, m_2 \in S_i$ and $m_1 \neq m_2$.

This is, roughly, Fine's proposal with respect to the inter-coordination of idiolectal expression-types containing names, as far as I understand it.[26] The proposed characterization of "$\approx$" is

[24] Of course, distinct idiolectal expressions can be directly derived from a single source (i.e. an idiolectal expression may be the source of multiple idiolectal expressions). That is, given m_1, m_2, m_3, it is possible that $m_1 >> m_3$ and $m_2 >> m_3$.

[25] Alternatively, we could define a transmission path in terms of a sequence of *links* $< l_1, \ldots, l_n >$ where each l_i ($0 < i < n$) is a link $< m_{i-1}, m_i >$ such that $m_i >> m_{i-1}$. See Kamp 2022: 23-24 for a related defnition.

[26] I have a hard time evaluating Fine's general proposal, which otherwise looks very interesting—note that I am using the parts of Fine's proposal that I understand, when I modify F&M's coindexing relation across idiolects in order to account for the Paderewski cases. Fine tries to define "the supervenience base from which all questions of *same-use* etc. are to be settled" in terms of a *manifold of names* (Fine 2007: 138 note 9). In mathematics, more specifically in the field of geometry, there is a certain structure called a manifold. It is a structure that makes it possible to differentiate functions defined on the structure. Unfortunately, Fine seems to mess up some notation

incomplete. And there are problems with this characterization. For example, it seems that two speakers at opposite ends of a transmission path may fail to coordinate their idiolectal names if an incoherence occurs somewhere in the middle of the path (hence the "if" in the principle (1)). However, the general solution pattern is hopefully clear enough. Inter-coordination between idiolectal names determines facts about inter-coordination between *token uses* of those names. Principles (1)-(2)-(3) become:

> **(1*) Similar assignments**
> Speakers' idiolectal expression types containing names ('expressions', for short) can be in the *same-use* relation. Two speakers believes a similar assignment iff they assign the same semantic value to expression-types that are in the *same-use* relation.
>
> **(2*) Understanding**
> A hearer understands a name-involving utterance only if the expressions across the representations of speaker and hearer are in the *same-use* relation.
>
> **(3*) Same-use across distinct discourses**
> Idiolectal expressions can be in the *same-use* relation across distinct discourses. The *same-use* relation gives rise to chains of *weak de jure coreference*.
>
> **(4*) Direct derivation and same-use**
> Two expressions from different idiolects are in the *same-use* relation if (i) one speaker's expression is directly derived from the other speaker's expression or vice versa or (ii) there is a *coherent* transmission path from one to the other.[27]

Let us recap. F&M imply that co-indexing is an equivalence relation that can hold across idiolects. Here, I believe the authors are too quick. Paderewski cases suggest that speakers need not represent the *same* expressions for understanding; and that a speaker can directly derive an expression type from another speaker without coming to *share* the same expression type. Accordingly, I proposed to model the alleged "co-indexing" relation between expressions of different idiolects in terms of a non-transitive *same-use* relation, borrowing from Fine (2007). So, in my view, Fiengo & May can afford coordination as recurrence of expression type only within idiolects, but not across idiolects.[28] I now turn to F&M's view about *de dicto* attitude

and forget to define all the concepts. For instance, he first states that N is a function from I to the set $\{N_i : i \in I\}$, but just below, in point 2, N is an element of N_i. Moreover Fine forgets to define what ${}_I N_i$ and what $\cup_i$ is. Additionally, Fine does not really define the relation >>, he just states some of the properties of the relation. Finally, I don't quite understand the notion of a "common use", and the posited coordination links between the "common use" and idiolectal names, see Fine 2007: 109-110.

[27] Note that condition (i) is redundant because each direct derivation constitutes a coherent transmission path of length 1.

[28] F&M *could* afford coordination as recurrence of expression type across idiolects, by incurring further commitments governing the relation. I have in mind a metalinguistic version of alignment. On this construal, it is not the case that Peter share any of his "Paderewski"-expressions with non-confused speakers, because Peter's expressions are misaligned with the "Paderewski"-expressions of non-confused speakers.

reports.

2.5 Ascribing assignments

In communication, the assignments believed are typically not part of what is said. Rather, they are part of the common background information. However, according to F&M, in certain linguistic contexts, assignments come to the foreground and enter into the truth conditions of the utterances. This is the case with (non-trivial) identity statements, and *de dicto* reports, on their view. As F&M understand them, such reports are said *de dicto* precisely because they involve the (covert) attribution of assignments to the reportee. For example, consider the following reports used *de dicto*:

(2) Ivan believes San Sebastian is beautiful
(3) Ivan believes Donostia is beautiful

F&M analyze (2) and (3) used *de dicto* as follows:

(2a) Ivan believes [[San Sebastian$_1$ is beautiful] and ["San Sebastian$_1$" has the value San Sebastian$_1$]]
(3a) Ivan believes [[Donostia$_1$ is beautiful] and ["Donostia$_1$" has the value Donostia$_1$]]

According to (2a), for the report (2) used *de dicto* to be correct, it must be the case that Ivan would agree that San Sebastian is referred to by the "San Sebastian"-expression, and is beautiful. For example, given a context in which Ivan formed the belief that San Sebastian is beautiful on the basis of pictures labelled "San Sebastian", (2) is true. Suppose further that Ivan is also related to San Sebastian under the name 'Donostia' but does not believes that 'Donostia' translates 'San Sebastian', and does not believe that Donostia (under this name) is beautiful. Then given such scenario, the report (3) is false *de dicto* (and true *de re*).

2.6 Making sense of name-involving *de dicto* reports in terms of (≈)

This subsection makes use of the relation (≈) —defined below in remplacement of F&M's co-indexing relation across idiolects— in order to characterize the correctness conditions of name-involving *de dicto* reports. I have already proposed an account of *de dicto* report in terms of pragmalignment, and indexed mental files. The solution proposed here is less general: it is restricted to attitudes involving names, and the reporter & reportee must share the same language. I am extracting this proposal from Fine 2007 (pp.103-104 & p. 113) — I only focus on what Fine calls *strict* de dicto *reading*.

Name-involving *de dicto* attitude reports (≈):
A name-involving attitude report is true *de dicto* only if for each *de dicto* occurrences in the scope of the report,

> **(i)** each of the idiolectal names N_i, N_j used by the reporter ($1 \leq i < j \leq k$) is in the *same-use* relation to the idiolectal names M_i, M_j that the reportee uses to represent the associated part of the attributed content;
>
> **(ii)** $N_i = N_j$ iff M_i=M_j

Condition (i) requires inter-coordination (≈) between each idiolectal names used by the reporter and the corresponding idiolectal names used by the reportee to represent the associated part of the attributed content, where (≈) is defined as above. Condition (ii) requires that the occurrences in the scope of the report be coordinated for the reporter iff they are coordinated for the reportee. To illustrate, consider Kripke's Peter again. Consider the two reports below.

(a) Peter thinks Paderewski is a musician.

(b) Peter thinks Paderewski is a politician.

According to the account formulated here, we may say that there are true *de dicto* readings of report (a) and report (b) above. Since there is no occurrence N_j / M_j to consider in the report (a) or in the report (b), condition (ii) is trivially satisfied in both cases. Likewise, an enlighted speaker can think true *de dicto* readings of the following reports:

(c) Peter believes that Paderewski is musically gifted.

(d) Peter doesn't believe that Paderewski is musically gifted.

By contrast, the report:

(e) Peter thinks Paderewski is a musician and Paderewski is a politician

is only true *de re* if the reporter is not identity-confused in the way Peter is, because condition (ii) won't be satisfied. One problem with this account is that it seems that even an enlighted reporter could *think* a true *de dicto* reading of (e) by mimicking Peter's respective individual uses (i.e. involving two uses such that $N_i \neq N_j$). However, the account cannot explain how this is possible. Relatedly, one may legitimately wonder what it is (on the considered view) that makes the use of an enlighted speaker capable of expressing a *de dicto* reading of the complement clause in (a)–(d): that there is any difference from a *de re* reading of the same reports does not seem in any way captured by the account. Accordingly, I prefer the solution in terms of pragmalignment/indexed files proposed in the previous chapter.

Interestingly, Fiengo & May have the resources to account for the possible true reading of (e) made by an enlighted speaker. This is because on their view, a reporter need not believe the *de lingua* belief she attributes to the reportee; however, she must believe *that the reportee believes it*. Hence we solve the tension between the Singularity principle, and possible employments of expressions in *de dicto* reporting which do not reflect the reporter own *de lingua* beliefs—e.g. an enlighted speaker reporting certain of Peter's beliefs about Paderewski, as in (e). In this

respect, F&M analysis of *de dicto* report is reminiscent of the analysis in terms of *indexed* mental symbols, but with MOPs construed as words. But by no means are the views equivalent. I will now argue that the metalinguistic analysis is wanting.

2.7 Problems with the account in terms of *de lingua* beliefs

2.7.1 Why this logical form?

I start with a *prima facie* problem with F&M's treatment of attitude reports.[29] Consider the report (3) above. The conjonction in the *analysans* (3a) seems to be logically equivalent to the following:

(3b) $\exists x$ ("Donostia" has the value x & Donostia = x & x is beautiful)
which is equivalent to:
(3c) "Donostia" has the value Donostia & Donostia is beautiful.

The problem is that Ivan apparently believes (3c), for he believes of Donostia that it is beautiful (although not under this name), and he believes that "Donostia" refers to Donostia. (Ivan might not be able to provide a definite description to refer to Donostia, and he might not be able to distinguish Donostia from another city, but this is not required to believe the relevant assignment).

F&M intend to block such counterexamples with the following move. They say that (3a) should be distinguished from (3c), because in (3c) the relevant expressions need not be co-indexed (as the absence of numerical subscripts indicates), and so the truth-conditions associated with (3c) allow that John holds two separate independent beliefs, whereas the relevant expressions *are* coindexed in (3a), as a result *one* and only one belief is ascribed (*de dicto*) to Ivan. I think the move does solve the problem, *pace* Ostertag 2007.[30] However, there is a question. Why does an attitude report like (3) used *de dicto* systematically have truth-conditions as represented in (3a), as opposed to something it can convey in *certain* contexts? F&M do not tell us. In fact, I will now argue that it is false to think that reports used *de dicto* always have such partly metalinguistic truth-conditions.

2.7.2 True *de dicto* reports without assignment ascription

It seems that there are cases in which one can make a true *de dicto* report in terms of a given NP-expression even if the reportee is not disposed to believe the relevant NP-assignment. This possibility is straightforward in the case of non-linguistic thinkers. But the problem occurs even for subjects who speak the same language as the reporter, as the following example (due to Saul 1998) illustrates:

[29] I am following Ostertag 2007 here.

[30] One problem with Ostertag's diagnosis is that he takes semi-formal statements like (3b) and (3c) to be equivalent to F&M's statements of truth-conditions for *de dicto* reports. But such paraphrases are not equivalent with F&M's proposed truth-conditions for *de dicto* reports: co-indexing patterns are not respected.

> NICOLE: Nicole has met Superman at a party. Nicole does not read newspapers, or watch TV, or talk to very many people, and she has never heard about the hero Superman. In fact, she is rather puzzled by the outfit worn by the man that she meets. But she finds him witty and urbane. Unfortunately, she never learns his name. Oddly, Nicole is also acquainted with him under the name 'Clark Kent.' Clark has the office next to hers at the newspaper. She has seen him at work, in dull attire, behaving like a shy, harried reporter. She is not particularly impressed by his social skills. Importantly, Nicole does not make a connection between the man next door and the man at the party - in fact, she believes the man at the party to be much more interesting than Clark next door. Rebecca might report the events of the party with sentence (12):
>
> **(12)** Nicole believes that Superman is witty and urbane.
>
> (Saul 1998: 276; mentioned in Ostertag 2007)

Consequently, F&M's analysis of the truth-conditions for the *de dicto* report, namely

> **(4)** Nicole believes [[Superman$_1$ is witty and urbane] and ["Superman$_1$" has the value Superman$_1$]]

is not a good way to state the truth-conditions of the report (12) used *de dicto*. F&M's treatment of *de dicto* reports cannot explain why (12) is true, because Nicole does not have the word 'Superman' in her idiolect. Since one can make a true *de dicto* report without ascribing an assignment about the linguistic expression used in the scope of the report, it is not true that the truth-conditions of *de dicto* reports always include the *de lingua* component. F&M face similar problems with respect to reports about non-linguistic thinkers, or reports in a language that is not shared by the reportee.[31] A related objection may be phrased in the 'material' mode: it seems that MOPs are simply not reducible to linguistic expressions (a point Frege already made in the quote above). Phrased in the framework I am assuming in this thesis, we *can* have mental files that are not labelled with names, including mental files of persons. In such cases, one usually invokes a descriptive entry to access the relevant information stored in memory. Hence at least some MOPs are not linguistic expressions.[32]

2.7.3 Non-linguistic MOP needed

Here I argue that even believing an assignment may involve a non-linguistic MOP. To believe an assignment, a thinker not only needs to represent a word: she needs to represent an

[31] See the sententialist view in Higginbotham 2006, who avoids the problem by stipulating that believing a sentence is to have a belief wich has the same content as the sentence. See also Richard 1990 for the view that the complement-clause of an English report involves a proposition enriched with English names (these enriched propositions he calls "public Russellian Annotated Matrices"/"sentential" propositions). Crucially, the reportee need not understand the English vocabulary enriching the proposition expressed by the complement on this account. Instead, the report puts constraints on a "correlation function" taking us from the public RAM to the *private* RAM that is the way the reportee entertains the proposition.

[32] Relatedly, thinkers do not need to share a linguistic expression in order to inter-coordinate on a topic: they might do with an *ad hoc* signal, or with some non-linguistic form of communication.

object, hence arguably needs to have a MOP to think about the object. Let me explain. An assignment is a metalinguistic statement about an expression-type. It involves the *mention* of the expression-type it is about, and attributes a semantic value to that expression-type. The semantic value is expressed by a referential *use* of that expression-type (referring to whatever is its referent). Hence an assignment involves not only a word, but pairs it with the object that is the referent of that word. In order to believe an assignment then, thinkers must have a concept for the relevant object. *This* concept cannot be the expression-type which is paired with the object.

A similar reaction to F&M's view is expressed in Recanati 2016, who writes:

> We don't have to treat an anaphoric pronoun as the same expression as its antecedent to acknowledge that *recurrence* is what ultimately grounds coreference *de jure*. What recurs, arguably, is not (or not necessarily) a *linguistic* representation but a *mental* representation. (Recanati 2016: 9)

Recanati here suggests that what coindexing reflects is not the recurrence of a linguistic expression, but the recurrence of a (possibly non-linguistic) mental representation. Expressions may serve as *labels* for MOPs, but they are not themselves MOPs (except for purely deferential MOPs, as I will tentatively suggest).[33]

In the mental file framework, assignments are represented by metalinguistic entries of the form "is called *NP*". Expression-types (in the sense of Fiengo & May) can be co-indexed across the mental files of different thinkers. Figure 5.4 depicts interpersonal coindexing in the mental file framework.[34] For reasons adduced above, I prefer to say that idiolectal expression-types can stand in the (possibly intransitive) *same-use* relation, as depicted below (each rectangle stands for a mental file):

[33] Expression can be the referent of a mental file, when the mental file is about a word. See Gasparri 2015 for the outline of an account of mental words as lexical files.

[34] It should be noted that there is a notion of coindexing between *files* in the mental file theory. In Recanati 2013b, Recanati introduces a numerical index on files. Two files share a numerical index (are coindexed) just in case they belong to the same sequence of files. We may say that coindexed files participate in the same distributed file (where a distributed file is a sequence of file connected by weak coreference de jure relations). I elaborate on this notion in the second part of this chapter.

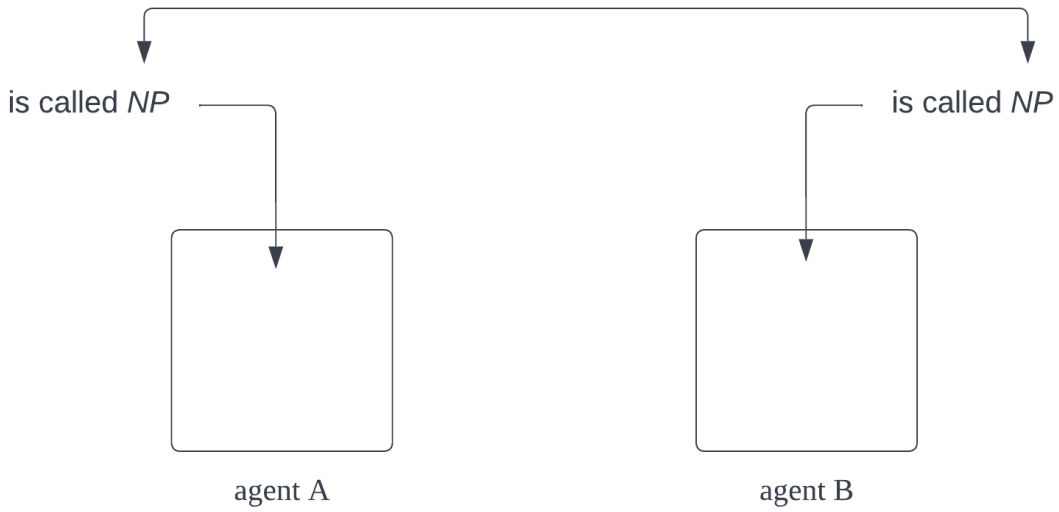


Figure 5.4 – Interpersonal coindexing in the mental file framework

Here is the mental file counterpart of the intransitive (≈) structure in the BOB–PETER–ANNA CASE (see Figure 5.3 above):

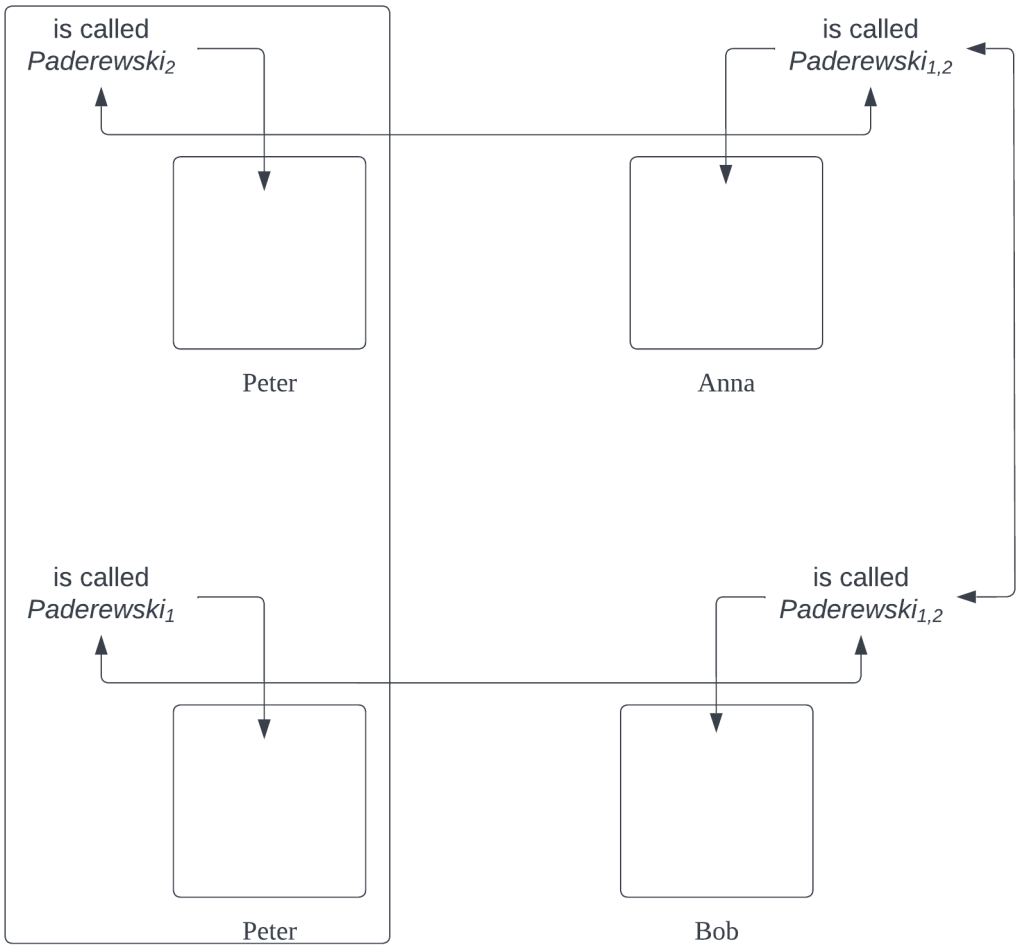


Figure 5.5 – Intransitive (≈) in the MF framework

The last objection against MOPs construed as linguistic expressions (namely, that believing an assignment involves a non-linguistic MOP) can be challenged. To see this, consider *purely deferential concepts*. For example, I overhear the name "Tuvalu" in a conversation. (To flesh out the example, imagine that I overhear an utterance of "I was in Tuvalu on vacation, it's beautiful" said by a passenger on the metro). I understand that it is the name of a place (a place that I don't know). My concept TUVALU is purely deferential. A purely deferential concept arguably *is* nothing but (or not much more than) the representation of a word (Recanati 2001, Millikan 2000).[35] How can an agent represent an assignment with respect to a word whose use she does not understand? Presumably, the assignment pairs a *mention* of the word with whatever is referred to according to the deferee by a *use* of this word (the deferee may be distinct from the person who transmitted the word, see Thuns 2017). When the *use* of a word is purely *deferential*, it arguably involves a *mixed quote* (See Recanati 2001, Shan 2007). So, an

[35] In the case considered, the deferential concept is clearly also the concept of a place.

assignment with respect to such a word will involve a mixed quotation.[36] As a result, in order to believe an assignment for a 'deferential' word, a thinker need not deploy a MOP other than a representation of the word in question.[37] However, not all concepts can be purely deferential, otherwise the thinker could not entertain any content. Hence the objection that assignments involve non-linguistic MOPs is in general valid, and concepts cannot be reduced to words.[38]

I have raised problems for the metalinguistic view of Fiengo & May (2006). A metalinguistic semantics of attitude reports seems inadequate. This reflects the fact that, in general, MOPs are not words (unless for purely deferential MOPs, perhaps). Rather, words may serve as *labels* of the MOPs. Moreover, co-indexing across the representations of different agents has another status than co-indexing in the intra-subjective domain. Speakers can be wrong as to whether they are using the same expression in a discourse context. So, co-indexing of expressions across idiolects is not transparent. Moreover, I have suggested that the Paderewski cases exhibit intransitivity in the interaction between coindexing within idiolects and the alleged relation of coindexing across idiolects. I have proposed to replace co-indexing across idiolect with a non-transitive relation of *same-use*, borrowed from Fine (2007).

From syntax to semantic appearances Even if it turns out that expression-types are, strictly speaking, not shared, speakers believe that they are using the same words, and they enjoy mutual appearances of meaning sameness. They take themselves to mean the same, and to use the same words. Perhaps this is what matters. These *de lingua* beliefs, and these mutual semantic appearances, might help make it the case that things happen as if meanings were shared. That is the line of thought I explore in the next section, moving the explanation of samethinking at the level of semantic phenomenology — in terms of what appears *de jure* the same to subjects. Note that, by moving the account to the realm of semantic appearances, we do not need to commit on controversial principles like shareability, or a particular view of the individuation of words. Instead, we can focus on the appearances speakers enjoy, and describe the representational traditions to which these appearances give rise.

3 Sharing semantic appearances (Schroeter 2012)

What is it that links together different uses of a representation by different agents or at different times into a common practice of using that representation to mean something? Schroeter (2012) has an interesting answer to this question. She says that *mutual appearances of meaning sameness* in speakers is what *prima facie* links different token uses into a shared representational tradition.

[36] It is part of Fiengo & May's view that expressions are both *mentioned and used* in *de dicto* reports or non-trivial identity statements. In that sense, expressions are mixed quotations in those linguistic contexts.

[37] I use '*deferential* word' to mean: word whose use one does not understand but defers to another speaker.

[38] Here, I am obviously not dealing with concepts at the sub-personal level, related to e.g. the functioning of perception, motor control or navigation, with respect to which the remark trivially applies. This is because I am dealing with concepts distributed through linguistic transmission. Samethinking is primarily a relation between such concepts.

These very appearances help make it the case that things happen as if meanings were shared.

In this section, I follow Schroeter in describing the phenomenon of *boostrapping through mutual semantic appearances*. However, my interpretation of this phenomenon is different from hers. In particular, I believe that we need not interpret this phenomenon in terms of meaning *sameness*: on my view, we do not share meanings other than referents or network content (i.e. when the network has no origin). I emphasize this difference in due course. First, I introduce Schroeter's account, which has two layers: one layer is the layer of semantic phenomenology, which may be stabilized and distributed throughout a human population. The other layer is the output of the post-hoc procedure of sorting out the appearances, and which aims at assigning a unique semantic value to the whole network of these appearances.

3.1 Intrasubjective semantic appearances

Schroeter takes as basic each thinker's object-level perspective on their meanings. She thinks the distinction between coreference de jure vs. de facto is best illuminated by the first-person experience on meaning sameness:[39]

> Sameness of meaning requires not just coreference, but a certain sort of subjective appearance of coreference – what I'm calling the appearance of de jure sameness. (Schroeter 2012: 3)

To illustrate, suppose you try to recall what you know about tigers. You think: *tigers have dark vertical stripes*. You also think: *tigers are the largest living cat species*. (You don't need to think these thoughts in those very words). The fact that your thoughts are about the same topic —tigers— is somehow presented as obvious to you, as part of the very contexture of the whole sequence. Moreover, the fact that your thoughts are about the same topic is typically not subject to correction. A corollary of this is that the question "how do you know that your thoughts are about the same topic" would be very hard to answer —compare with the question "how do you know that you are conscious?": it seems as if you just know. I have just informally introduced the three main epistemic features that Schroeter takes to be defining of the appearance of de jure sameness: (1) obviousness; (2) incontrovertibility and (3) epistemic primitiveness. Schroeter defines the appearance of meaning sameness in terms of its epistemic profile understood as the conjunction of (1)-(2)-(3):

> **Appearance of meaning sameness:**
> Two token representations e_1 and e_2 appear de jure coreferential to a given thinker iff it is *obvious*, *incontrovertible* and *epistemically basic* to that thinker that e_1 and e_2 corefer.[40]

[39] In this respect, Schroeter's approach is akin to theories that aim at grounding intentionality in phenomenal consciousness. See Tim Bayne and Michelle Montague 2011 & Bourget and Mendelovici 2019 for an overview, Kriegel 2011 for interesting outlines.

[40] I use 'representation' to mean either a linguistic or a mental representation.

Schroeter's view is a *bootstrapping* view: the very semantic *appearances* help make it the case that things happen as if meanings were shared. Accordingly, Schroeter's intrasubjective criterion of de jure coreference features the experience of de jure sameness, as follows:

> **Intrasubjective meaning sameness:**
> Two token representations are de jure coreferential for a thinker (roughly, to be qualified) iff they appear de jure coreferential to that thinker.

As indicated in parenthesis, this definition should be qualified. The qualification has to do with the fact that appearances of de jure sameness can be misleading. For example, I used to believe that Barry Smith the ontologist was Barry Smith the philosopher of language. All my past encounters with tokens of the form 'Barry Smith' were associated with an experience of *de jure* sameness. Until I realized that they were in fact two different philosophers. So appearances of meaning sameness are fallible – they are at best a reliable indicator of meaning sameness. Because the appearance of de jure sameness is fallible, one cannot define de jure coreference solely in terms the appearance of de jure sameness. The appearance of de jure sameness, albeit usually veridical, is coreference as *presupposed* by the subject—the presupposition may be false.[41]

Let us call the relation of meaning sameness in the intrasubjective domain, R_{intra}. Schroeter suggests that R_{intra} is transitive:

> Ordinary reasoning seems to commit us to the transitivity of *de jure* sameness in thought and talk. For instance, when you rely on standing beliefs about tigers in a stretch of conscious reasoning, you're not just committed to those beliefs pertaining *de jure* to the same topic. You're also implicitly committed to those beliefs pertaining *de jure* to the same topic as the past judgements from which they derive, and to the other past attitudes on which those past judgements were based – even if you no longer remember those attitudes. (Schroeter 2012 note 18)

Schroeter here introduces the notion that thinkers are *committed* to their (possibly forgotten) past attitudes on one topic being de jure coreferential with their current attitudes on that topic. This talk of commitment introduces a distinctive normative element to the picture. This notion of commitment to de jure corefer with a diachronically or socially extended representational tradition is important, and will resurface more explicitly in Schroeter's final stage criterion of meaning sameness in the interpersonal domain. As we shall see, thinkers not only commit to corefer with their past uses, they also commit to corefer with the (presumed) community use.

[41] Contrast with Fine's notion of strict coreference (=coreference required in virtue of one's semantic knowledge), which Fine takes to be factive. Lawlor 2010 objects to Fine's notion on the ground that de jure coreference is arguably not factive; Fine answers in Fine 2010. Liwekise, in Fiengo & May (2006)'s framework, while an agent cannot be wrong as to whether two expressions are co-indexed in her idiolect, she may we wrong about the identity of words *qua* words of a shared language. Of course, speakers can also be wrong with respect to the assignments they believe.

Given this relation of intrasubjective *de jure* sameness (R_{intra}), attitudes that are R_{intra}-connected thus belong to "historically extended bundles of attitudes demarcated by the subjective appearance of de jure sameness" (p.15).[42] [43]

3.2 Intersubjective semantic appearances

Just as we are disposed to treat our own tokens and thoughts as *de jure* coreferential, we are disposed to treat *others'* tokens and thoughts as *de jure* coreferential with our own. Consequently, one may then define meaning sameness across thinkers in terms of the *mutual* appearance of de jure sameness. Interpersonal appearances of meaning sameness typically involve utterances of linguistic expressions. (We often do think in words for ourselves as well, that is, using inner speech. One idea is that we do so in order to commit to the community use, essentially for epistemic reasons. More on this shortly). Importantly, since we are dealing with mutual phenomenology here, we don't have to stick to Fiengo & May's conception of words. Schroeter's first-pass relational criterion may be formulated as follows:

> **Intersubjective meaning sameness (first pass)**:
> Two token expressions e_A and e_B produced by thinker A and thinker B resp. are de jure coreferential (roughly, to be qualified) iff A and B have *the mutual appearance of meaning sameness* with respect to their respective sub-utterances e_A and e_B.

Let us call the relation *__mutually appears to mean the same as__ to X and Y*, (R_{inter}).[44] It may be cashed out as follows:

> **Mutual appearance of meaning sameness (R_{inter}):**
> It appears to A that A's subutterance *de jure* corefers with B's subutterance and it appears to B that B's subutterance *de jure* corefers with A's subutterance.

Intersubjective meaning sameness is defined in purely individualistic terms, namely in terms of what appears *de jure* the same to each individual. Note that the qualification having to do with the fallibility of the appearance of de jure sameness applies, *a fortiori* in the interpersonal domain. The mutual appearance of meaning sameness is at best a very realiable, but not infallible, indicator of meaning sameness. A sufficient amount of contrary evidence can override such appeareances for the participants (the experience is, I take it, familiar: e.g. you thought

[42] Prosser 2020 proposes a similar picture, when he suggests to conceive of mental files as *continuants*. However Prosser adopts a stage view on the persistence of mental files (as opposed to a 'worms' view identifying files with the bundles) and thus calls his view 'a stage theory of mental files'. Recanati endorses this stage/continuant model of mental files in Recanati 2016. As already mentioned, two file stages are co-indexed just in case they are part of the same mental file continuant (a distributed file). *Belonging to the same continuant* should not be confused with *identity*.

[43] In linguistics, there are usage-based theories of language cognition in which networks and the notion that speakers continuously map inputs into equivalence classes, feature prominently. For instance, Bybee and Beckner 2015 proposes that many aspects of language cognition are formed on the basis of experienced tokens by speakers. This experience is somehow cumulative. Bybee explains, using networks, how e.g. quasi-stability in speech conventions emerge over time thanks to this kind of ongoing and dynamic usage-based linguistic representations.

[44] Where X and Y are schematic letters ranging on persons.

you were being told about Bob the cousin, but in fact it was about Bob your neighbor). In some cases, mistaken appearances are not discriminated by the speech participants in the discourse context. Such cases thus demand a specific metasemantic repair, because the mistaken appearances defeat the representational tradition. I say how to deal with such cases in due course.

Another reason why the relation of mutual appreance of meaning sameness cannot be equated with meaning sameness simpliciter is that it fails to be transitive, because the relation requires thinkers to interact. According to a now familar move, to make her criterion non-interactive, Schroeter puts forward the *ancestral relation* of the relation of mutual appearance of *de jure* sameness. Let 's call it (R^{+}_{inter}):

> **Being reachable along a chain of mutual appareances of meaning sameness (R^{+}_{inter}):**
> Two token expressions e_A and e_B stand in R^{+}_{inter}/the ancestral relation of R_{inter} iff there is an ordered set $< e_A, \ldots, e_B >$ such that each member stands in R_{inter} to its successor.

We may then define *de jure* coreference à la Schroeter in terms of R^{+}_{inter}:

> **Intersubjective meaning sameness (second pass)**:
> Two token expressions attached to different speakers *de jure* corefer (roughly, to be qualified) iff they stand in R^{+}_{inter}.

Schroeter provides the following informal glosses on the relation (R^{+}_{inter}):

> Even if two English speakers have never met, there will be chains of apparent *de jure* sameness relations that indirectly link them together. (Schroeter 2012: 16-17)

> The automatic mechanisms for understanding others' speech as *de jure* coreferential with our own have the effect of extending the prima facie unit of interpretation from the individual to the group (p.15)[45]

The second quote may suggest that (R^{+}_{inter}) captures the ($\rightharpoonup$) relation examined in chapter 4, which obtains between two mental symbols belonging to two different thinkers just in case (roughly) the joint communicative dipositions of the thinkers with these symbols is an equilibrium. But this is not quite correct, as I shall explain below. Before I do this, I present Schroeter's rationale (which I endorse) for focusing on semantic appearances to explain samethinking. It is important to understand that Schroeter's focus on semantic appearances is guided by two desiderata she thinks should govern any theory of meaning. I present them in turn.

[45] One can think of these automatic mechanisms roughly as the strategies of expression and construal defined in chapter 4. I qualify this claim in a minute.

3.3 Two reasons to direct the explanation of samethinking in terms of semantic appearances

3.3.1 The accessibility constraint

Schroeter puts forward the following methodological claim: any theory of meaning should individuate meanings in such a way that the semantic appearances competent speakers experience should afford a direct and reliable access to meaning sameness. Said differently, the metasemantics should individuate meanings in such a way that sameness of meaning is easily accessible to language users.

Observe that the accessibility constraint is not a *transparency* constraint. The accessibility constraint does not require that language users' perspective on meaning sameness should be *infallible*. It merely requires that appareances should be *reliable*. This is compatible with meaning sameness not being transparent.

The bootstrapping feature of Schroeter's relational criterion of meaning sameness is designed to accommodate the accessibility constraint. On Schroeter's view, the fact that two token uses appear to mean the same to the participants, helps make it the case that they mean the same. So meaning sameness is accessible, and appearances are reliable. In general, epistemic relational criteria are good candidates for satisfying the accessibility constraint, because they make the epistemology of meaning sameness partly constitutive of meaning sameness (e.g. Dickie & Rattan 2010, Onofri 2018, Prosser 2019). That is not true of non-epistemic relational criteria. Consider Cumming's relational criterion for sameness of content. My token use of 'Hesperus' may appear to mean the same as what you mean with this word, but, on Cumming's criterion, we may still fail to express the same content with this word if our concepts are misaligned. Misalignment is a factor which may prevent agents from sharing content that may easily elude the agents' awareness (if my previous arguments are correct, see chapter 3 section 8). So Cumming's criterion arguably does not meet the accessibility constraint.[46]

Before I move on to the next constraint put forward by Schroeter, I will briefly compare Schroeter's proposal with a neighboring view in the literature.

Comparison with Devitt's causal-historical model Devitt (1981, 2001) proposes that the meaning of a name is (roughly) the causal-historical chain grounded in the bearer of the name:

> MEANINGS AS MODES [= the thesis that the meaning of a word is its property of referring to something in a certain way, its mode of reference] together with EXTERNALISM [=the thesis that some words, including names and natural kind

[46] It should be noted that, on Cumming's (2013b) view, alignment is *not* required for deference, but only for sharing content finer than reference. Now, to be fair with Cumming, what Schroeter is after with her notion of 'meaning sameness' might only be referential coordination.

> terms, refer in virtue of causal relations that are partly external to the head] yield the shocking idea [=the thesis that the meaning of some words, including names and natural kind terms, are causal modes of reference that are partly external to the head]. (...) For example, the mode for 'Mark Twain' is the property of referring by means of causal chains grounded in Mark Twain and involving the sounds, inscriptions, and so on, that constitute the history of the name's use to designate Mark Twain; and the mode for 'Samuel Clemens' is similar but involves the sounds, inscriptions, and so on, of this different name. If another externalist theory is right for a word, its mode will be its property of referring by some other sort of causal relation to the external reality. (Devitt 2001: 476-477)

Devitt's characterization of the meaning of names is different (although not obviously extensionally different) from Schroeter's characterization in terms of semantic appearances. Devitt says that two representations have the same meaning just in case they belong to the same causal-historical network. It is not clear how the pronounciation+spelling (i.e. the word shape) of a word such as a name, is supposed to contribute to the individuation of its meaning on Devitt's view. Is it the case that a causal chain links only tokens of the same linguistic continuant? If that it so, then Devitt's view is similar to Schroeter's (as I interpret her). At any rate, Devitt's definition implies that transparency of meaning is lost, because it takes empirical investigation to determine that different nodes belong to the same causal-historical network (as showed in chapter 3). For example, in Devitt's framework, agents in a Paderewski case cannot tell that they are in fact thinking the same meaning (think of Peter and "Paderewski"). But again, Schroeter's account does not clearly differ with Devitt's on this point: Peter's attitudes and token uses of "Paderewski" all belong to the same network of semantic appearances, whereas Peter is not disposed to have appearances of meaning sameness with respect to two subsets of them.[47]

3.3.2 The flexibility constraint

Different speakers may samethink despite large differences in the *conceptions* they attach to their respective concepts of the object; where a conception may be roughly defined as "a summary description of the extension of the concept" (Rey 1985: 298 cited in Murez 2021). Knowledge of the meaning of a word should be compatible with a very sparse or mistaken conception about the referent. This methodological constraint has been forcefully put forward by proponents of 'direct reference': in general, people can mean what they say even when they have an incorrect or very partial grasp of meaning (see e.g. Kripke's argument from ignorance against descriptivist theories of meaning in (1980: 81-82), Putnam (1975), Burge (1979)). In Schroeter's words:

> An account of [samesaying] must not impose implausibly rigid constraints on competent speakers' substantive understanding of the subject matter as a precondition

[47] Similar remarks apply to the view of Sainsbury & Tye 2012.

> for apparent samesaying. (Schroeter 2012: 9)

Schroeter's phenomenological relational criterion for meaning individuation meets the flexibility constraint. Conceptions do not play a central role in the mutual appearances of meaning sameness. (This is not to say that conceptions never play a role: conceptions *might* play a role e.g. when participants have to disambiguate cospelled word forms. See Gray 2016). So, speakers diverging in their conceptions of the subject matter can still samethink on this account. Williamson nicely expresses the spirit of Schroeter's view in the following quote:

> A complex web of interactions and dependences can hold a linguistic or conceptual practice together even in the absence of a common creed that all participants at all times are required to endorse. This more tolerant form of unity arguably serves our purposes better than would the use of platitudes as entrance examinations for linguistic practices. (Williamson 2007: 125)

Still, not anything goes. Like I said, appareances of meaning sameness are not always reliable. In some cases, speakers experience mistaken appearances, and the representational tradition can be defeated. I now turn to this.

3.4 Defeaters of the representational traditions

3.4.1 The need for similarity

Imagine that Mrs Malaprop tells you:[48]

> (1) Allegories can be dangerous

You apparently understand her utterance. Next time you spot an allegory in a text, you might recall Mrs Malaprop's statement. As it turns out, your presupposition of meaning sameness is wrong. You realize this when Mrs Malaprop goes on to say (Schroeter 2012: 17):

> (2) Allegories live on the bank of the Nile

You reason that Mrs Malaprop is not so crazy as to think that figures of speech live in rivers. Rather, she must have a very idiosyncratic – and divergent – use of the word *allegory*. What she means by this word might be what you would express with "alligator", perhaps. We may describe the situation in terms of Cumming's communicative policies. Let a be the symbol Mrs Malaprop expresses with 'allegories', and b the symbol you deploy to interpret this word. It is not the case that $a \rightharpoonup b$, because it is not the case that your joint strategy is an equilibrium. Clearly, Mrs Malaprop would benefit to modify her communicative strategy. Here it is very

[48] Mrs. Malaprop is a famous character from Richard Brinsley Sheridan's 1775 play *The Rivals*. As Schroeter's example (which I am reusing) illustrates, Mrs. Malaprop frequently misspeaks (to comic effect) by using words which do not have the meaning that she intends but which sound similar to words that do. This character has been famously put to philosophical use by Donald Davidson's article on malaprops, Davidson (1986). Another literary character famous for producing funny and unknowingly witty malapropisms is *Dogberry* in Shakespeare's play *Much Ado About Nothing*, Shakespeare 1599.

clear that Cumming's networks of communicative policies are an idealization. The communicative failure between you and Mrs Malaprop is registered by a missing connection, at the competence level, between your symbol and Mrs Malaprop on the network of communicative policies. Mrs Malaprop's mental symbols are thus isolated on the Cummingian networks. By constrast, Schroeter's ground relation is not so idealized. You *are* related by a relation of mutual appearances of meaning sameness with Mrs Malaprop.[49]

The phenomenological personal-level seemings of meaning sameness which Schroeter emphasizes in her proposed relation of (inter-)coordination, are arguably the phenomenological signature, at the personal level, of some kind of implemented algorithms for tracking conceptual/meaning sameness in thought and in utterance interpretation at the subpersonal level.[50] We may employ a less idealized notion of joint communicative strategy than Cumming's to capture the functional dispositional counterpart (the 'automatic linguistic parsing') of Schroeter's relation ___*mutually appears to mean the same as*___.[51] Let us write it ($\searrow$). Let x and y two mental symbols belonging to two different agents X and Y. Then $\ulcorner x \searrow y \urcorner$ indicates that X is disposed to express x with a word that Y is disposed to interpret as y. Unlike Cumming's ($\rightharpoonup$) relation, there is no requirement that the strategy be an equilibrium. In particular, ($\searrow$) is not reference- or subject matter- preserving. For example, with respect to the Mrs Malaprop example, we can write:

$$a \searrow b$$

because you are *prima facie* disposed to interpret a as b given the strategy of expression Mrs Malaprop attaches to a. We define the counterpart notion for Cumming's COMMUNICATIVE PATH as the transitive closure of ($\searrow$), simply written $(\searrow)^+$. Then, while Mrs Malaprop's lexicon is isolated on the Cummingian networks, it is connected to all the other agents' lexicons by $(\searrow)^+$. I sugget that ($\searrow$) is a better candidate than ($\rightharpoonup$) as the *dispositional counterpart* of Schroeter's relation of mutual appearances of meaning sameness. The ($\searrow$) relation translates the relation of apparent meaning sameness in terms of a dispositional characterization featuring agents' grammars.[52]

[49] In the context of this chapter, I am not criticizing Cumming by saying that his ground relation is idealized. The important point is that we *need* to idealize at one level or another in order to make sense of the representational practices. As we shall see, Schroeter idealizes *post-hoc* at the level of whole networks, where Cumming idealizes *ante rem*, at the competence level.

[50] The same mechanism might be responsible for semantic sameness tracking in utterance interpretation and in inner speech, or it might not.

[51] Imagine that language engineers devise a very sophisticated linguistic agent, call it BOB, capable of coordinating its uses of linguistic token expressions with human users. You can think of BOB as an implemented dialogue system that collaborates with humans. Moreover, I assume that there is nothing it is like to be BOB. In particular, BOB does not experience any appearance of de jure sameness. Intuitively, BOB will be able to samesay with human language users nevertheless, because it is designed to do just that. However, Schroeter is committed to say that BOB cannot samesay with other linguistic agents, because BOB does not experience any appeareance of de jure sameness and consequently, its tokens cannot be R-related to human tokens. So having a functional characterisation of the *appear* de jure *the same as* relation is desirable in order to account for human-machine semantic interactions.

[52] Schroeter seems to have something like ($\searrow$) in mind. Textual evidence:

Besides malaprops, all kind of performance errors can happen that makes mutual semantic appearances inaccurate. For example, the hearer may mishear what the speaker says. Like malapropisms, this kind of defeater will be usually easy to spot. However, some defeaters are less easily discernable. For example, the British use "corn" to apply to any kind of grain, whereas Americans take "corn" to apply to maize (the two do not corefer). A British and an American speech participants might fail to realize the difference in their use in the context of one interaction. Moreover, plausibly, American "corn" originated from the British "corn" so that a chain of apparent de jure sameness connects the two types of use (S&S 2014: 16).

So mutual appearances of meaning sameness do not guarantee meaning sameness, as interactions with Mrs Malaprop testifies. What is lacking, despite the mutual appareance of de jure sameness between you and Mrs Malaprop, is what Schroeter calls a sufficient degree of '*congruence*' in your respective understanding and histories of use:

> Being appropriately connected up in the relevant way [by relations of apparent *de jure* sameness] within a continuous representational tradition is not sufficient for [speakers'use of terms to be] samesaying. (. . .) *There must also be enough congruence* in [speakers'] understanding, environment, and histories of use to warrant a univocal interpretation of their presumed representational practice. (2012: 18, my italics)

3.4.2 Schroeter's final criterion

Accordingly, here is Schroeter's resulting criterion of meaning sameness:

> **meaning sameness (third pass)**: Two token expressions t_A and t_B as used by two different speakers A and B *de jure* corefer iff:
>
> (a) t_A and t_B stand in R^{+}_{inter}
>
> (b) there is enough *congruence* in A and B's understanding, environment and histories of use with respect to t.

Thus the automatic mechanisms for understanding others' speech as de jure coreferential with our own have the effect of extending the prima facie unit of interpretation from the individual to the group (Schroeter 2012 p.15)

The idea then would be to demarcate shared representational practices by tracing out networks of apparent de jure sameness relations within a given linguistic community grounded in *individuals' stable dispositions to understand their acquaintances as samesaying*. Each individual, after all, implicitly intends her own use to coordinate not just with those she has already met, but also with the people that those acquaintances intend to coordinate with, etc. In this way, each individual in a community will be hooked up to the very same network of linked representations that together constitute a communal representational practice. The very same network can then figure as the prima facie unit of interpretation for each member of the community. This approach can yield stable, shared units of interpretation for all members of the community (op cit)

Condition (b) amounts to the requirement that there be no defeaters to the apparent *de jure* coreference, and says that the source of any possible defeater lies in the lack of congruence in speakers's understanding, environment or histories of use. With this extra-constraint, Schroeter's criterion is reference- and -subject-matter- preserving. For short:

> Two token expressions t_A and t_B *de jure* corefer iff they are connected by a non-defeated path of mutual appearances of meaning sameness.

The above criterion tells us how to compare any two token uses. But it is not an individuation criterion for community-wide meanings. When does a chain of mutual appearances of meaning sameness support a shared meaning? In practice, it is not straighforward how we should apply the above criterion to individuate community-wide meanings. For example, one extrapolation of the above definition is the following:

> ***Topic continuity:** A community-wide representational tradition composed of the ordered set of token representations $< e_1, \ldots, e_n >$ *supports a shared meaning* iff (i) every member of the set is connected by a chain of mutual appearances of meaning sameness to any other, and (ii) the chain does not defeat the representational tradition.

> ***Topic discontinuity:** A chain of mutual appearances of meaning sameness defeats the representational tradition if some use in the set is incongruent with the other uses.

Clearly, this reading of Schroeter's relational criterion is too strong. The mere presence of one incongruent node would defeat the whole representational tradition, and spoil the possibility of samethinking for the rest of the language users. Another reading is thus required. To individuate community-wide meanings, we need to enrich the theory with a post-hoc interpretation procedure at the level of whole networks of apparent de jure sameness. Applying the congruence constraint holistically actually involves some kind of intricate hermeneutical business. I explain what this operation consists of in a dedicated sub-section below. Before I do this, I will close my comparison between Schroeter's and Cumming's views: is alignment necessary for congruence?

3.4.3 Alignment and congruence

*Inter*personal appearances of meaning sameness typically involve linguistic expressions. It is plausible to think that each representational tradition involve exactly one linguistic continuant together with explicitly anaphoric devices. I don't mean that linguistic expressions cannot undergo a gradual shift in their semantic or syntactic (e.g. phonetic) properties, and I don't mean that appearances of meaning sameness cannot possibly occur via different expression types (in the orthodox sense, not Fiengo and May's). Schroeter seems to want this because she wants appearances of de jure sameness to secure *direct logical relations*:

> The appearance of de jure sameness is crucial to determining direct logical relations among thought contents. The premise [Hesperus appears in the evening], for instance, does not logically entail the conclusion [Phosphorus appears in the evening]: minimal logical coherence does not require the subject to accept the conclusion whenever she accepts the premise, even though the truth of the premise metaphysically guarantees the truth of the conclusion. Similarly, relations of apparent de jure sameness are necessary for direct logical contradictions among thought contents. The fact that it's logically contradictory for the subject to accept that Hesperus appears in the evening and to accept that Hesperus does not appear in the evening depends on the apparent de jure sameness of topic: from the subject's perspective it seems obvious and incontrovertible that both thoughts pick out the very same thing (Hesperus) and attribute the very same property (appearing in the evening). (Schroeter & Schroeter 2016: 6-7)

This passage suggests that appearances of meaning sameness hold between *explicitly* coreferring expressions. So, in general, *different* names cannot trigger appareances of meaning sameness. This is reminiscent of Taylor's characterization of names as "devices of explicit co-reference" (Taylor 2021); and Fiengo & May's characterization of coreference *de jure* as the recurrence of expression-type (again, we don't have to agree with F&M that an anaphoric expression and its antecedent share a syntactic type to make this point). Of course, when it is common ground that e.g. "Hesperus" and "Phosphorus" corefer (in F&M's terminology, if the participants believe the same relevant *translation statements*), these terms will be in general inter-substitutable for the participants. But this does not mean that it will be *obvious*, *incontrovertible* and *epistemically basic* for them that the terms corefer.[53] In the terminology of F&M, the fact that one expression is believed to translate another, non-coindexed expression does not make them corefer *de jure*.[54] Be that as it may, the fact that I am identity-confused with respect to Venus does not seem to prevent my uses of "Hesperus" and "Phosphorus" from belonging to the respective representational traditions.

[53] But see Kaplan (2012), who writes:

> I am tempted to push this line of thought further [that we should distinguish between linguistic and psychological MOPs], to the conclusion that those who, like myself, first heard the names "Hesperus" and "Phosphorus" in a context in which we immediately learned that they named the planet Venus, have only a single way of having the planet in mind (although we have three names for it). This is because I already had Venus in mind, and when I was told about Hesperus and Phosphorus I immediately assimilated them to Venus. (. . .) If this is correct, there is really no saying whether the cognitive content of the three names is or is not the same. It will be the same for some people at some times and different for some people at some times. (Kaplan 2012: 138)

What Kaplan is suggesting here is that psychological MOPs are sometimes *coarser-grained* than linguistic MOPs. Recanati (2019) is following Kaplan here. More genenerally, in the mental file framework, coreference *de jure* is explained by the deployment of the same file. As a result, if a thinker has his VENUS-file labbelled with the three names, then the prediction is that those three names corefer *de jure* for that thinker. In this respect, Schroeter's phenomenological criterion might be more fine-grained. She is closer to the classical Fregean criterion than Kaplan or Recanati, in this respect.

[54] However, whether non-cospelled but covalued expressions can trigger appearances of de jure sameness is an issue that semantic phenomenologists might want to leave open. For example, it is less obvious that a *native bilingual* cannot experience appearances of meaning sameness between words of different languages.

Schroeter's relational criterion for meaning sameness, namely, *being connected by a chain of* undefeated *mutual appearances of meaning sameness*, is thus *prima facie* insensitive to patterns of alignment or misalignment between agents. For example, I may have two unlinked files for Venus, one labelled *Hesperus*, the other labelled *Phosphorus*. This does not prevent each of my files from being embedded in the relevant representational traditions. These representational traditions are disjoint: it seems that no two token uses of "Hesperus" and Phosphorus" are related by a chain of appeance of meaning sameness. (As noted, otherwise we loose the distinction between mere coreference and *de jure* coreference, if the characterization of the distinction is phenomenological).

One might find this view too liberal, and decide to incorporate *alignment* as a condition for congruence. On this more stringent view, agents must have their relevant concepts *aligned* in order for their uses to be *congruent*. Such a construal would certainly go against the spirit of Schroeter's proposal: the organizing desiderata of the proposal is to satisfy the *accessibility constraint* (i.e. the subjective appearance of meaning sameness should be reliable) and the *flexibility constraint* (i.e. the requirement for competence with a term must be minimal). This is because alignment makes it hard to share content, and is typically not apparent to speakers (as I have argued in chapter 4): alignment is typically not part of the experience of meaning sameness. Moreover, if all what is required to share a non-defeated representational tradition is *deference*, then clearly alignment is too strong (Cumming 2013b: 395).

Still, it might be thought that alignment may be required for congruence in the *Paderewski type of cases*, in which a thinker associates *two* mental symbols with just one common currency name. A speaker in the position of Kripke's character Peter will have all her unlinked token uses on the *same* representational tradition. Such a speaker makes the representational tradition to *fork* without reference shift. Two of her token uses will be connected by a chain of mutual appearances of meaning sameness, but the thinker is not disposed to have appearances of meaning sameness with respect to these two uses. This is analogous to the network configuration I have presented in section 3 of chapter 3.[55] In the Paderewski cases, one option is thus to apply the

[55] I reproduce a version of the figure 3.2 here:

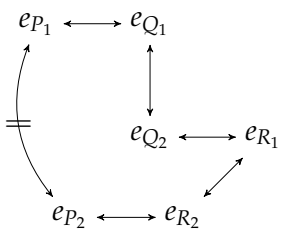


Figure 5.6 – A forking representational tradition

following rule:[56]

> **Non-forking rule**
> Forking R_{inter}–paths defeat the representational tradition.

However, the *non-forking rule* does not yet provide a way to deal with the Paderewski cases. We need to know *how* such a rule ought to be applied. For example, here is one way one could construe the rule:

> **(1)** If you find that the network is forking, then the network does not support shared meaning.

But **(1)** is not a good way of construing the non-forking rule: the mere presence of a forking path (perhaps due to a single agent on the network who happens to be confused on a single occasion) would spoil the possibilities of samesaying and samethinking for all the rest of the agents on the network, on this interpretation of the rule. Clearly, we should prefer a solution à la Perry in terms of pragmalignment along threads (as I have argued in the chapter 5), or a solution à la Fine in terms of an intransitive notion of *same-use* (see below) and more generally, a solution in terms of the notion of same-saying defined on proper portions of the network, to a solution on which same-saying is completely discarded. Another way of construing the non-forking rule is as follows:

> **(2)** If you find that the network is forking, then refine the default partition in such a way that there is samesaying only in non-forking paths.

This way of dealing with the Paderewski cases is equivalent to the way I have proposed to modify Onofri's criterion (see chapter 4 section 5). The problem with this way of applying the non-forking rule is that it seems arbitrary. In particular, what should we do with the token representations that get isolated? It seems that Schroeter's metasemantics has no ressource to ground content-attribution for these sets of token representations.

Formally, this way of dealing with the Paderewski cases is also equivalent to Fine's notion of *coherent referential path* —see above. Fine does not have the problem of grounding content-attribution, because what a coherent referential path defines is not a shared meaning, but a collection of idiolectal names that can be said to stand in the *same-use* relation. Because the relation of *same-use* is allowed to be intransitive, it does not ground shared meaning.

I think the best solution is to combine Fine's insights, and Perry's insights. We discard fine-grained shared meanings, and we explain everything with an intransitive *same-use* relation. Transparency is preserved, at the cost of Shareability. In the Paderewski cases, we predict that two token uses are samesaying if the idiolectal names governing those tokens are connected

[56] See Prosser (2020: 15; note 19) where this idea is suggested.

by a coherent (i.e. non-forking) referential path. Coherent referential paths do not amount to sharing meanings.

Moreover, again when *referential coordination* is what is required, then we can safely take Peter to be referentially coordinated with enlighted speakers. Both of Peter's mental symbols refer to Paderewski. Peter is deferring to the representational tradition, even if he takes himself to be deferring to two distinct traditions.

Having proposed a solution to the Paderewski-cases on behalf of Schroeter, I will now explain the metasemantic procedure by which one assigns a community-wide semantic value to the networks of semantic appearances.

3.5 A post-hoc vindication of community-wide shared meanings

3.5.1 Metasemantic infrastructure vs superstructure

When speakers think and speak, either together or on their own, their uses of token expressions and their correlative attitudes are clustered by the relation of apparent *de jure* coreference, both within and across speakers. As a result, the relation of apparent *de jure* sameness determines a partition of the set of token representations. The elements of such a partition – the connected components induced by the relation of apparent *de jure* sameness over time and between speakers – are the relevant *inputs* into the metasemantic interpretation:

> The unit of interpretation – the unit that demarcates the input into interpretation and the unit to which the interpretation applies – [are] the historically extended bundle of attitudes demarcated by the (...) relations of apparent *de jure* sameness among different speakers in [one's] linguistic community. (Schroeter 2012: 15)

So, the diachronically and socially extended networks of R–connected token representations do not, by themselves, *determine* whether the relation of coreference actually holds between any two given R–connected token representations on the network. Instead, the relation of apparent *de jure* sameness between token uses and associated attitudes merely provide the relevant *inputs into interpretation*, without *determining* what the interpretation of these attitudes should be. I express this design feature of Schroeter's account by saying that her metasemantics is '*post-hoc*'. Again, by this I mean that the assignment of semantic value operates on default (but defeasible) units *already clustered* by the relation of apparent *de jure* coreference (the ground relation R).

I propose using the following terminology: R–networks correspond to what we may call the metasemantic *infrastructure*, namely, the facts about word uses and occurrent attitude states that ground meaning and reference. On the other hand, the 'post-hoc' character of meaning assignment to the networks of semantic appearances makes it belong to the metasemantic

superstructure, namely, a reflexive level about meaning and reference that includes our *theories* about them. Cappelen 2018 has a notion of *metasemantic superstructure* which appears congruent with mine. He defines it as follows:

> Think of the metasemantic superstructure as consisting (at least in part) of our beliefs, hopes, preference, intentions, theories, and other attitudes about meanings and reference (what they are and what they ought to be). (Cappelen 2018: 58-59)[57]

Here is Schroeter's gloss on what I call 'metasemantic superstructure':

> The metasemantic *theory* of a connectedness model seeks to assign semantic values, not to token uses of expressions considered in isolation, but to a *whole bundle* of different uses that are related by apparent de jure sameness. The default goal is to assign a single univocal interpretation for that entire bundle of uses taken as a corporate body. So it is no accident that the appearance of *de jure* sameness is reliable – for the metasemantic *theory* explicitly aims to vindicate these appearances. Only if no such univocal interpretation is available will the metasemantic *theory* be forced to an interpretation that violates the relations of apparent de jure sameness. In a nutshell, then, the appearance of de jure sameness demarcates default units of interpretation. If all goes normally, these appearances will be veridical. (Schroeter 2012: 14, my italics)

Community-wide meanings & Variance When things go well, the metasemantician just has to project an obvious semantic value to the tradition as a whole. For example, consider this sentence use:

> (3) Red is my favorite color

Consider the representational tradition that links all the states pertaining *de jure* to *this* subject matter, the color red. (Agents need not think in words to think about that color). Now, it is at least conceivable that each speaker associates a different intension/extension to the concept RED. One might argue from this that "red" will systematically contribute *distinct* truth-conditions from one speaker to another. (And perhaps the same is true for a single thinker across time). Some philosophers argue from considerations like this to the idea that truth-conditions are never shared. For example, Abreu Zavaleta (2019) proposes that:

> **Variance:**
> Nearly every utterance is such that there is no proposition that more than one language user believes to be that utterance's truth-conditional content. (Abreu Zavaleta 2019: 2)

[57]Cappelen then argues that philosophers often overestimate their ability to define/change meanings via the metasemantic superstructure, which is somewhat disconnected from the metasemantic base (namely, the first-order metasemantic facts). I agree with Cappelen here. I will think of the metasemantic superstructure *constructs* as convenient mutual reifications on the basis of the first-order facts about word uses and occurrent attitude states.

Even if VARIANCE is true, the Schroeterian metasemantic procedure enables to identify 'post-hoc' a single "meaning" for all uses of *red* (I don't mean a single intension/extension). That will be the case if all uses are *congruent*. We don't have to reify a semantic object capturing what's common to all uses of *red* (there might be no such thing). However, we can identify a univocal representational tradition. This suggests that, even if there is no single intension/extension shared by distinct uses of RED, this does not defeat the representational tradition as a whole.

3.5.2 Dealing with incongruence

Malaprops were an example of cases in which things don't go normally and where the metasemantics is forced to an interpretation that violates the appearances of de jure sameness. When things go wrong, appearances are to be *revised*. Given a particular network[58], the goal of the metasemantic theory is to assign a plausible univocal meaning to the largest possible (or *maximal*) subgraph in which all nodes are reachable from every other. *Maximal* means such that one could not find another node anywhere in the graph such that it could be added to the subgraph and all the nodes in the subgraph would still be connected. Accordingly:

> **Goal of the metasemantic interpretation:**
> For each network, seek to assign a maximally plausible univocal meaning (in particular, when applicable, a unique referent) to the largest possible subgraph.[59]

Said differently, when it turns out that appearances of de jure sameness are not veridical, the goal of the metasemantic interpretation is that each cell *in the revised partition* be assigned a single shared meaning (and in particular, when applicable, a unique referent).[60] Revised partitions thus resemble the *objective* notion of co-indexing of expressions across speakers' idiolects in Fiengo & May's framework.

Schroeter thinks of this metasemantic disambiguation procedure as a kind of best commonsense interpretation that one should accept given rational interpretive norms:

[58] I am being relaxed with my uses of 'graphs' for 'networks' and vice versa, although it should be noted that of course *networks* are *concrete* because they are universes of interconnected tokens distributed in space and time, whereas *graphs* are *abstract* mathematical structures that may be used to *represent* networks (and more generally, any kind of pairwise relations between objects).

[59] A subgraph S of a graph G is a graph whose set of nodes and set of links are all subsets of G. All the links and nodes of G might not be present in S; but if a node is present in S, it has a corresponding node in G and any link that connects two nodes in S will also connect the corresponding nodes in G.

[60] There is a class of non-contextual singular terms such that they do not seem to refer, or it's not clear what they refer to, but with respect to which we want to talk of de jure meaning sameness or coordination. The network approach to meaning sameness is actually very potent to handle cases like this. This is because, even without referent, there is nothing mysterious to there being an originating use for terms of this class, and subsequent anaphoric uses on that originating use. See for instance Perry 2012 and Friend 2019. This class includes so-called empty names such as 'Santa Claus', kind terms such as 'unicorn', normative terms such as 'justice' or 'moral wrongness', and numerals such as '5'. Importantly, even though such terms do not refer or do not clearly refer, we use them just like referential expressions as far as meaning sameness is concerned: that is, we intend to use them in a way that is somehow anaphorical on other speakers' uses and our own past uses of these terms. So we still want to be able to apply the notion of meaning sameness or coordination to terms like this. Schroeter aptly uses the more neutral terms 'subject matter' or 'topic' that do not carry the assumption that terms refer. I feel free to to likewise.

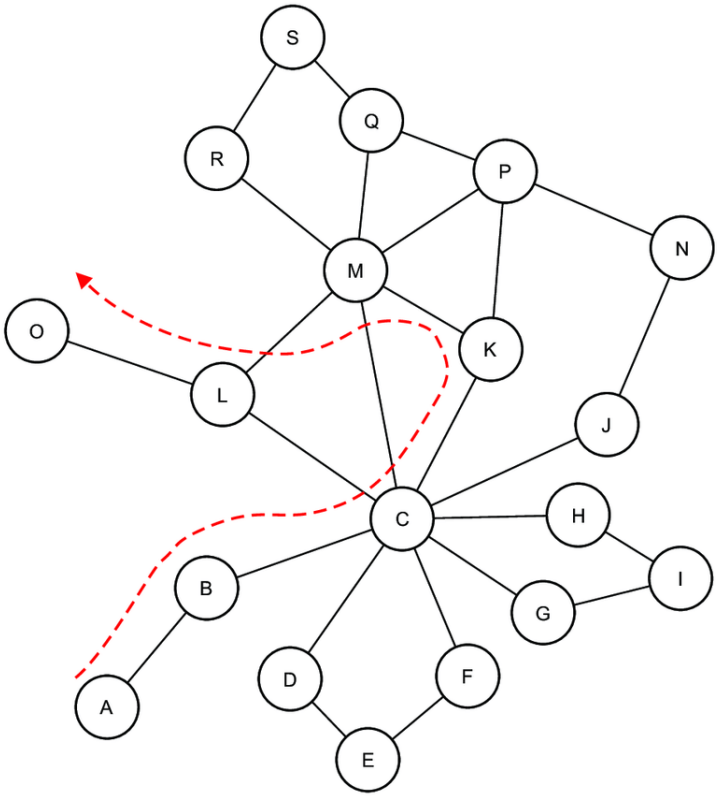


Figure 5.7 – Repartitioning a network of apparent de jure sameness according to congruence

> If the interpretive norms for determining semantic values cannot generate a single, univocal interpretation that vindicates the initial presumption of de jure sameness, then *our best commonsense metasemantics* tells us that these prima facie units must be revised. Just as we do in the Mrs Malaprop case, our metasemantic theory should seek to partition the prima facie unit of interpretation in such a way that each of the [cells] can be assigned an intuitively plausible univocal subject matter. (Schroeter 2012: 18, my italics)

In a slogan, on Schroeter's picture, shared meanings correspond to ideally disambiguated representational traditions. There are issues having to do with articulating the notion of BEST OVERALL DISAMBIGUATION / IDEAL DISAMBIGUATION. Is the best overall disambiguation, a disambiguation that one should accept on the basis of *ideal* reflection and *full* empirical information? Or is it something weaker? If it is something weaker, how much weaker and what is it exactly? I won't try to answer these worries here.[61]

[61] Likewise, this style of metasemantics carry problems typically associated with interpretationism. Among those problems, there are well-known *indeterminacy* issues associated with interpretationism. In at least *some* cases, multiple incompatible interpretive overall disambiguations may be equally good. As a result, there may be no fact of the matter which of those interpretations is really the correct one. Consequently, there being incompatible-but-equally-good interpretations for some units of interpretation may leads to a kind of anti-realism about meaning. In fact, Quine and Davidson, for instance, explicitly acknowledged and endorsed this consequence of their respective versions of interpretationist holism.

3.5.3 Normative continuity from base to superstructure

There is a noteworthy parallel between the normativity inherent to what speakers do *qua* participants in the networks (what each speaker commits to individually), and the normativity inherent to what the Schroeterian metasemantician does in order to reach an ideal disambiguation of the representational traditions. A Schroeter remarks,

> Speakers of a public language, normally intend their own use of an expression *to conform to the best interpretation* of a presumed communal practice. (Schroeter 2012: 16, my italics)

And again:

> *Just as we do* in the Mrs Malaprop case, *our metasemantic theory* should seek to partition the prima facie unit of interpretation in such a way that each of the [cells] can be assigned an intuitively plausible univocal subject matter. (...) This final disambiguating move ensures that our theory of meaning makes plausible referential assignments that are compatible with *the referential standards that we should hold ourselves accountable to* in rational inquiry into the real defining characteristics of familiar topics." (2012: 18, my italics)

This parallel between the way we ordinarily interpret others and ourselves, and the metasemantic interpretivist procedure is important, because it suggests that the normative aspect inherent to the (holistic, post-hoc) metasemantic interpretation of the inputs given by the ground relation *is already present* in ordinary speakers. In particular, there seems to be a distinctively normative element attached to the activity of *participating in the networks*. As Schroeter says,

> Speakers of a public language, *normally intend* their own use of an expression *to conform to the best interpretation* of a presumed communal practice.[62]

We may express this as follows:[63]

> **Norm of univocity:**
> The very purpose of the practice of participating in the networks is to warrant a shared *univocal* representational tradition. Accordingly, in participating in the networks, speakers are committed to use their terms in a way that makes best sense of the communal practice.

[62] id. my italics.

[63] Schroeter's metasemantic 'post-hoc' procedure is vulnerable to *the normative fallacy* (Campbell 1970, Moravcsik 1998 both mentioned in Thuns 2021, 3.3). In a nutshell:

> **The normative fallacy**
> One cannot infer what a word means from normative intuitions about what one thinks the word ought to mean.

I won't discuss this worry here.

Speakers have a deferential attitude vis à vis the representational traditions they participate in: they commit their own use to be aligned with the (presumed) community use. One's use is taken to be corrigible in light of evidence from other speakers in the shared practice. I examine more closely this deferential feature of language use in the penultimate section of the chapter.

The interconnectedness between speakers who all intend to use a referential word according to a presumed community use constitutes a *file-network* composed of all the mental files deferentially connected in the minds of the agents participating in the same representational tradition. As a first approach, we can think of such a distributed file as the network of mental files involved in the attitudes bundled by chains of mutual appareances of meaning sameness.

However, I believe we should understand the representational traditions in terms of the collective epistemic goal that they serve. The goal of the representational traditions is to accumulate information on encyclopedic entries of general interest. Arguably, there is no separate linguistic labor for 'elm', and for, say, the French word 'orme' (even though it is at least possible that there is no speaker for whom "elm" and "orme" appear *de jure* the same). Accordingly, the distributed files in which individual files participate *are not fine-grained*: they are as coarse-grained as referents, when the speech community is not collectively in a Frege case. I examine more specifically the deferential dimension related to the epistemic purpose of the representational traditions in the next section.

4 File-networks again: the Human mental encyclopedia

The Human encyclopedia is the store of human knowledge. The Human *mental* encyclopedia is the totality of human knowledge stored in human brains. Sharing words increases the quantity of information stored up, and the degree of accessibility of the "memory" of the distributed human mental encyclopedia (see Wegner 1995 for *a computer network model of human transactive memory*).[64] The normativity for the correctness conditions of the use of words Schroeter talks about, has to do with the *epistemic goal* of the representational traditions: maintaining, transmitting and accumulating knowledge about things that are of mutual interest. This section proposes to understand semantic deference in light of this collective epistemic goal.

There are two aspects of the deference to community use that I want to bring to bear on each other. First, there is *semantic* deference in the sense that the reference of a word is fixed at the community level (at the level of the representational tradition as a whole, as Schroeter says). This is the idea of the *division of linguistic labor* Putnam was talking about (Putnam 1975). Second, there is *epistemic* deference to experts, producers or more generally knowledgable

[64] As Wegner writes,

> Human beings in pairs and groups form message-passing, directory-sharing memory systems. (Wegner 1995: 326 cited in O'Madagain 2018, earlier version, ms.)

people when seeking knowledge about an encyclopedic entry of shared interest, which I suggest is an important reason for using words in a deferential way (i.e. to defer in the semantic sense)—see O'Madagain 2018 for arguments supporting this idea. I propose that we should understand both kinds of deference in terms of each other. We defer in the semantic sense to the community as a whole (in particular, to the agents whose mental files contribute to fix the reference of the term), because we seek to accumulate information through testimony about encyclopedic entries of shared interest.[65] This is an *info-centric* conception of representational traditions.[66] Here is a relevant quote from Perry (2012) in which he characterizes the epistemic goal of representational traditions:

> Each person's mind is a pool of information about many objects, in the form of the [mental symbols] they have of those objects, and the ideas they associate with those symbols they have about the objects. These pools of information are accessible to us through language. In virtue of the indirect connection networks provide, the files of people whom we will never meet are connected to and influence our own. In virtue of written tokens-books, archives, recordings and such, the information stored in the minds that produced the utterances of which those tokens are traces is available to us. (Perry 2012: 198)

If one wants to benefit from testimony on encyclopedic entries of mutual interest, one should intend to use the words as they are used by the community. Deference makes possible the accumulation of information from other uses of the word.[67] For example, a good way to know more about some subject matter is to put a word that name it in a search engine. It is also a typical way people acquire knowledge by testimony these days.

4.1 Consumers vs producers

Each participant in a network is a more or less active member of the representational tradition. I said 'more or less' because not all language users contribute equally to (i) determining the reference of the word-meaning pair continuant and (ii) providing new information about the relevant topic. Relatedly, Evans (1982) draws a distinction between *producers* and *consumers* in a name-using practice.[68] The producers are the

> Core group of speakers who regularly and reliably recognize an individual, x, as *NN* (Evans 1982: 388)

Producers acquire information about x by interacting directly with x and transmit this information to others. Only producers can contribute *new* information into the practice.[69] Grounding

[65] *Pace* Devitt (2015: 216-217).

[66] Friend's (2014) term characterizing the approach originally developped in Evans (1973).

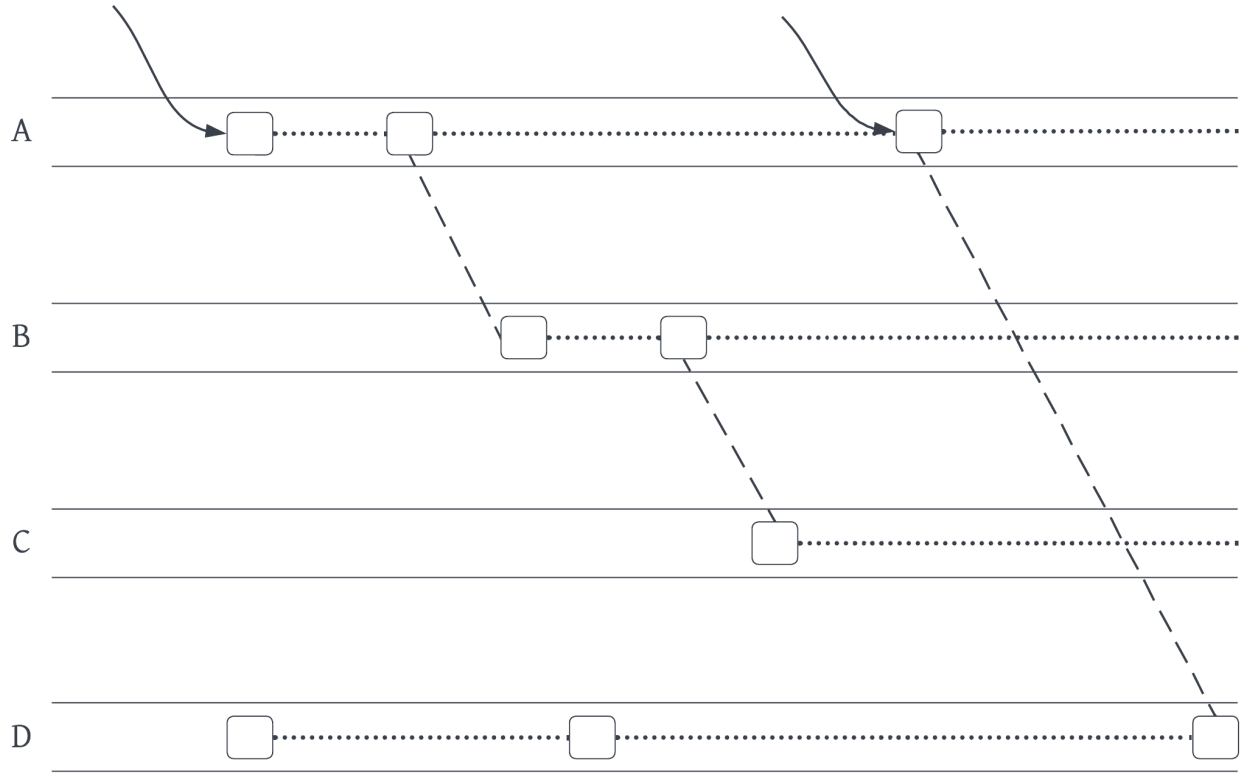
[67] This sense of 'normativity' is arguably very weak. I put aside issues having to do with rule-following and Kripke's (1982) reading of Wittgenstein on this matter here. See Glüer and Wikforss 2022 for an overview.

[68] I suggest that we may generalize this distinction to other kinds of singular terms below.

[69] When the referent is not present anymore in the speech community, participating consumers such as historians or archeologists are arguably analogous to producers. See below for a characterization of the relevant notions.

(the anchoring of the common currency name in a referent) takes place via the mental files of the producers. Evans (1982: 126) has an interesting diagram representing what he calls "a simple model of social informational system" (a file-network), slightly modified in figure 5.8.[70] *Unbroken diagonal lines* represent perceptual links; *broken lines* represent the transmission of information through communication; *dotted lines* represent the retention of information in memory, *squares* represent mental file *stages*; the left-right dimension stands for the temporal dimension; each continuant such as *A*, *B*, *C* represents the mental file continuant of a single agent. The whole thing depicts a distributed file multiply grounded (Recanati 2016: 126 and passim). Of course, in the real world, widely distributed files are much more complex than the one represented here.[71]

Figure 5.8 – Evans' model of a "social informational system" (Evans (1982))



Consumers are those introduced into the practice of talking about *x* without directly knowing *x*. Their *NN*–labelled mental files about *x* is parasitic on the files of producers. Following Dickie 2011, we may further distinguish between *participating* and *parasiting* consumers. The partic-

[70] What Evans calls a social informational system corresponds roughly to what Sperber calls 'social cognitive causal chains' (Sperber 1996, 2001).

[71] The chains may take different forms depending on many factors. For a taxonomy of transmission chains, see Morin (2015: 138-140). Morin explains that people copy, adopt, transform the elements of culture that they find attractive. This is the *Cultural Attraction Theory*.

ipating consumers have a *NN*-labelled mental file containing information about *x*, whereas the *parasiting* consumers have virtually no information about *x* and refer to *x* simply in virtue of intending to use *NN* the way it is used by the producers or participating consumers in the practice. In fact, it's not clear that *parasiting* consumers can entertain genuine singular thoughts about *x* using *NN*, as opposed to meta-linguistic thoughts of the form 'the individual named "*NN*" is such and such' (Goodman 2016). (Likewise, recall my earlier suggestion that *purely deferential concepts* may be identified with mental mixed-quotes.) Recanati 2016 calls *deferential file*, a file that is based solely on the name-using practice. As just mentioned, when a thinker has a purely deferential file about an object, he is thus a 'parasitic' consumer.

In 'parasitic' cases, I suggest that the MOP-role may be played by the representation of an idiolectal expression-type, roughly as Fiengo & May (2006) proposed. When the deferring thinker gets more information about the referent, the purely linguistic MOP will be converted into a regular encyclopedic file (Recanati 2001, 2016). When the thinker has a regular encyclopedic file, he is either a 'participating' consumer or a 'producer', depending on whether the thinker has the capacity to identify the referent *as* the bearer of the name (which is required to be a producer), or not. Here is Recanati's gloss about deferential files:

> When a name is used purely deferentially (as when one picks up a name overheard in a conversation), the individual mental file the language user associates with the name is a deferential file: a file based on a specific [Epistemically Rewarding] relation, that of being party to a proper name-using practice (Recanati [1997], 2000, 2001). Being party to a proper name-using practice (through acquiring the name from someone else) is an epistemically rewarding relation: one is in a position to gain information about the referent of the name through testimony (by attending to the name when it is used, or by using it oneself to elicit information from others). One has access to the distributed file of the community. Let us call that ER relation, made available by the mere sharing of words, the 'deferential relation'. (Recanati 2016: 128)

4.2 Division of linguistic labor & social grounding

The distinction between consumers and producers also seems relevant for other kind of terms, not just names. As mentioned in the introduction of this chapter, a speaker who is incapable of recognizing elms by sight, and know very few things about elms, may still mean elm with "elm". Such a speaker borrows the meaning of the word from his speech community, and in particular, the people who are able to discriminate elms and characterize them in substantive terms. Meaning borrowing may happen solely in virtue of a *linguistic* contact with other speakers who use the *word* 'elm'. Such a speaker is a consumer in the representational practice. By contrast, experts such as botanists, or knowledgable people about trees, can differentiate between elms and non-elms, and associate substantive information about elms. They are pro-

ducers in the representational practice, because they actively contribute to the linguistic labor.

The semantic content of a name, or a kind term, is (very roughly) determined by the files of the producers in the representational practice. More generally, the extension of words such as "elm" or "sofa" is determined socially (at the level of the whole file-network) rather than individually (Putnam 1975, Burge 1979). Virtually each peer in the network defers to the linguistic community as a whole. This seems true even for non-scientific, ordinary terms like "sofa" for which there is no clear need for experts. A file-network of language users is not a cluster of independent idiolects. Rather, speakers' idiolect refer *by coreferring* through deference with peer-idiolects.[72] Idiolects are interconnected because speakers pursue these deferential connections in order to accumulate information on shared subject matters. The interconnectedness between speakers (their files) give rise to a file-network in which each file is embedded. The network constitutes a distributed file managed by the community as a whole (Recanati 2016: 127-128).[73]

A deferential file-network resembles a *peer-to-peer informational network* where the total information believed about a mutually interesting encyclopedic entry is distributed across different minds.[74] The label 'peer-to-peer' evokes the fact that deference is among particular, local users, and there is no central authority with respect to how words should be used, and how information should be stored (see Figure 5.9):

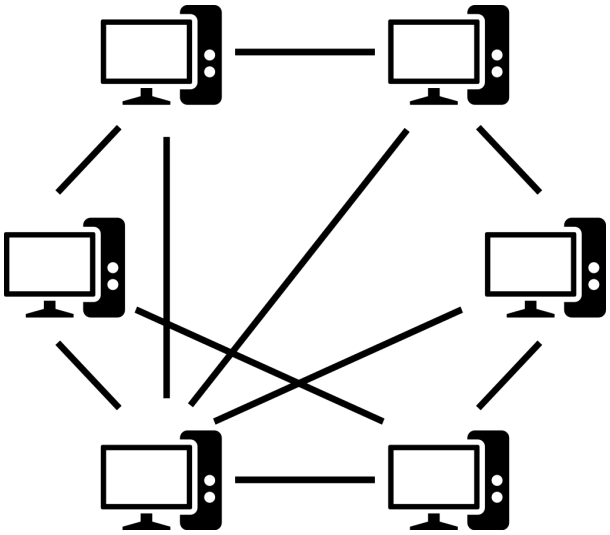


Figure 5.9 – A deferential network resembles a peer-to-peer (P2P) network.
Source: The 360 Degree, CC BY-SA 4.0, via Wikimedia Commons

[72]Coreference by deference seems to be a clear notion of coreference *de jure* in the interpersonal domain, see Prosser 2019. We don't have to construe deference in terms of shared MOPs, or even in terms of shared word (see above an argument against *Direct derivation and shared assignment*). In fact, as mentioned, an agent whose file is purely deferential is typically very different from the files of the producers. What matters is the connection between them.

[73]All this fits well with Schroeter's account I have examined above.

[74]I don't mean that *meanings* are themselves distributed.

Labelling our concepts (files) with public words get us a distributed filing system (as Wegner 1995 already suggests) where information about a mutually interesting encyclopedic entry can flow. Norms of correctness for the use of words exist at least in part because agents want to accumulate information about objects of shared interest using testimony. These norms are *conditional* on this epistemic goal. I suggest that words of a public language act as *network protocols*, in a sense suggested just below.

4.3 Words as Network Protocols

Paraphrasing Recanati (1993: 183), we may say that linguistic expressions have two functions in relation to mental files, and the peer-to-peer file-networks in which they embed. Words can be used to initiate mental files in a way that make them networked together, and they can be used to retrieve mental files in synch across peers, and gain or transmit relevant information through testimony from the P2P network:

> **Initating/networking function** Simply using a 'common currency' word creates a mental file on its referent. Recanati (2016) calls this relation, *the deferential relation*. In other words, merely sharing a public word give access to the distributed file of the community.
>
> **Access/synch up function** The label of a concept gives access to this concept. Uttering the label of a concept acts as an instruction to retrieve that concept so as to use/change it in the way required for the informational transaction. A label essentially serves as a rule *to synchronize memory addresses of different thinkers*.[75]

In such networks, interconnected nodes ("peers") share information amongst each other without the use of a centralized "administrative" system. Instead, each node is responsible for the reliability and consistency of the information associated with a distributed file. The social/psychological mechanisms involved may be compared to a coherence mechanism for a network file system. That is, the freedom to benefit from the division of linguistic labor also requires some work from each of us in order to keep tracking the same referent and its properties. In particular, one should:

> (i) strive to keep one's own mental file self-consistent with respect to the information it contains. We do not want to store two contradicting predicative entries within a single file. So, the way we maintain and update our mental file is governed by the principle of non-contradition, according to which objects cannot have and not have a given property. For any property ϕ, we typically assume that there is a fact of the matter as to whether we should add "ɪs ϕ" to the relevant mental file.[76]

[75] I am not implying that sharing a lexical item is sufficient or necessary for coordination.

[76] Priest 2006/1987 argues that there are true contradictions. I put aside "dialetheias" here, and they do not concern singular thoughts of the sort I am dealing with. Priest says:

> The paradoxes are all arguments starting with apparently analytic principles concerning truth, membership etc., and proceeding via apparently valid reasoning, to a conclusion of the form 'α and not-α'.

(ii) strive to keep one's own mental files consistent with the mental files *of others* with respect to the information they respectively contain, and be sensitive to incoming relevant evidence from other speakers (when we care about the topic). As Recanati writes,

> The community filters out information tentatively contributed to the distributed file by screening testimony and correcting tentative individual contributions when they do not fit. [Footnote — For example, if I tell you that Napoleon died a few years ago, you will act as a gatekeeper and do your best to prevent that piece of alleged testimony from entering the public file associated with the name 'Napoleon'.] In this way the community pools information from the interconnected individual files so as to build a coherent body of information about the reference of the distributed file. (Recanati 2016: 126)

Again, the "should" here expresses something like a duty conditional on the epistemic goal of accumulating information via testimony. It is at least conceivable that one might not care, and so not be "obligated" by deferential norms. On the other hand, it is at least *questionable* whether a speaker *could* mean something *non-deferentially* with a word from a public language (unless, that is, one is coining a *new* word, in which case one could use this word non-deferentially on the first occasion of use). Non-deferential uses of language seem at any rate very rare.[77]

Wikipedia & the public files Articulating the fine-grained details of the social mechanisms by which a community manages a distributed file is beyond the scope of this section. Let me make a few suggestive remarks. The distributed files are managed in a way that is in some important respects comparable to the practices of information production, management and utilization on the Internet (see e.g. Thagard 2001). I will use *Wikipedia* as an example. Wikipedia is a free online encyclopedia that depends on the collaborative effort of decentralized writers who contribute to this constantly increasing store of information. Wikipedia is decentralized in the sense that the ability to add information is completely open and public.[78] Where classical encyclopedias rely on academics to provide information, Wikipedia is more like a peer-to-peer network, giving this role to the public. (Of course, many 'Wikipedians' are in fact scholars). I suggest that distributed files are managed in a comparable way, at the community level, by decentralized agents. If an edit on Wikipedia is not validated by other contributors, it will be modified until a consensus is reached. If no consensus is reached, the disagreement is settled by an appointed expert. The dynamic evolution of content within *Wikipedia* is a significant difference from classical encyclopedias, which offer a more centralized and more static repository of information. I suggest that public files are more like wiki entries than they resemble classical

> *Prima facie*, therefore, they show the existence of dialetheias. Those who would deny dialetheism have to show what is wrong with the arguments—of every single argument, that is. For every single argument they must locate a premise that is untrue, or a step that is invalid. (Priest 2006: 9)

[77] *Pace* Chalmers 2012, see chapter 6 section 9; see also the excursus *Twin-Earthability and Internalism*.
[78] It is not decentralized in a distributed system sense, because it is based on a central database.

static centralized encyclopedias. Again, we may think of the informational content of public files in the same way. Just like with public files, the opportunity to provide misinformation on Wikipedia exists. However, on Wikipedia, the transparency of edits makes it straightforward for honest writers to identify and rectify changes (contributions are transparent through a time-stamped history of all edits made visible to all users). There is of course no equivalent of this for *public files*, because there is no central database registering all the contributions of language users. But the mechanism of epistemic vigilance, and the consensus-based edit management, is essentially the same (see Sperber et al 2010). In a sense, Wikipedia encyclopedic entries are a materialization of the public files. We are all, *qua* producers or participating consumers, more-or-less-expert 'Wikipedians' of a sort regarding the topics we introduce in conversations.

On the file-network picture, samethinking involves real world connections between agents. In the last section, I briefly address how we can make sense of samethinking without causal link, in part by capitalizing on Schroeter's notion of *congruence*.

5 Samethinking without causal link

Sandgren 2019 complains that causal models of samethinking are incomplete, because they cannot account for intentional identity without causal link. But it is important to note that meanings can be distributed *without transmission*. Sandgren mentions the following case to illustrate intentional identity without causal link:

> Two mathematicians, Leibniz and Newton, quite independently of one another, both come up with the idea of a mathematical function *integration* (though they use differentwords to refer to it) *that plays almost the same role* in their respective mathematical reasoning. (Sandgren 2019: 3682, emphasis mine)

As Morin concurs:

> Some distributions owe nothing at all to transmission—like the various inventions of agriculture, or Newton and Leibniz's two discoveries of differential calculus. These are cases of distribution without transmission. We shall only use the word *diffusion* to point at distributions that, in contrast, owe something to transmission. (Morin 2015: 23)

Two populations causally isolated from each other can each develop similar meanings in response to similar environments. So two causally independent representational traditions can converge on similar meanings. Echoing Schroeter 2012 and Morin 2015, I call configurations of overlapping representational traditions of this kind, *congruence without transmission*.

A representational practice's *distribution* is the set of points in space and time where the practice can be found (Morin 2015). Case of congruence without transmission is a kind of distribution

without causal link (in the relevant sense: for example, sharing a biological makeup is a causal link of a sort, but not in the intended *interactive/communicative* sense). Meanings can be distributed merely in virtue of the fact that congruent representations can develop naturally in organisms with a shared evolutionary history and ecology, giving rise to mental representations with similar representational functions. Congruence is a matter of similarity of representational function, as a result of a similarity in ecology, understanding, and history of use.[79] Thus, we answer Sandgren's (2019) objection against causal-historical networks to the effect that they cannot account for samethinking without causal links. Because we need an interpretationist layer in terms of *congruence* anyway (Schroeter 2012), we may explain samethinking without causal link in terms of congruence without transmission.

6 Taking stock

I began this chapter with the following issue: if concepts are not shared, how can we account for partial understanding/word learning/correctness conditions for the use of words? The answer I proposed is that, even if Shareability is false, and concepts (finer-grained than reference) are not shared, we can still account for partial understanding/word learning/correctness conditions for the use of words in terms of deference, for which only referential coordination is required. We made sense of this in terms of two *bootstrapping* claims:

> **(i)** the use of 'common currency' words trigger appearances of semantic sameness in language users;
>
> **(ii)** these semantic appearances make it the case that things happen as if meanings were shared, and give rise to representational traditions to which speakers intend to conform their use of words.

Thanks to (i) and (ii), we can make sense of surrogate community-wide meanings. At the metasemantic level, surrogate community-wide meanings are grounded in the mutual reifications of file-networks that stabilize lexicalized mental representation in a human population. When a speaker has distinct idiolectal expressions belonging to the same representational tradition, as in the Paderewski cases, I proposed that we use an intransitive *same-use* relation to preserve the transparency of idiolectal meanings. For the other kinds of defeaters (malapropisms, semantic shifts, etc.), the theorist can repartition 'post-hoc' the representational traditions and project the most obvious meanings onto each cells, along the lines offered by Schroeter 2012. I recommend to have an instrumentalist attitude attitude towards these constructs: they need not correspond to any semantic object that is shared.

[79] See Brigandt's related notion of the *epistemic goal of a concept*, Brigandt 2010.

6 Conclusion: What is samethinking?

I would like to conclude this dissertation by doing four things. First, I propose to step back from specific theories, and think about the issue at a greater level of generality, by delineating the solution space for the problem of characterizing samethinking (section 1). Next, I locate the model I have suggested in this dissertation within the delineated solution space, and draw some notable implications of this model (section 2). I then indicate two lines of research, which I think are worth pursuing in order to further develop the ideas presented in this thesis (section 3). Finally, I conclude by stressing a distinction that has emerged from this work between two important notions, which I believe have not been clearly distinguished in the literature (section 4).

1 The solution space, upon further examination

One idea that I hope will have emerged from this work is that samethinking is a rich and complex issue. Its complexity has been under-appreciated until recently in the philosophy of mind and language. For a long time, samethinking was simply assumed to be *semantic resemblance*. Under the influence of Fodor's argument—that resemblance views need a notion of identity but leave it undefined hence are not viable—this conception is less popular today than it was then.[1] It is fair to say that the current received view understands samethinking as identity. In fact, most theories of content are *based* on an understanding of samethinking as identity. As such, they simply assume that sentences used in context have contents which are both what a speaker expresses and what a hearer understands.[2] For the last ten years however, there has been a burgeoning and newly evolving literature on the issue of samethinking. The solution space for the problem of samethinking has been noticeably enriched in recent years. Samethinking is emerging as a topic of importance in its own right. I have critically examined

[1] Fodor 1998, Fodor & Lepore 1992.

[2] One interesting exception is the MSDRT ("Mental Discourse Representation Theory") of Kamp (e.g. Kamp 2013, 2015, 2021, 2022) and Maier (e.g. Maier 2016a, 2016b), which converge on the framework suggested in this dissertation.

Figure 6.1 – Solution space

important recent proposals in this dissertation. Here, I want to gain perspective on the issue. My aim is to provide a comprehensive map of the solution space.

Taking advantage of the theoretical distance provided by the work carried out in this thesis, this section is concerned with adequately defining general classes into which views of samethinking might fall. Accordingly, the discussion will be cast at an asbtract level.[3] I will proceed by splitting the question "What is samethinking?" recursively into smaller questions, using a divide-and-conquer strategy along a binary decision tree (depicted in Figure 6.1). To my knowledge, there has not yet been an attempt to map the entire space of views on the problem of samethinking.[4] In this section, my goal is methodological. I offer this as a modest meta-metasemantic contribution, as it were.

Starting from the root node at the top and following a path down the tree, one gradually refines views through a sequence of classificatory decisions, and arrives at specific views by reaching down to a terminal node (i.e. a node that cannot be divided any further). Let us start at the top of the tree. The first classification decision is the most abstract and general question about samethinking.[5] The question "What is samethinking?" presupposes that samethinking *exists*. Some theorists might want to deny just that. The most general question about samethinking is therefore about its very existence. Is samethinking a real phenomenon? *Realism* says yes, *Eliminativism* says no.

Realist vs Eliminativist models Eliminativism is the view that samethinking is not a real phenomenon. I can think of two ways to be eliminativist. A first way is to say that the only robust notion of "samethinking" just is coreference; samethinking defined as a relation more stringent than coreference is an illusion. We may call such a view, the *radical Referentialist view* (already cursively mentioned in the general introduction, n.3). According to eliminativism, speakers can communicate using a term iff they attach the same referent to it. Likewise, the view says that there is no robust notion of agreement and disagreement finer-grained than coreference.

[3] The level of generality for each classification decision is determined (roughly) by the corresponding level of tree depth (tree depth is the length of the longest path from the root to a terminal node. Here in Fig 6.1 the tree depth is 6).

[4] But see Gray 2017 for a survey of the relationist views; Gray 2022; Schroeter 2012; Fine 2007 and Schroeter & Schroeter 2016 for useful characterizations of the organizing distinction between relationist models ("binding"/"connectedness") and intrinsicalist ("matching") models.

[5] A decision tree thus exhibits the law of Port Royal's *Logic* (Arnauld & Nicole 1662/1993) according to which "comprehensions" (what we would call "intensions" today) and extensions are inversely related:

> L'autre sorte d'addition, qu'on peut appeler détermination, est quand ce qu'on ajoute à un mot général en restreint la signification, et fait qu'il ne se prend plus pour ce mot général dans toute son étendue, mais seulement pour une partie de cette étendue ; comme si je dis : *Les corps transparents*, *les hommes savants*, *un animal raisonnable*. Ces additions ne sont point de simples explications, mais des déterminations, parce qu'elles restreignent l'étendue du premier terme, en faisant que le mot de corps ne signifie plus qu'une partie des corps, le mot d'homme, qu'une partie des hommes, le mot d'animal, qu'une partie des animaux. (Arnauld & Nicole 1662/1993: 59)

So, the more depth a category has in the tree, the richer its "comprehension", the smaller its extension.

The only robust notion of interpersonal (dis)agreement there is, is extensional isomorphism. Relatedly, the view denies that *de dicto* attitude reports induce an hyperintensional linguistic context. Instead, truth conditions of attitude reports are all coarse-grained. In other words, there is only one unique *de re* reading for attitude reports. Has this view been defended? Russell (1903) might come close to it. And so-called *Naive Russellians* (Salmon 1986, 1989; Soames 1987, 1989, Braun 1998) might come close to it as well.

Another possible version of eliminativism corresponds to *Ephemeralism* applied to the interpersonal domain (Woodfield 1991, Casasanto & Lupyan 2015, both cited in Murez 2021). It says that different agents never think the same, be it at a time or across time. One may think of ephemeralism as a kind of *Heracliteanism* applied to cognition (a particularly radical version might say that one never (or almost never, perhaps) refer to the same entity twice).[6] According to one construal of this view, samethinking solely consists in our experience thereof; but the appearances of *de jure* sameness are always *illusory*. One motivation for *Ephemeralism* is *Holism* together with the claim that Shareability (**SHAR**) would be a necessary condition of samethinking, if it existed. (Note this is different from the view that combines Holism and the claim that samethinking is similarity, to be introduced later on). *Holism* is a radical version of functional role theory about the individuation of concepts. Functional role theories individuate concepts in terms of their inferential relations to other concepts in the mind of a subject.[7] We can think of the conceptual repertoire of a subject as a network (which might include mental entities other than concepts as well, such as affects, mental images, etc), and each concepts as individuated in terms of their position in the network (Pollock 2020, a recent advocate). Functional role theories can be more or less holistic depending on the proportion of the network they take to be relevant for individuating a given concept. A *holistic* theory says that concepts are individuated by their relation to *all* other cognitive entities in the network (Pollock 2020). In other words, for any given concept, the *total* functional role of the concept is responsible for the identity of that concept (e.g. Schneider 2009 is a proponent of a holistic functional role theory of concepts). Note that Holism *per se* does not entail *Ephemeralism*. It entails *Ephemeralism* only on the assumption that (**SHAR**) is necessary for samethinking (or would be necessary for it, if it existed). A *non-eliminativist* holist might explain samethinking in terms of *similarity* of functional role, or something along this line. One might construe *is-similar-to* as a relation between vehicles, or contents, and that relation is very likely intransitive.

It is not clear how ephemeralism applied to the interpersonal domain might explain the intuitive contrast between successful communication and miscommunication, the contrast between non-

[6] One famous fragment of Heraclitus says:

> You cannot step twice into the same rivers; for fresh waters are ever flowing in upon you. (Fragment 12)

Which seems to imply the particularly radical version of Ephemeralism.

[7] For this reason, one might challenge that this family of views belongs to intrinsicalist views. I will come back to that when discussing the definition of "Relationist model".

genuine and genuine (dis)agreement, and the contrast between *de dicto* and *de re* attitude reports. It may be that ephemeralism is best combined with the aforementioned *radical Referentialist view*. It is incumbent on the *radical Referentialist view* to explain away our intuitions about samethinking, perhaps as a pragmatic side effect. I take this view to be a last resort, when one has exhausted the candidate views.[8]

Uniform vs Mixed models Let us move to the *Realist* views. Once we say that samethinking is a real phenomenon, the most natural type of question at this current depth of the classication seems to ask what kind of relation samethinking is. However, this is not (I suggest) the best type of question to ask if one wants to proceed at the level of abstraction required at the current tree depth of the classification. Instead I propose that we are faced with the following question: Is samethinking a *uniform* phenomenon? This question seems vague as long as "uniform" has not been defined. The opposite of "uniform" I call "mixed". A model of samethinking is *mixed* when there are different domains such that they each involve a *different* theory of samethinking, according to that model. A model of samethinking is uniform iff it is not mixed.[9] What counts as a domain? For example, we have the intersubjective vs the intrasubjective domain. We have the domain of indexical *de se* thoughts vs the domain of non-*de se* thoughts. We have the domain of communicative events vs the domain of attitude reporting events, and so on. Mixed models thus offer differential explanations of samethinking depending on the domain considered. They vary according to (i) the selection of domains they think should receive a differential treatment, (ii) the set of uniform models they think should be mixed, and (iii) the distribution of these models over the selected domains. Since mixed models are composed of specific views that constitute the uniform models, we need to be clear about what the uniform models are before we can examine the mixed models (an extensive category).[10] I turn to the uniform models.

[8] An analogy: illusionism about phenomenal consciousness—the view that we seem to be phenomenally conscious but we are not— is a view one possibly accepts only when one is convinced that no other non-illusionist view on phenomenal consciousness is viable (e.g. Kammerer 2019).

[9] We may stipulate that eliminativism is a uniform model but with a relation that is neither external nor internal (i.e. the next classification decision following the path down the tree) because it is the empty relation.

[10] In principle all views of the solution space can be combined, that is $\binom{n}{p}$ possible combinaisons with n representing the number of combinable views, and p is the number of selected domains in need of a differential explanation. (On the presented version of the tree, n =12, see figure 6.1) To make the space of views more vivid, we can represent the $\binom{n}{p}$ with a double entry table: in column the values of p, in row, those of n. A schema of the table is depicted in Table 6.1:

Values of p / Values of n	Domain 1	Domain 2	...	p
(SCI)				
(SVI)				
...				
n				

Table 6.1 – Possible Mixed views

Relationist vs Intrinsicalist models The two topmost categories of uniform models are, at this stage of the dissertation, well known. The classificatory decision here is whether samethinking is understood as an *external* relation (such as marriage) or rather as an *internal* relation (such as similarity). Unsurprinsingly, this division is a central theme in the current debate about samethinking (e.g. Fine 2007, Heck 2012, Schroeter 2012, Schroeter & Schroeter 2016, Gray 2017, Valente & Verdejo 2021, Gray 2022). My focus here is methodological. I will not go back over the arguments pro or contra specific views. Rather, I propose that we gain perspective by going back to the issue of the definition of the attribute that decides membership to the Relationist family (the Intrinsicalist family, resp.). Doing this is useful, because there are open problems with the mainstream definition of Relationism (Intrinsicalism, resp.). The input definition here is:

> **Relationism**
> Samethinking is an external relation between representations.

This definition is often invoked (under one guise or another) by participants in this debate (Fine 2007, Schroeter 2012, Gray 2017, 2022, Valente & Verdejo 2021). Consider, for example, this passage from Fine:

> According to this view — which I call "Semantic Relationism" — the fact that two utterances say the same thing is not entirely a matter of their *intrinsic* semantic features; it may also turn on semantic relationships among the utterances or their parts which are not reducible to those features. We must, therefore, recognize that there may be *irreducible* semantic relationships, ones not reducible to the intrinsic semantic features of the expressions between which they hold. (Fine 2007: 3)

This quote is about *semantic* relationism (a particular version of relationism). But one can remove the occurrences of "semantic" and get a generic characterization.[11] Likewise, the relation does not have to be restricted to language. It can be extended to thoughts. Fine's characterization relies on the intrinsic/extrinsic distinction. (I am using the term "external" instead of "extrinsic", but the idea expressed is the same as Fine's). When is a representational feature

To illustrate, if one thinks the distinction between intra-personal and inter-personal determines the set of the relevant domains in need of differential explanation, then $p = 2$. Significant mixed models in the current debate include Gray 2022's *Minimal Fregeanism*, which offers an explanation of samethinking in terms of content identity with respect to essentially indexical thoughts, but in relationist terms for non-*de se* thoughts. Its core idea is that relationism explains cognitive significance better than identity models as far as non-*de se* thoughts are concerned. However, Gray finds that we need identity models to explain cognitive significance with respect to *essentially indexical* thoughts. Discussing Gray's proposal is for another occasion. For a recent edited collection on the *de se* and sametinking involving *de se* thoughts, see Garcia-Carpintero & Torre 2016. On the virtues of hybridizing relationist and intrinsicalist models, see what Valente & Vejrdejo 2021 say with respect to what they dub "weakly relational" models. Another salient option is to say that the relation of coordination is transitive in the *intra*personal domain, but not in the *inter*personal domain (Fine 2007, and the construal of the mental file theory in terms of 'mental filing' in Gray & Goodman 2021 can be used in a similar fashion, if combined with intransitive relationism in the interpersonal domain). Or, some views might want the relation of coordination to be characterized in epistemic terms in the *inter*personal domain, but in non-epistemic terms in the intrapersonal domain. And so forth.

[11] Gray uses the term "representational" instead of "semantic", which has the advantage (at this level of generality) of being neutral between semantic and non-semantic characterizations.

intrinsic, and when is it *extrinsic*? A first pass at the notion is to say that a representational attribute is intrinsic to the representation it is an attribute of, when it does not depend on relations to *other* representations. Along this line, Gray writes, "intrinsic representational features are those which can be stated without reference to another representation" (Gray 2017: 4). But as he rightly remarks, this characterization is description-dependent. Recall, for instance, the Equivalence Class Fregean strategies of Dickie & Rattan 2010 or Cumming 2013a, 2013b. These theorists define a representational feature (a "sense" or a "content") in terms of a relation between representations. Through an "abstraction principle', they turn this relation-based feature into an intrinsic one. So we want something stronger, perhaps using an ontological notion, such as the metaphysical notion of *grounding*. I am note sure how to state the relevant condition. We might try to say that a representational property is intrinsic in the targeted sense *when it can be instantiated by a given representation independently from other representations*.

But this looks too strong, at any rate poorly formulated, for the following reasons. This criterion, in addition to the fact that it is not very clear, seems to have bad classificatory consequences. For example, a representation having the *conceptual role* that it does is not intrinsic to it as defined, because it depends on its relation to other representations: it could not have a conceptual role in isolation from other representations (remember *Holism* mentioned above). However, do we want to count conceptual role theories as *relationist*? It is not clear, because some conceptual role theories might want to characterize samethinking as *type-identity* (or, alternatively, as *similarity*) between representations individuated in terms of their conceptual role. On the face of it, both versions would be a clear version of intrinsicalist model.

We could try to say the following. Even though the individuation of mental representations in conceptual role theories might involve an external relation (in the sense that it involves other representations), the fact that a representation token is of a certain type is an intrinsic representational feature, even if the type happens to be individuated through an external relation. This is because the term "conceptual role" can be used to refer to the (putative) categorical basis of the disposition to interact with other representations in certain predicable ways. In other words, one might take the property of having a conceptual role to be *grounded* in some intrinsic property.[12] As an upshot, we get a construal of conceptual role theories according to which they are not relationist, as desired.

But we have a similar problem with respect to other views for which this response does not seem available. For example, this characterization of "intrinsic" (resp. "extrinsic") might suggest that e.g. *social externalism* is a relationist theory. Social externalism tells us, roughly, that the meaning of an expression e as used by a speaker A depends on how e is used in A's community. On the face of it, this makes the individuation of meanings relationist, because the individuation involves an explicit connection to other speakers' uses in the speaker environment. Now,

[12] I am indebted to Michael Murez.

one might have thought that social externalism was a particular theory of content, subsumed under the family of views I call "Semantic Idendity models".[13] Similar considerations apply to *Originalism*. Roughly, originalism says that two mental occurrences are uses of the same concept just in case they stem from the same originating use. This seems to be an external relation as defined, because of the reference to another use. However, originalists seem to be an instance of identity model, as they explain samethinking in terms of concept identity. Moreover, note that the type of response I have mentioned with respect to conceptual role theories does not seem to be available when it comes to theories like social externalism, or originalism. Real world connections between agents and their representations seem to be part of what grounds the relevant representional properties, on such views. Perhaps originalism and social externalism are, after all, relationist views.[14] But this is not how these views are standardly characterized, and understood. So either the proposed characterization makes bad classificatory predictions, or originalist and social-externalist metasemanticians are in fact relationalists (despite the way they seem to characterize themselves). This was my remarks on the definition of relationalism. I will now say a word on the alternative decision pathway in the solution space, where some comments are in order for the diagram to be readable. (I will go back to Relationism in the next section, when I locate the model suggested in this thesis within the solution space).

Similarity vs Identity models Why is *Syntactic Atomism or Molecularism* a terminal node belonging to *Identity* models, whereas *Syntactic Holism* is a terminal node belonging to *Similarity* models? As already alluded to, holism is often thought to be incompatible with Shareability (Fodor & Lepore 1992, Fodor 1998, Schneider 2011). When Holism is not combined with the *Eliminativist* claim that samethinking *would require* identity if it existed, Holism belongs to *Similarity* models. Now, both atomism (the view that concepts are individuated independently of one another/are not individuated by their inferential role) and molecularism (very schematically, the view that concepts are individuated independently of some but not all other concepts) are both compatible with Shareability.[15] Observe the *shared sub-tree* between Similarity Models and Identity Models. Whether one chooses *similarity* or *identity*, one is faced with the same question, namely, Is samethinking a relation between contents? Semantic approaches answer 'Yes'. Syntactic approaches answer 'No'. Syntactic Atomism or Molecularism implies the *Language of Thought Hypothesis* (LOTH), which says (very roughly) that cognitive processes involve mental representations equipped with logical structure (in particular, Boolean connectives and

[13] In fact, I defend this classification—see Fig 6.1.

[14] In effect, the reader will notice strong similarities with the structure of Perry's and Schroeter's views (examined in chapter 4 & 5 respectively), and the structure of originalism and social externalism.

[15] Atomistic LOTH is not well suited to account for samethinking in the interpersonal domain, however. In fact, this is one of the reasons for the growing interest in relational models of samethinking. The reason is that there is no straightforward syntactic intersubjective individuation method for typing thoughts across agents (see e.g. Aydede 1998). Given this, a possible reaction is to combine a syntactic criterion of samethinking in the intrapersonal domain with a semantic, formal or epistemic criterion of samethinking in the interpersonal domain. As already mentioned, Cumming 2013b is an instance of this kind. But the framework defended in this dissertation belongs to this kind as well.

quantification).[16] In contrast, Syntactic Holism is really a mixed bag. It is compatible, but does not imply, the LOTH (see e.g. Schneider 2011). For example, Churchland (1998) is an instance of Syntactic Holism but without LOTH. I will end my gloss on the delineated solution space by saying a few words on the (pseudo) terminal node "Theories of content". Obviously, it is not the case that this node cannot be divided any further. On the contrary, it includes a host of various theories of content, such as descriptivism, or 2D semantic theories[17]

I have presented a method for delineating the solution space of the samethinking problem, discussed definitional issues, and reviewed important decision pathways in this solution space. I will now situate the suggestions made in this thesis within this solution space.

2 Formal intransitive relationism

In this dissertation, I have suggested that identity models rely on a misleading idealization. That makes the conception suggested in this dissertation a Relationist model. But my main interlocutors have been the Equivalence Class Fregeans (in particular Dickie & Rattan 2010, Prosser 2019, Onofri 2018, Cumming 2013a, 2013b, Fiengo & May 2006, Schroeter 2012). Let me characterize the suggested model of this dissertation in light of the classificatory decisions introduced earlier, by focussing on the Relationist subtree (Figure 6.2).[18]

In Chapter 3, I defended both (i) that the alignment relation is not a legitimate constraint on successful communication and attitude reporting, and (ii) that the alignment relation *is* a necessary constraint on any relational criterion in order to satisfy both Frege's Constraint and Shareability. Figure 6.3 is a roadmap which provides the reader with a bird's eye view of the argument deployed against the constraint of alignment. In this dissertation, however, I have assumed that Frege's constraint was a constraint on the individuation of thoughts.[19] The conjunction of the preceding claims entails a rejection of Shareability. It is shown that an alternative model is viable, in which agents samethink without sharing thoughts. As a matter of fact, Pragmalignment, the pragmatic version in place of alignment defined in chapter 3 & 4, is not a transitive relation, and therefore does not support Shareability. That makes the model a Formal Intransitive Relationist one.

How does the IB-joint attentional criterion of communicative success defended in chapter 2

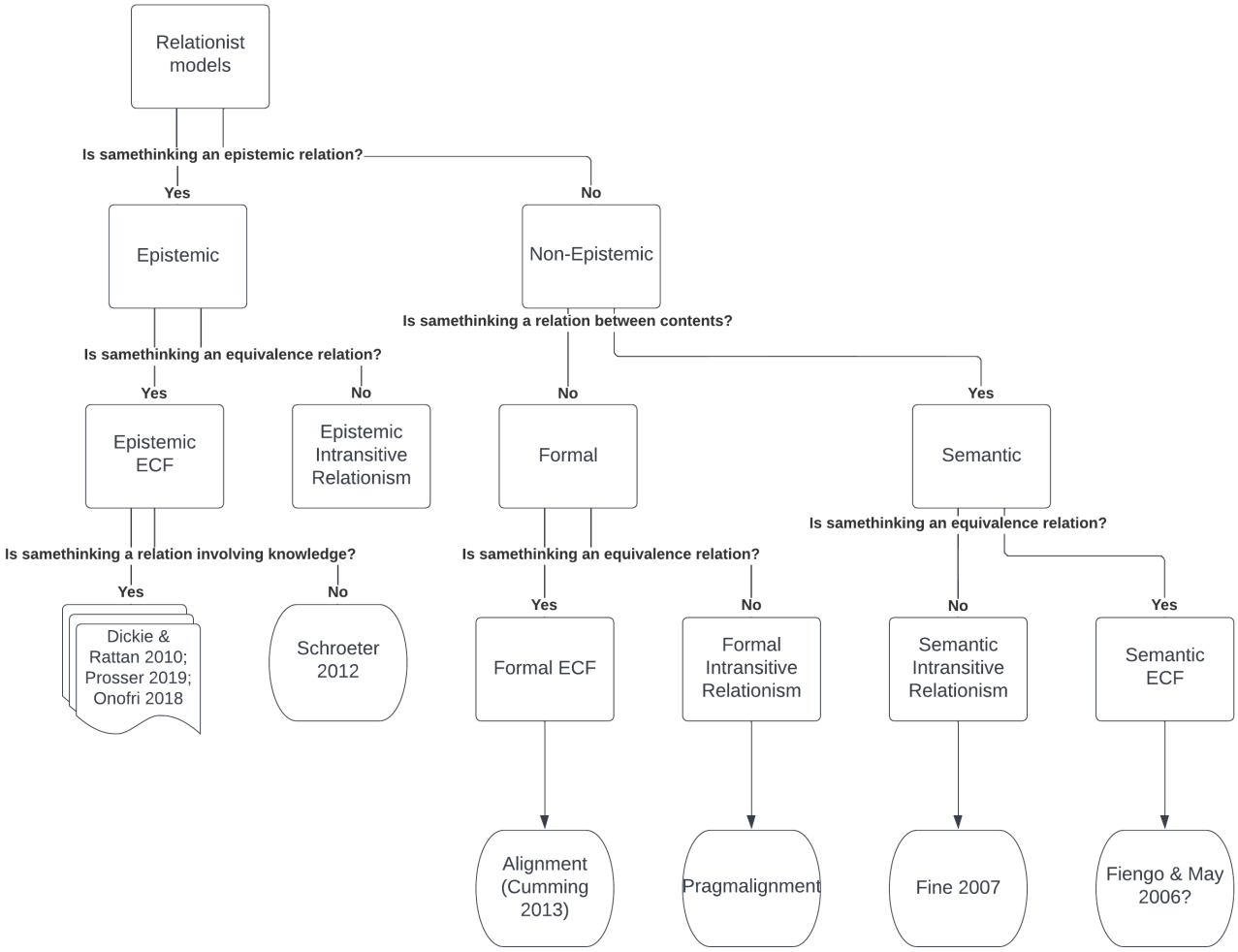
[16] Rescorla 2019, Piantadosi et al 2016.

[17] Main contemporary descriptivist proponents include Jackson 1998, 2010, Peacocke 1998, 2008, Chalmers 2002, 2006, 2011, 2012. On two-dimensional theories of content see Stalnaker 1978, Chalmers' aforementioned articles and e.g. the edited collection Garcia-Carpintero & Josep Macià 2006.

[18] Although Fiengo & May present their view as syntactic, my reason for considering it as a semantic view is that they have a very rich and unorthodox conception of syntax, which makes their *explanans* relation a semantic relation by standard criteria. See also the importance of what they call 'Assignments' (a semantic notion) for the characterization of their proposed *explanans* relation, as discussed in chapter 5.

[19] See Almotahari & Gray (forth), Gray 2022 and Speaks 2013 for discussion.

Figure 6.2 – Relationist models



relate to Pragmalignment? First, observe that the former is compatible with the latter. In fact, I have proposed the IB-joint attentional criterion as an alternative to ECF (of the epistemic variety)—and I have proposed pragmalignment as an alternative to ECF (of the formal variety). Moreover, there is a linguistic theory that, if true, supports both the IB-criterion and Pragmalignment (or so I argued).

This is the *Givenness Hierarchy* theory (GH), according to which speakers make implicit assumptions about the degree of activation of the representation of the intended referent in the minds of their interlocutors. This theory supports the existence of an interesting class of *inference-based* features attached to linguistic expressions (such as pronouns or determiners) that indicate whether a referent is present in the common ground and its degree of accessibility in the memory/attentional states in the hearer's mind—as assumed by the speaker. Therefore, (GH) supports the notion that discourse participants can have common knowledge that the hearer

Figure 6.3 – Argument against *alignment* as a background condition for samethinking

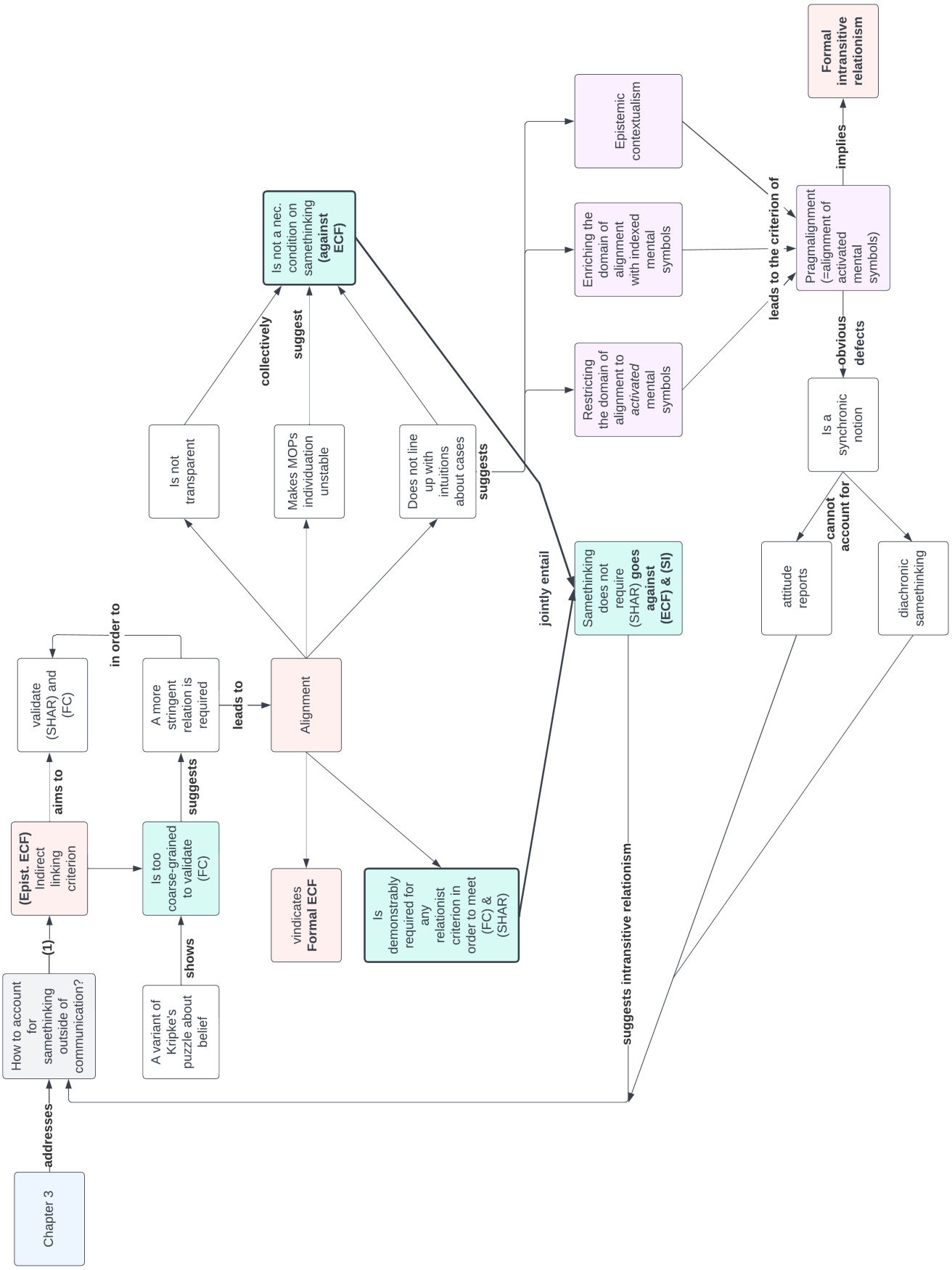
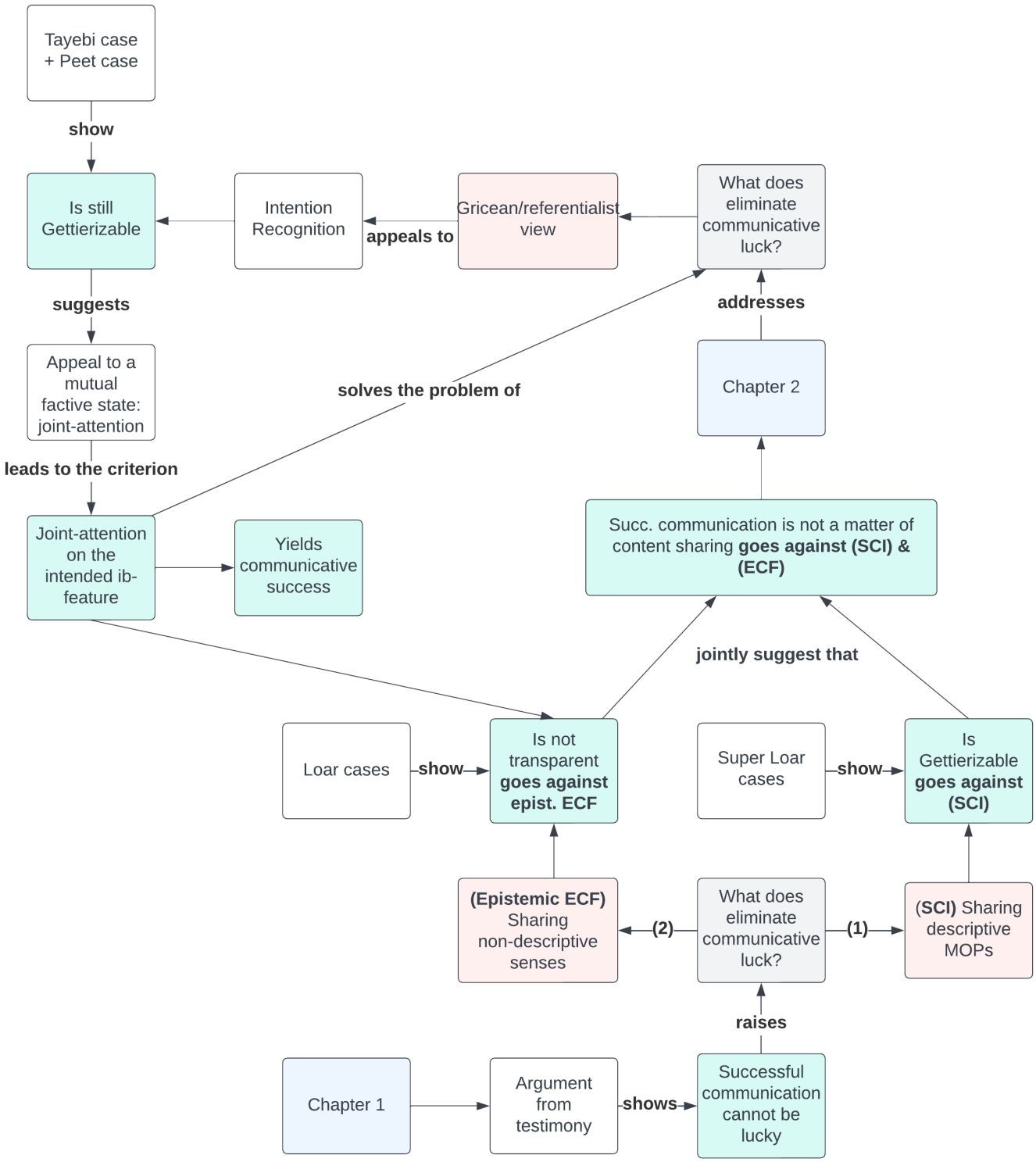


Figure 6.4 – IB-joint attentional criterion as an alternative to identity models & Epistemic ECF



is recovering the correct interpretation as a result of jointly attending to these IB-features.[20] But because (GH) suggests that the cognitive statuses (as assumed by the speech participants) play a role in establishing the common ground in communication, it supports (by the same token) the idea behind Pragmalignment. Figure 6.4 is a visual synopsis of the argument against identity models and Epistemic ECF that leads to the IB-joint attentional criterion.

The discussion in this thesis may have seemed preoccupied with mere technical details. For example, I raised issues such as whether it is legitimate to distinguish a level of *sense* understood as equivalence classes of modes of presentation; whether *alignment* is a background condition for successful communication and reporting; whether we could make sense of a an intransitive *same-use* criterion between expressions from different idiolects; or, in what way the standards for communicative success might be context-sensitive, etc. However, that impression is mistaken. Samethinking is a basic and pervasive phenomenon, integral to a range of social, linguistic and cognitive phenomena. As a result, investigating the nature of samethinking has important theoretical ramifications. Let me mention a few examples.

Issues related to samethinking are important for understanding the nature of *disagreement*, including metalinguistic disagreement, and central epistemological issues related to the recognition of disagreement.[21] *Metalinguistic disagreements* are disagreements about how an expression is, or should be used. The term *metalinguistic negotiation* is sometimes used to refer to the latter kind of metalinguistic dispute (Plunkett & Sundell 2013, 2021). For example, people involved in a dispute about whether Bitcoin is money might be communicating their diverging views about how "money" should be used.

I have defined samethinking in part in terms of its involvement in genuine disagreements. But what about metalinguistic negotiations? Under which conditions can thinkers have *genuine* disagreements while being engaged in metalinguistic negotiations about which concept is expressed by the expressions at issue? How can we tell when a metalinguistic disagreement is somehow substantive? Are there any theories of samethinking in the solution space that are better suited than other theories of samethinking to account for this?

Relatedly, a theory of samethinking might help providing criteria for when and why attitudes share or do not share a subject matter. Hence it is potentially relevant in the many places in philosophy where the metasemantic notion of topic continuity/stability plays a central role. For example, in normative philosophical areas such as aestetics and ethics; debates surrounding the semantics/pragmatics interface, or the metaphysics of free will.

Issues about samethinking bears on *social epistemology* as well. Semantic deference is related to

[20] Féry and Krifka 2008 (cited in Hedberg 2013) call these features "Givenness features".
[21] See Bryan & Frances 2019 for an overview on the epistemology of disagreement.

epistemic deference. (I have defended a particular view of this relation in chapter 5, but other views are possible). Therefore, theorizing about samethinking sheds light on the metasemantic infrastructure behind epistemic deference. It bears on the epistemology of collective agents, and the social dimension of collective knowledge. This includes scientific knowledge.[22] More generally, the philosophy of samethinking bears on the interaction between language and social reality. Importantly, a theory of samethinking seeks to tell us when communication is successful, as such it is a chapter of the *epistemology of testimony*.

In the next section, I outline some implications of the ideas presented in this thesis for other debates in disciplines related to the philosophy of language and mind. I indicate two lines of research I am particularly interested in, and which I think are worth pursuing in order to further develop the ideas presented in this dissertation.

3 Directions for future research

3.1 Theories of samethinking & cultural epidemiology

How do theories of samethinking and theories of *how representations become cultural* (i.e. widely distributed) relate? Samethinking is the psychosocial foundation of the creation and transmission of collective knowledge and cumulative culture. Therefore, one may think of the philosophy of samethinking as a chapter of foundational cognitive anthropology. As I see it, the philosophical investigation on samethinking has essentially the same objects of study as *cultural epidemiology*, but with a method, theoretical angle and ingredients at a much higher level of abstraction (Sperber 1996, Morin 2015, Buskell 2017, Heintz 2018, 2019).[23] Now, does the best explanation of *cultural transmission* favor certain theories of *samethinking* in the solution space over others—and if so, which ones?

It turns out that the framework suggested in this dissertation interestingly converges with a recent and influential model of cultural transmission. I have in mind the *Cultural Attraction Theory*. According to this framework, very roughly, representations are diffused by being *altered*. A major proponent is Morin (2015). In his work, Morin wants to explain how causal historical chains stabilize the representations they spread, thus accounting for cultural stability (i.e. traditions). We can identify (as a first pass) the representational traditions presented in chapter 5 as an instance of the causal historical chains Morin deals with. Morin emphasizes that the disposition to copy faithfully does not explain how representational variants change in their distribution and form over time. Cultural representations, he argues, are more likely to be the result of repeated *transformations*.

[22] See e.g. Zollman (2007, 2013) for different models of communication in scientific communities, and an exploration of their epistemic consequences.

[23] Sperber has a foot in both approaches, see e.g. Sperber 1994, 1997. I believe that the interface between the philosophy of samethinking and cultural epidemiology is fertile and needs to be explored. I did not explore the interface between the two disciplines as I would have liked in this dissertation. I hope to do so in future work.

What is the convergence I am talking about? As I see it, it is the following. On the one hand, according to the cultural attraction theory, stabilized representations are not the result of faithful copying. Rather, representations are altered during the process of acquisition. On the other hand, on the view of samethinking I have defended in this thesis, communication is not the replication of thoughts. Now, the fact that communication is not the replication of thoughts would explain at least *some* of the transformations involved in cultural transmission.

I began this dissertation with an evocation of Frege's picture. The fact that "mankind possess a common treasure of thoughts which is transmitted from generation to generation" (Frege 1892: 188 in Martinich 1996) implies, thought Frege, that communication is the transmission of content (since transmission of thoughts passes for the most part through communication). Now we can see that this picture might be bad cultural epidemiology. Work is needed to connect the philosophy of samethinking with cultural epidemiology in detail, and explore what may be a theoretically fertile interface. This is a project I hope to continue to pursue in further research. This brings me to the second theoretical area that I believe it would be interesting to explore in order to develop the material in this thesis.

3.2 Theories of samethinking & the nature of human–AI communication

Is successful communication between humans and dialog systems the same phenomenon as successful communication between humans? What would it take for a dialog system to samethink with human conversational partners? Under which conditions could a machine be said to *understand* a speech act? More generally, what does it take for a silicon-based representational system and a human to have their semantic representations in the samethinking relation? In particular, could a neural language model samethink with a human mind, and why? As these questions hopefully illustrate, theories of samethinking have implications for debates about the nature of human-artificial intelligences (AI) semantic interactions. Are there any samethinking theories in the solution space that are better able than others to shed light on this question?

Samethinking is a basic, in part pre-theoretical, notion. We need to keep its *explanans* broad and flexible in order to account for putative samethinking between humans and AI. For example, we might not want to say that we *share* thoughts or mental representations with AI. However, even if we do not share mental representations or thoughts with AI, it is plausible that their representational states could be in the samethinking relation to our mental states. If an adequate theory of samethinking allows for samethinking without thought sharing, all other things being equal this is an advantage when it comes to theorizing human-AI semantic interactions. Therefore, I speculate that *Relationism*/rejecting (SHAR) adds flexibility which turns out to be useful when it comes to theorizing about human-machine semantic interactions. In this respect, we can also speculate that non-epistemic relationalist models seem to have an advantage over models that centrally feature unreduced epistemic or phenomenological notions.

Of course, the suggestions I made in this thesis seem far from being applicable to human-AI semantic interaction as is. For example, a main notion I introduced (pragmalignment) involves cognitive statuses of mental representations—a notion that has no straightforward counterpart in AI. Similarly, I have repeatedly emphasized the meta-representational dimension of samethinking—a capability whose implementation in AI raises many interesting issues. We may call *Human-AI pragmatics*, the domain that investigates the conditions of *AI speech acts understanding*. Delineating this theoretical domain and its agenda is a project I hope to undertake in further research.

Another promising line of research consists in providing functional characterizations of intentional and epistemic notions involved in theories of samethinking, with the aim of accounting for human–AI samethinking. Relatedly, a promising task is to methodically examine the potential of different models of samethinking in the solution space to be applied to human-AI semantic interactions, a project I hope to undertake in further research too.

I have outlined how theories of samethinking, and in particular the ideas presented in this thesis, can connect with other debates in philosophy of language & mind and related disciplines. In the final section of this conclusion, I highlight an important distinction that has emerged from this work between two notions not clearly distinguished in the literature.

4 Coda: $\text{coordination}_{\text{int}}$ and $\text{coordination}_{\text{ext}}$

In the literature, the relation that underlies successful communication, genuine agreement and disagreement, is often called *coordination* or *coreference de jure*, and presented as "rational engagement" or in terms of "trading on identity" (Dickie & Rattan 2010, Schroeter 2012, Prosser 2019, Recanati 2016, Gray 2017). Let me quote relevant passages. Here is Prosser:

> Transparent communication [...] involves trading on identity and requires no interpretive premises. Transparent communication is analogous to [an argument whose validity depends on the possibility to trade on identity] and involves shared MOPs. (Prosser 2019: 10)

And here is Schroeter:

> The default interpretation generated by automatic linguistic parsing is to simply take others' words at face value—as samesaying *de jure* with your own use of those words. (Schroeter 2012: 6)

Here is Dickie & Rattan:

> Consider a situation in which you and I are using a demonstrative to refer to an object to which we are jointly attending. You say:

> That is $F_{(\text{at } t_1)}$
>
> I says:
>
> That is $G_{(\text{at } t_1)}$
>
> Each of us hears and understands what the other says. So either of us would be *warranted* in moving to the conclusion:
>
> Something is both $F_{(\text{at } t_1)}$ and $G_{(\text{at } t_1)}$.
>
> In this case, (. . .) my use of 'that' and your use of 'that' must share a sense. It is because the tokens of 'that' share a sense that (assuming that each of our initial utterances expresses knowledge) we can use each other's claims to extend what we know without going through the extra step of establishing that the same object is in question for both speakers. (Dickie and Rattan 2010: 147; italics mine)

Lastly, here is Gray:

> When two singular terms are coordinated, they represent their referential content "from the same perspective". In an informal sense, the coordination of two singular terms reflects the representational *presupposition* that they share referential content. (. . .) Inferential transitions whose truth-preserving character depends on the identity of referential content across particular positions in their premises require that those positions are coordinated. (Gray 2019: 3)

To a first approximation, it seems that all these authors are targeting the same notion, namely, speakers' disposition to *trade on identity*, disposition which makes coreference in some sense non-accidental, or "de jure". Authors who invoke this notion often explicitly connect it with the transmission of knowledge in communication, as the relation that makes speech participants' thoughts non-accidentally coreferential.[24]

But there are *two* notions of "being warranted to trade on identity" or "making coreference non-accidental" in the vicinity. The authors I have quoted do not all speak of the same notion. One notion is externalist, and non-transparent. Let's call it $\text{coordination}_{\text{ext}}$. The other notion is internalist, and transparent. Let's call it $\text{coordination}_{\text{int}}$.[25] $\text{Coordination}_{\text{ext}}$ and $\text{coordination}_{\text{int}}$ are playing different roles. To see this, let's return to the Loar case. In the Loar case, there is

[24] For a radically non-epistemic approach to coordination, see Pryor 2017. I have not discussed this type of approach in this thesis, particularly because I have been concerned with the interpersonal domain, in particular, communication. But that is hopefully for another occasion. See also Simchen (2017) for an interesting critic of what he calls the *interpretationist* type of metasemantics. The latter is closely related to the approach I have engaged with in this thesis.

[25] Recall my reconstruction of D&R's argument for shared senses as equivalence classes of MOPs, where (I argued in chapter 1) they are actually equivocating between the two notions:

> **(1)** On the Moderate Fregean view, the contrast between an intersubjective situation in which there is rational engagement, and one in which there is no rational engagement, is *not* marked in terms of sameness and difference in sense (by definition of the Moderate Fregean view)

coordination$_{int}$, because the speaker and hearer are mutually presupposing that the object the speaker is thinking about is the object the hearer is thinking about. That is, speaker and hearer are trading on identity of each other's thoughts. Still, communication fails. That means that, although there is an important sense in which coordination$_{int}$ makes coreference non-accidental (in favorable cases), coordination$_{int}$ is not *sufficient* to eliminate communicative luck. In effect, it is a matter of luck that the speech participants' thoughts corefer in the Loar case.

In which sense does coordination$_{int}$ contribute to make coreference non-accidental, then? Roughly, to put it in the terms of this thesis, coordination$_{int}$ is to samethinking what epistemic justification is to knowledge (where justification is construed in such a way that it is possible for a person to be justified in believing a proposition that is in fact false; and where samethinking is construed as the relation, whatever it is, that explains successful communication). So, coordination$_{int}$ does not *guarantee* samethinking. But coordination$_{int}$ makes coreference non-accidental in the same way that justification is what distinguishes a justified true belief (JTB) from a "luckily true belief" (LTB).

In contrast, I will think of coordination$_{ext}$ as the relation, whatever it is, that is required in addition to coordination$_{int}$ and coreference to output samethinking in communication. Hence coordination$_{ext}$ is akin to the notion of externalist "warrant", or "epistemic entitlement", namely, the notion used by epistemologists for the ingredient, whatever it is, that makes one have knowledge in addition to JTB and eliminates Gettier cases.[26] Coordination$_{ext}$ is thus akin to the notion of "warrant" used by epistemologists for the ingredient, whatever it is, that makes one have knowledge in addition to JTB and eliminates Gettier cases. This notion is often characterized as an externalist type of epistemic justification, in the sense that whether an agent is "warranted" is determined by facts that are independent of the reasoning abilities he or she may or may not have, and that the agent need not be able to recognize. In other words, Coordination$_{ext}$, being an externalist relation, fails to be transparent (as illustrated in the Loar case).

There are several ways to characterize transparency (Wikforss 2015). One way involves the notion of knowledge and is thus concerned with *epistemic access*. On this way of characterizing transparency, the non-transparency of coordination$_{ext}$ reads: it is not the case that, if coordination$_{ext}$ does not obtain, then the speech participants are able to have common knowl-

(2) But difference in sense can explain why there isn't rational engagement in the intrasubjective domain (i.e. in Frege cases) iff rational engagement is explained in terms of sameness of sense in the intersubjective domain (Thesis of the excessive focus on the '*multiplying role* at the expense of the *consolidating role*')

(3) Therefore **(a)** the Moderate Fregean view is unstable, and **(b)** there is intersubjective rational engagement iff there is shared sense.

As far as this reconstruction of their argument is correct, in the premiss 2, D&R are equivocating between the two different construals of 'rational engagement' I have pointed out: the first occurrence expresses coordination$_{int}$ whereas the second occurrence expresses coordination$_{ext}$. See chapter 1 and end of chapter 2 where I put forward the distinction.

[26] "Warrant" is not necessarily a factive notion, if we define it as what it is enough to add to the *true* justified belief to give knowledge: it is not necessarily from the "warrant" that the factivity of knowledge comes.

edge (or, more weakly: are both able to know) that coordination$_{ext}$ does not obtain. Another way of characterizing transparency involves *functional role*. On this way of characterizing transparency, the non-transparency of coordination$_{ext}$ can be expressed along the following lines (I am drawing on Murez 2022): it is not the case that if all speech participant reason as if their thoughts are the same (/are about the same object), then their thoughts are the same (/are about the same object).

To recap, coordination$_{ext}$ is the relation that is *uninstantiated* in the Loar case (so there is no samethinking), and coordination$_{int}$ is the relation that is *instantiated* in the Loar case (so there is a mutual presupposition of coreference).

In order to better understand these notions and their relationship, it would be interesting to know whether samethinking should be understood in terms of coordination$_{int}$ or in terms of coordination$_{ext}$. More generally, the issue whether coordination is factive is part of the debate about the primacy of notions that imply *success*, such as knowledge, or veridical perception, as opposed to notions such as belief or hallucination. The question whether coordination is factive is somewhat similar to that of whether belief should be understood from a (factive) notion of knowledge. Moreover, some authors, in particular Fine (2007, 2010), and Lawlor (2010) have a debate about the (non-)factivity of coordination (see Recanati 2016, chapter 2 for an overview). On the face of it, if one says coreference is factive (Fine), and the other says it is not (Lawlor), one of them must be wrong. It would be interesting to know which one.

One reason to understand samethinking in terms of coordination$_{int}$ has to do with the putative role of coordination$_{int}$ in successful communication, and (dis)agreement—in the terminology of this dissertation, coordination$_{int}$ is often thought to explain *samethinking* (as the passages quoted above illustrate). Let us focus on communication. There is an intimate connection between communication and (the transmission of) *knowledge*. In chapter 1, I argued that successful communication involves the knowledge of what was said on the hearer's part. Knowledge is generally assumed to be factive in epistemology. But coordination is assumed to play a central role in the analysis of successful communication.

Should samethinking be defined in terms of coordination$_{ext}$ or in terms of coordination$_{int}$, then?

My answer is that we need both notions anyway. We need coordination$_{int}$ insofar as we need to describe the Loar cases as cases where participants mutually trade on coreference of each other's thoughts (just like Gettier cases are cases where a true belief *is* justified). As Lawlor (2010) puts it,

> What we want is to interpret confused utterances in such a way that we can see how, first, a confused utterance is not just a crazy or unintelligible utterance; and second, the subject's reasoning with the proposition(s) expressed by confused utterances is

often good reasoning. (Lawlor 2010: 489)

But we need coordination$_{ext}$ to explain why communication fails in the Loar case.[27] Coordination$_{int}$ is necessary, but not sufficient, for samethinking. Coordination$_{int}$ is arguably transparent, but not factive; Coordination$_{ext}$ is factive, but not transparent. On the one hand, we need to explain psychology and behavior. On the other hand, we need to explain success and knowledge of what is said. No single relation plays both roles at the same time. Therefore, I recommend separating two types of coordination, one of which is externalist and non-transparent, but explains success and eliminates luck; while the other is cognitively significant, internalist and (arguably) transparent.

[27] I have proposed a particular view of what coordination$_{ext}$ consists in chapter 2 in terms of joint-attention on the intended ib-features.

7 Résumé substantiel en français

Cette thèse de doctorat en philosophie a pour thème les ressorts de la compréhension mutuelle dans la communication singulière (portant sur des objets particuliers), et la nature du partage des représentations entre les agents. Elle part de l'hypothèse qu'il existe une relation plus stricte que la coréférence entre les représentations mentales des agents, impliquée dans la communication réussie, l'accord, le désaccord, et les rapports d'attitude *de dicto* (où deux représentations coréfèrent si et seulement si elles réfèrent au même objet). J'appelle cette relation putative *samethinking*. Le problème central de ce travail doctoral est de spécifier les caractéristiques et l'explication de cette relation. Classiquement, la relation de samethinking a été conçue comme une relation de similarité entre les pensées. Sous l'influence d'un argument dû à Fodor & Lepore (1992) — selon lequel les théories qui invoquent la similarité ont besoin d'une notion d'identité mais ne la définissent pas et ne sont donc pas viables — cette conception est moins populaire aujourd'hui. De nos jours, le samethinking est conçu majoritairement comme de l'identité entre les pensées. En fait, la plupart des théories du contenu sont basées (parfois de manière implicite) sur une compréhension du samethinking comme identité. En tant que telles, elles supposent simplement que les phrases utilisées en contexte ont un contenu qui est à la fois ce que le locuteur exprime et ce qu'un auditeur saisit dans la compréhension. Cependant, au cours des dix dernières années, la question du samethinking a fait l'objet d'une littérature florissante et en pleine évolution. L'espace des solutions au problème du samethinking s'est considérablement enrichi. Une nouvelle conception du samethinking comme relation externe s'est développée. On peut donner une idée de la notion de relation externe avec un exemple emprunté à Gray (2017). La relation d'être "*âme-sœur*" est une relation interne, car elle repose entièrement sur une adéquation entre les propriétés respectives des relata. Par contraste, la relation d'être *marié* est externe : elle dépend de si les relata sont dans une certaine relation légale (irréductible à leur propriétés respectives). La thèse peut être conçue comme un argument cumulatif pour l'idée que le samethinking est une relation *externe* et *intransitive*. Une explication particulière de cette relation est proposée en termes de réseaux intersubjectifs causaux-historiques de concepts, réseaux dont la relation de base est définie en termes d'attention conjointe et du statut psychologique des concepts en contexte.

Cette thèse est divisée en deux parties. La première partie traite du samethinking dans la communication. La seconde partie traite du samethinking en dehors de la communication, c'est-à-dire dans l'attribution des pensées et dans l'accord et le désaccord.

La PARTIE 1 — "Le samethinking dans la communication" — est constituée de deux chapitres. Le premier expose un problème qui doit être résolu et s'oppose à deux solutions proposées. Le second défend une solution alternative. Dans le CHAPITRE 1 — "Communication, contenu, et les cas de (Super)-Loar" — je soulève le problème suivant : dans quelles conditions les gens communiquent-ils avec succès, étant donné que ce n'est pas simplement en s'accordant sur le bon contenu référentiel ? Je clarifie en quoi consiste le succès communicationnel. En particulier, je déploie un argument selon lequel une communication réussie ne peut pas être chanceuse. Ensuite, j'explique pourquoi la conception selon laquelle le succès communicationnel consiste à penser un contenu identique de la part du locuteur et de l'auditeur n'est pas satisfaisante. Puisque la coordination sur du contenu référentiel n'est pas suffisante, comme le montre le cas Loar présenté dans le chapitre (une variante communicationnelle des cas Frege — un cas Frege étant un cas dans lequel un penseur peut, sans irrationalité, attribuer au même individu (au même référent) des propriétés objectivement contradictoires, ou adopter des attitudes incompatibles envers lui), le contenu qui doit putativement être saisi pour le succès communicationnel est plus fin que la référence.

J'examine deux conceptions majeures de la communication comme transmission de contenu à grain fin. La première je l'appelle "la théorie frégéenne standard". Elle énonce que les participants au discours communiquent avec succès à propos d'un objet *o* ssi (en gros) ils déploient les mêmes modes de présentation descriptifs pour *o* dans la production et la compréhension de l'énoncé, respectivement. En m'appuyant sur Buchanan (2013) et Tayebi (2013), j'utilise des intuitions sur les cas pour montrer que cette condition n'est pas suffisante : il est toujours possible que les participants partagent le même contenu référentiel sous le même mode de présentation descriptif, mais qu'ils le fassent par chance. S'ensuit mon examen de la deuxième conception de la communication comme transmission de contenu à grain fin, que j'appelle "la théorie frégéenne sophistiquée". Cette théorie conçoit les modes de présentation comme non descriptifs, en particulier leur référence est déterminée par des relations causales avec l'environnement. De plus la théorie a une conception relationnelle des contenus partagés, selon laquelle un contenu partagé est une classe d'équivalence de mode de présentations (MOPs) non descriptifs convenablement reliés les uns aux autres dans une situation (où la relation pertinente est *externe*).

Parce que la relation de partage de mode de présentation est externe, les participants peuvent ne pas se rendre compte quand la relation pertinente n'a en fait pas lieu. J'exprime cela en disant que les contenus partageables postulés sont en partie opaques, ou non-transparents. En

partie parce que ces contenus postulés ne sont pas transparents, je soutiens qu'ils créent plus de problèmes qu'ils n'en résolvent. En particulier, ce niveau de contenu n'est pas explicatif de la communication réussie. La *pars destruens* peut être résumée dans les termes du dilemme suivant : soit l'identité de contenu est "gettierizable" (Gettier 1963), c'est-à-dire qu'on peut y arriver par chance (dans la conception frégéenne standard), soit la différence de contenu plus fin que la référence n'est pas transparente (dans la conception frégéenne sophistiquée). Ce dilemme nous donne des raisons de penser que nous ne devrions pas concevoir la communication réussie en termes de contenu partagé à grain fin. Le résultat de ce chapitre est que la condition qui élimine la chance quant à la coréférence dans la communication doit être comprise comme une condition causale et non comme une condition sémantique.

Dans le CHAPITRE 2, intitulé "De ce qui pourrait éliminer la chance communicationnelle" et qui est le chapitre central de la première partie, j'examine une autre solution candidate importante au problème et j'explique pourquoi elle est également inadéquate. Puis, en m'appuyant sur cette solution, je propose ma propre solution. Le chapitre déplace le centre de la discussion vers l'idée que la communication est une question de reconnaissance des intentions. Un thème central est l'idée que le plan référentiel d'un locuteur (à savoir, son plan pour que son auditoire pense à un certain objet) inclut typiquement l'intention que *certaines caractéristiques* de l'énoncé ou du contexte soient utilisées dans la façon dont l'auditeur reconnait ce sur quoi le locuteur a l'intention de communiquer. En m'inspirant de Buchanan (2013), j'incorpore cette idée dans la condition anti-chance suivante : l'auditeur doit interpréter l'énoncé du locuteur en vertu de l'attention portée à l'information que le locuteur souhaite que l'auditoire utilise afin de reconnaître le référent (j'appelle cette information "*ib-feature*", en suivant Schiffer (à paraître a/b)).

En m'appuyant sur la littérature, je présente deux cas (Tayebi 2013, Peet 2016) montrant que cette condition n'est pas une solution générale au problème. J'introduis ensuite l'attention conjointe (un phénomène centralement étudié en psychologie interpersonnelle et développementale) comme mécanisme de sécurité communicationnelle. Je distingue deux types de communication : *déictique* où l'objet dont on parle est présent et observable dans la situation de discours ; et *non-déictique*, où l'objet n'est pas présent ou pas observable dans la situation de discours. J'explique ensuite comment l'attention conjointe peut être utilisée pour analyser le succès communicationnel dans les deux types de communication. Le critère auquel je parviens est (en gros) le suivant : l'interprétation par l'auditeur de l'énoncé du locuteur est entièrement gouvernée par la bonne manière d'utiliser l'aspect du contexte ou de l'énoncé prévue par le locuteur comme base à la reconnaissance du référent, aspect sur lequel le locuteur et l'auditeur ont une conscience jointe, et comme résultat non-déviant de ce qui précède, l'auditeur reconnaît le bon référent.

L'idée qui sous-tend ce critère de la communication réussie est la suivante. L'attention conjointe fournit une sécurité coréférentielle parce qu'il s'agit d'un état mutuel *factif* — un état

dans lequel les participants au discours ne peuvent se trouver que s'ils ont réellement leur attention sur le même objet avec la conscience commune qu'ils le font. Lorsque cela se produit, le locuteur et l'auditeur réfèrent *ensemble*, pour ainsi dire. L'attention conjointe sur les aspects du contexte planifiés par le locuteur (les *ib-features*) fait que chaque élément d'information contextuelle utilisé dans l'interprétation de l'énoncé est non seulement mutuellement connu, mais (en gros) communément connu (où x est mutuellement connu parmi un ensemble d'agents si chaque agent connaît x ; alors que x est communément connu parmi un ensemble d'agents si x est mutuellement connu parmi cet ensemble d'agents, et il est mutuellement connu parmi cet ensemble d'agents que x est mutuellement connu parmi cet ensemble d'agents, et ainsi de suite à l'infini). Par conséquent, le locuteur et l'auditeur ont une connaissance commune que l'auditeur produit l'interprétation correcte, et la chance est éliminée. J'appelle ce critère "critère d'attention conjointe" du succès communicationnel (*ib-joint attentional criterion*). J'explique pourquoi ce critère est un premier pas vers une approche du *common ground* (c'est-à-dire, très brièvement, l'ensemble des propositions et des références supposées être déjà partagées entre les participants au discours) qui est moins intellectualiste que les conceptions dominantes en la matière. En conclusion, je propose quelques réflexions sur la question suivante : si l'approche par l'attention conjointe est sur la bonne voie, comment le *common ground* est-il établi dans la communication qui n'est *pas* face-à-face ? Enfin, je compare la solution proposée à la conception frégéenne sophistiquée examinée au chapitre 1.

La PARTIE 2 —"Le samethinking en dehors de la communication"— traite du problème suivant : Qu'est-ce que le samethinking entre des penseurs différents qui n'interagissent pas ? Cette partie, qui se compose de trois chapitres, procède de manière similaire à la première partie : elle considère différentes conceptions de la conception du samethinking hors communication comme identité de contenu, et explique pourquoi elles ne sont pas satisfaisantes (chapitre 3) ; puis elle défend progressivement un modèle alternatif (chapitres 4 & 5).

Le CHAPITRE 3 — "De l'alignement au pragmalignement (ou alignement pragmatique)" — examine la communication impliquant des noms propres comme pierre de touche pour l'examen des théories du samethinking en dehors de la communication. Ce chapitre fait donc la transition entre les deux parties de la thèse, et en constitue une pièce maîtresse. Comment la communication impliquant des noms propres peut-elle nous conduire au samethinking en dehors de la communication ? Pour illustrer, si vous connaissez le nom "Napoléon", c'est parce qu'il vous a été transmis de manière communicationnelle. Le chemin de transmission s'origine dans une utilisation initiale du nom qui en établit la pratique d'utilisation. Tous les utilisateurs du nom "Napoléon" sont reliés entre eux par un tel chemin de transmission. J'observe qu'à première vue, l'appartenance au réseau semble garantir le partage d'un concept. Si une locutrice est compétente avec le nom "Napoléon", on peut dire qu'elle a une connaissance commune de Napoléon —et qu'elle partage le concept NAPOLÉON— avec tous les utilisateurs de "Napoléon" (c'est du moins ce que l'on est en droit de supposer de prime abord). En d'autres termes,

lorsqu'il s'agit de pensées impliquant un nom propre, la relation *même-concept-que* semble pouvoir se réduire à l'appartenance à un même chemin de transmission d'utilisation du nom.

Le chapitre commence par l'examen d'une théorie due à Onofri (2018) qui précise cette idée (une idée commune aux modèles causaux-historiques du samethinking). Je montre que l'appartenance à un même chemin communicationnel, lorsqu'elle est interprétée comme un critère relationnel de *partage des pensées*, est trop grossière pour rendre compte de la signification cognitive et de la transparence des pensées : un tel critère identifie des pensées qui sont différentes pour leurs penseurs.[1] Ce critère contrevient donc à la *Contrainte de transparence*, à savoir l'idée que le sujet doit être en mesure de déterminer *a priori*, par simple introspection, si les pensées qu'il forme et les concepts qui y figurent sont les mêmes, ou s'ils sont différents les uns des autres.[2] Il contrevient corrélativement à la *Contrainte de Frege* — un critère de différence pour les concepts qui incorpore la *Contrainte de transparence*.[3] Je propose ensuite une modification du critère d'Onofri (2018) qui résout techniquement le problème. J'explique que le critère qui en résulte est stipulatif : il semble exclure arbitrairement des agents des chaînes communicationnelles seulement afin de rétablir une compatibilité avec la *Contrainte de Frege*. Pour y remédier, il faut au minimum prouver que la clause stipulant l'exclusion des cas Frege des chaînes communicationnelles est nécessaire pour expliquer le succès communicationnel *lui-même*.

Cela nous conduit à la question suivante : un locuteur dans un cas Frege vis-à-vis d'un objet *o* peut-il communiquer avec succès à propos de *o* avec un interlocuteur qui n'est pas dans le cas Frege pertinent ? Une réponse *négative* à cette question impose une condition d'*alignement* sur la communication référentielle réussie. L'alignement s'obtient entre les répertoires conceptuels des agents si et seulement si (très brièvement) les dispositions communicationnelles des agents relient leurs concepts d'une manière *biunivoque*. Dans la deuxième partie du chapitre, à la suite de Cumming (2013a,b), je montre que l'alignement est nécessaire pour tout critère relationnel d'individuation des concepts afin de satisfaire à la fois la *Contrainte de Frege* et la *Partageabilité des concepts* (à savoir l'idée que les concepts sont partagés dans la communication, l'accord et

[1] Je distingue les chemins de transmission composés de liens qui conduisent à la création d'un *nouveau* concept chez l'auditeur, ceux composés de liens qui conduisent à l'ajout d'une *association* du concept avec un nom propre, et ceux composés de liens qui mènent à l'utilisation d'un concept *existant déjà labellisé* par le nom propre en cours d'utilisation dans le contexte. Seuls les deux premiers types de chemins sont des chemins de *transmission de la référence* à proprement parler.

[2] Contrairement à ce que peut laisser penser cette caractérisation informelle de la transparence, la transparence des concepts ne nécessite pas que le sujet soit capable de conceptualiser ses propres concepts ou de faire des jugements métaconceptuels à leur sujet. Il s'agit plutôt de capturer l'intuition qu'il est typiquement immédiatement évident pour un penseur qu'il pense en partie *la même chose* à chaque fois qu'il déploie le même concept, et qu'il est immédiatement évident qu'il pense *des choses différentes* à chaque fois qu'il déploie des concepts différents.

[3] Le nom "Contrainte de Frege" est dû à Schiffer (1990). Voici une formulation possible de cette contrainte directement inspirée de Schiffer :

> **Contrainte de Frege :** Si un sujet minimalement rationnel *S* croit simultanément d'un certain objet *o* qu'il est *F* et qu'il n'est pas *F*, alors *S* pense à *o* moyennant deux concepts distincts.

le désaccord).[4] Par conséquent, la question susmentionnée a un statut crucial, car elle décide si les pensées sont partageables. C'est un autre aspect important pour lequel je considère la communication comme pierre de touche pour l'examen des théories sur le samethinking *simpliciter* dans ce chapitre. Rejeter l'alignement comme condition nécessaire de la communication réussie revient *ipso facto* à rejeter la *Partageabilité* des pensées et des concepts qui y figurent.

Le reste du chapitre présente une série d'arguments contre l'alignement comme condition nécessaire à la communication réussie. Premièrement, je soutiens que l'alignement n'est pas transparent, donc le contenu partagé basé sur l'alignement n'est pas transparent. Deuxièmement, et de manière connexe, je soutiens que l'alignement rend instable l'individuation des concepts. En particulier, l'alignement des concepts est relatif à des ensembles d'agents, mais l'individuation intrapersonnelle des concepts ne devrait pas l'être. Troisièmement, je soutiens que l'alignement produit des prédictions erronées. En particulier, je fais valoir des cas où les agents peuvent savoir ce qu'un locuteur *désaligné* a dit, et où la communication réussit. Dans ce genre de cas, *la coréférence est transparente* quand bien même le contenu partagé basé sur l'alignement des concepts ne l'est pas.

A des fins exploratoires, je critique le caractère nécessaire de l'alignement des concepts pour la communication par une autre voie (plus controversée). Je soutiens en particulier que *si* les normes de réussite de la communication sont sensibles au contexte, alors l'alignement n'est pas une condition nécessaire à la réussite de la communication. Je poursuis en défendant que les normes de réussite de la communication *sont* sensibles au contexte. En supposant que *savoir ce qui est dit* implique d'être capable d'exclure toutes les alternatives pertinentes, *quelles* alternatives sont pertinentes dépendent du contexte de la conversation. Je suggère deux conceptions spécifiques de cette sensibilité au contexte. La première conception que je propose est celle de l'empiètement pragmatique (*pragmatic encroachment*), selon lequel (en gros) les normes pour savoir ce qui est dit peuvent dépendre des coûts pratiques d'une mauvaise compréhension. Ma discussion culmine dans une tentative de donner une tournure pragmatique à la contrainte d'alignement, et qui constitue la deuxième conception de la sensibilité au contexte des normes de réussite communicationnelles que je propose, à savoir la théorie du *statut psychologique*. En m'appuyant sur la théorie linguistique de la Hiérarchie informationnelle (*Givenness Hierarchy*), j'observe que le statut cognitif d'un concept (c'est-à-dire, en gros, son degré d'accessibilité dans la mémoire et les états attentionnels des interlocuteurs — tel qu'il est assumé par le locuteur) joue un rôle important dans la communication linguistique. Selon (GH), chaque fois que les locuteurs utilisent des pronoms et des déterminants, ils font des suppositions implicites sur les

[4] Voici une formulation de la *Contrainte de Partageabilité* qui est souvent acceptée de manière implicite par les défenseurs du caractère partageable des concepts :

> **Contrainte de Partageabilité :** Si un penseur A communique de manière réussie la pensée singulière que *o* est *F* à un penseur B, ou bien si A et B sont en accord (ou en désaccord) authentique sur le fait que *o* est *F*, alors A et B partagent un même concept pour *o*.

statuts cognitifs que l'objet en discussion a dans l'esprit de leurs interlocuteurs (par exemple, Hedberg 2013, Féry & Krifka 2008). Ces statuts cognitifs aident à définir une notion de *pertinence* appliquée aux concepts : les concepts pertinents sont, selon moi, ceux qui présentent un certain degré d'accessibilité (à savoir les concepts activés dans la situation de discours). Cependant, la notion standard d'alignement est aveugle à la notion de pertinence ainsi définie, ce qui donne lieu (selon moi) à de mauvaises prédictions. Avec cette notion de pertinence appliquée aux concepts, je définis une version pragmatique de la contrainte d'alignement restreinte au domaine des concepts activés. J'appelle la contrainte résultante "pragmalignement" ou "alignement pragmatique" (*pragmalignment*), et j'illustre son fonctionnement. Je fais valoir que l'alignement pragmatique fait des prédictions plus intuitives sur les cas que la notion standard.

Après avoir soutenu que le domaine de la notion standard d'alignement était trop large, je suggère qu'il est également, dans un sens important, trop étroit. La représentation de la perspective d'un agent désaligné est, selon moi, un moyen tout à fait valable de se coordonner avec succès avec son auditeur dans le cadre d'une communication désalignée. Autrement dit, on a des raisons de vouloir intégrer les concepts métareprésentationnels des agents dans le domaine de la relation. Je propose une définition de l'alignement pragmatique qui intègre cette idée, en m'appuyant sur la théorie des fichiers mentaux et en particulier le fragment qui concerne les fichiers *indexés* (Recanati 2012, 2016). C'est la dernière transformation pragmatique de la contrainte d'alignement que j'explore dans ce chapitre. En conclusion, je souligne une forte limitation de l'alignement pragmatique ainsi défini : il s'agit d'une notion *synchronique*, et arrimée à des contextes particuliers. Par conséquent, en l'état, cette notion est incapable de rendre compte du samethinking dans l'attribution des pensées (comme lorsque je rapporte des croyances d'Aristote), ou dans l'accord et le désaccord entre des agents qui n'interagissent pas.

Le CHAPITRE 4 — "Le pragmalignement en action : les rapports d'attitudes et d'énoncés" — généralise la relation d'alignement pragmatique au samethinking diachronique et transcontextuel, comblant ainsi le vide signalé à la fin du chapitre 3. Une explication, sans contenu partagé, du samethinking dans les rapports d'attitudes et d'énoncés, et dans l'accord et le désaccord, est présentée et défendue. Pour ce faire, j'examine les réseaux de fichiers mentaux associés à l'utilisation de noms dans les chaînes causales-historiques, plus précisément la description qu'en fait Perry (2012). Perry les appelle des *réseaux de fichiers intersubjectifs*. Ce chapitre explique comment Perry définit une relation de samethinking sans alignement en termes de réseaux de fichiers. La solution de Perry implique un partitionnement supplémentaire du réseau — qu'il appelle *thread* — permettant d'identifier le fichier d'un agent impliqué dans un contexte discursif ou mental particulier, et la manière dont ce fichier est utilisé ou mis à jour dans ce contexte. Je souligne la convergence significative de la notion de *thread* de Perry avec la notion de pragmalignement — et la théorie du statut psychologique — introduites dans le chapitre précédent. J'utilise cette notion pour généraliser le pragmalignement aux rapports d'attitude diachroniques et contrefactuels. Lorsqu'ils rapportent les

attitudes d'un penseur dans un cas Frege vis-à-vis d'un certain objet, les attributeurs ont à l'esprit les manières particulières qu'a le penseur de penser à l'objet. Ce faisant, ils distinguent implicitement un fil (*thread*) dans le réseau de fichiers, qui explique la sensibilité des rapports d'énoncés et d'attitudes au statut psychologique de fichiers mentaux particuliers. Les concepts en relation de samethinking sont les concepts qui sont connectés le long d'un fil (*thread*) dans le réseau de fichiers (ou bien qui sont co-activés en contexte, comme défini au chapitre précédent).

Je mets en exergue un aspect à mon avis crucial de la théorie de Perry (qui utilise souvent une terminologie qui rend malheureusement cet aspect peu saillant) : dans les configurations *désalignées*, le partage de contenu relativement à un *thread* donné n'équivaut pas à une identité de contenu à grain fin, car (comme le montre le chapitre 3) les agents qui sont dans un cas Frege par rapport à un référent introduisent des informations parasites supplémentaires qui ne sont pas égalées par les agents désalignés. Perry est, je suggère, mieux interprété comme un relationniste intransitif (c'est-à-dire comme concevant le samethinking comme une relation externe intransitive, voir l'espace logique Fig 6.1). En capitalisant sur le chapitre précédent, je propose une caractérisation de la façon dont les locuteurs ciblent implicitement les fils (*threads*) dans les réseaux de fichiers intersubjectifs lorsqu'ils attribuent des pensées : ils le font, je suggère, en indexant des fichiers sur la perspective de l'attributaire. Je définis cette idée et j'illustre son fonctionnement sur l'énigme de Kripke concernant la croyance. La dernière partie du chapitre traite de l'accord et du désaccord sans interaction. J'oppose le *contextualiste modéré*, selon lequel (en gros) les questions d'accord et de désaccord sont entièrement décidées par les dispositions communicationnelles des agents, au *contextualiste radical*, selon lequel ces questions impliquent irréductiblement un interprète. Je suggère que Perry est commis à cette dernière position, et je propose quelques réflexions sur les coûts et les avantages de chacune de ces positions.

Le CHAPITRE 5 — "Participer aux traditions représentationnelles" — aborde le problème suivant. Si les concepts ne sont *pas* partagés, comment se fait-il qu'un locuteur puisse *apprendre*, *se tromper* ou avoir une connaissance *partielle* de la signification d'un mot ? Le chapitre commence par proposer une typologie de la distribution des concepts (c'est-à-dire l'étendue et la manière dont les concepts sont répandus dans une population, dans un sens théoriquement neutre de "répandus"), et situe les *significations* des mots dans cette typologie comme étant les concepts qui sont *largement répandus par la communication*. Le reste du chapitre propose une explication méta-sémantique pour rendre compte de la possibilité d'apprentissage, d'erreur et de saisie incomplète de la signification des mots, sans significations partagées autres que les extensions (c'est-à-dire référent, classe, propriété, etc.). L'explication méta-sémantique proposée repose sur deux hypothèses, tirées de Schroeter (2012). La première hypothèse est que l'utilisation des mêmes mots déclenche des *apparences de ressemblance sémantique* chez les utilisateurs de la langue. (Comme je le précise dans le chapitre, je ne veux pas impliquer que nous ne pouvons pas avoir une caractérisation *fonctionnelle* de ces apparences sémantiques mutuelles.) La deuxième hypothèse est que ces apparences sémantiques font en sorte que les choses se passent

comme si les significations étaient partagées, et donnent lieu à des traditions représentationnelles. Les locuteurs ont l'intention de conformer leurs utilisations des mots à ces traditions représentationnelles supposées — on parlera de *déférence* au sens technique de Putnam (1975).

Je propose une conception particulière de ce que sont ces traditions représentationnelles. En suivant Recanati (2016) et Schroeter (2012), je propose que ce qui sous-tend la déférence sémantique au niveau méta-sémantique sont des fichiers distribués de pair-à-pair (*peer-to-peer*) gérés au niveau de la communauté. Comment sont-ils gérés ? Je suggère que la façon dont les entrées encyclopédiques de *Wikipédia* sont gérées reflète assez bien les mécanismes sociaux par lesquels la communauté gère un fichier distribué, et j'en mentionne quelques-uns. Pourquoi les locuteurs s'engagent-ils à accorder leur usage lexical à leurs usages passés et à l'usage (supposé) de la communauté ? En m'inspirant de O'Madagain (2018), j'explore l'idée selon laquelle les gens défèrent en un sens sémantique afin de déférer en un sens épistémique : la recherche de connaissances sur des entrées encyclopédiques d'intérêt partagé est une raison importante pour utiliser des mots de manière déférentielle.

Dans la CONCLUSION GÉNÉRALE — "Qu'est-ce que le samethinking ?" — j'offre une taxonomie synoptique de l'espace logique des théories du samethinking. En profitant de la distance théorique fournie par le travail effectué dans cette thèse, ce chapitre s'attache à définir des classes générales dans lesquelles les diverses théories du samethinking pourraient s'inscrire, et discute des critères de décision d'appartenance à ces catégories générales en les rattachant à la littérature récente. Je situe ensuite le modèle que j'ai proposé dans cette thèse dans l'espace de solution ainsi délimité, et je tire quelques implications notables de ce modèle. J'indique ensuite deux lignes de recherche qui, selon moi, méritent d'être poursuivies afin de développer davantage le modèle proposé dans cette thèse.

Je conclus cette thèse en distinguant deux notions importantes qui, selon moi, n'ont pas été clairement distinguées dans la littérature. Une notion est externaliste, et non transparente. Je l'appelle la *coordination externe*. L'autre notion est internaliste, et transparente. Je l'appelle la *coordination interne*. Ces notions jouent des rôles différents. Je soutiens que nous avons besoin des deux notions. Nous avons besoin de la coordination interne dans la mesure où nous devons décrire les cas Loar comme des cas où les participants exploitent mutuellement, et de manière intelligible, la coréférence de la pensée de l'autre (tout comme les cas Gettier sont des cas où une croyance vraie *est* justifiée). Plus généralement, nous avons besoin de la coordination interne pour rendre compte de la psychologie et du comportement des agents. Mais nous avons besoin de la coordination externe pour expliquer pourquoi la communication échoue dans le cas de Loar, et plus généralement, pour expliquer le succès et la connaissance de ce qui est dit. Dans cette thèse, j'ai proposé une analyse particulière de la coordination externe dans la communication singulière en face-à-face, en termes d'attention conjointe à l'aspect du contexte que le locuteur veut que l'auditeur utilise pour reconnaître le référent.

Bibliography

Akhtar, N. and Gernsbacher, M. (2007). Joint attention and vocabulary development: A critical look. *Language and Linguistics Compass*, 1:195–207.

Almotahari, M. and Gray, A. (forthcoming). Frege cases and bad psychological laws. *Mind*.

Arnauld, A. and Nicole, P. (1662/1993). *La logique ou l'art de penser (ou "Logique de Port-Royal")*. Gallimard Collection Tel.

Asher, N. (1986). Belief in discourse representation theory. *Journal of Philosophical Logic*, 15(2):127–189.

Asher, N. (1987). A typology for attitude verbs and their anaphoric properties. *Linguistics and Philosophy*, 10(2):125–197.

Austin, J. L. (1975). *How to Do Things with Words: The William James Lectures Delivered in Harvard University in 1955*. Oxford University Press UK.

Aydede, M. (1998). Fodor on concepts and Frege's puzzles. *Pacific Philosophical Quarterly*, 79(4):289–294.

Bach, K. (1997). Do belief reports report beliefs? *Pacific Philosophical Quarterly*, 78(3):215–241.

Bach, K. (2006). What does it take to refer? In Lepore, E. and Smith, B., editors, *The Oxford Handbook of Philosophy of Language*, pages 516–554. Oxford University Press.

Bach, K. and Harnish, R. M. (1979). *Linguistic Communication and Speech Acts*. Cambridge, MA: MIT Press.

Bartlett, F. C. (1932/1995). *Remembering: A study in experimental and social psychology*. Cambridge university press.

Barwise, J. (1988). Three views of common knowledge. In *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 365–379, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Bayne, T. and Montague, M. (2011). *Cognitive Phenomenology*. Oxford University Press UK.

Berto, F. and Nolan, D. (2021). Hyperintensionality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

Bliss, R. and Trogdon, K. (2021). Metaphysical Grounding. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition.

Block, N. (1995). On a confusion about a function of consciousness. *Brain and Behavioral Sciences*, 18(2):227–247.

Blumberg, K. (2018). Counterfactual attitudes and the relational analysis. *Mind*, 127(506):521–546.

Bochner, G. (2010). Perry on reference and reflexive contents. *Language and Linguistics Compass*, 4(4):219–231.

Bochner, G. (2021). *Naming and Indexicality*. Cambridge University Press.

Boghossian, P. A. (1994). The transparency of mental content. *Philosophical Perspectives*, 8:33–50.

Bourdoncle, R. (2022). Shareability of thought and Frege's constraint: A reply to Onofri. *Inquiry: An Interdisciplinary Journal of Philosophy*.

Bourget, D. and Mendelovici, A. (2019). Phenomenal Intentionality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition.

Braun, D. (1998). Understanding belief reports. *Philosophical Review*, 107(4):555–595.

Braun, D. (2017). Indexicals. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2017 edition.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Brigandt, I. (2010). The epistemic goal of a concept: Accounting for the rationality of semantic change and variation. *Synthese*, 177(1):19–40.

Brigandt, I. (2013). A critique of David Chalmers' and Frank Jackson's account of concepts. *ProtoSociology*, 30:63–88.

Brooks, R. and Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of child language*, 35(1):207–220.

Bruner, J. (1983). Child's talk: Learning to use language. *Child Language Teaching and Therapy*, 1(1):111–114.

Buchanan, R. (2013). Reference, understanding, and communication. *Australasian Journal of Philosophy*, (1):1–16.

Burge, T. (1979a). Individualism and the mental. *Midwest Studies in Philosophy*, 4(1):73–122.

Burge, T. (1979b). Sinning against Frege. *Philosophical Review*, 88(3):398–432.

Burge, T. (1983). Russell's problem and intentional identity. In Tomberlin, J. E., editor, *Agent, Language and the Structure of the World: Essays Presented to Hector-Neri Casteneda, with His Replies*, pages 79–110. Hackett, Indianapolis, IN.

Burge, T. (1993). Content preservation. *Philosophical Review*, 102(4):457–488.

Burge, T. (1998). Memory and self-knowledge. In Ludlow, P. and Martin, N., editors, *Externalism and Self-Knowledge*. CSLI Publications.

Burgess, A. and Sherman, B. (2014). *Metasemantics: New Essays on the Foundations of Meaning*. Oxford University Press.

Buskell, A. (2017). What are cultural attractors? *Biology and Philosophy*, 32(3):377–394.

Bybee, J. and Beckner, C. (2015). Language use, cognitive processes and linguistic change. *The Routledge handbook of historical linguistics*, pages 503–518.

Byrne, A. and Thau, M. (1996). In defence of the Hybrid View. *Mind*, 105(417):139–149.

Camp, J. L. (2002). *Confusion: A Study in the Theory of Knowledge*. Harvard University Press.

Campbell, J. (1987). Is sense transparent? *Proceedings of the Aristotelian Society*, 88:273–292.

Campbell, J. (1994). *Past, Space, and Self*. MIT Press.

Campbell, J. (2005). Joint attention and common knowledge. In Eilan, N., Hoerl, C., McCormack, T., and Roessler, J., editors, *Joint Attention: Communication and Other Minds: Issues in Philosophy and Psychology*, pages 287–297. Oxford: Clarendon Press.

Campbell, J. (2017). Joint attention. In Ludwig, K. and Jankovic, M., editors, *The Routledge Handbook of Collective Intentionality*. Routledge.

Campbell, T. D. (1970). The normative fallacy. *Philosophical Quarterly*, 20(81):368–377.

Cappelen, H. (2018). *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Cappelen, H. and Hawthorne, J. (2009). *Relativism and Monadic Truth*. Oxford University Press UK.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Carey, S. and Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference*, 15:17–29.

Carroll, L. (1895). What the tortoise said to Achilles. *Mind*, 104(416):691–693.

Carston, R. (2008). Linguistic communication and the semantics/pragmatics distinction. *Synthese*, 165(3):321–345.

Casasanto, D. and Lupyan, G. (2015). All concepts are ad hoc concepts. In Laurence, S. and Margolis, E., editors, *The Conceptual Mind: New Directions in the Study of Concepts*, pages 543–566. MIT Press.

Chafe, W. (1994). *Discourse, Consciousness, and Time*. University of Chicago Press.

Chalmers, D. (2002a). The components of content. In *Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press.

Chalmers, D. J. (2002b). On sense and intension. *Philosophical Perspectives*, 16:135–82.

Chalmers, D. J. (2004). Epistemic two dimensional semantics. *Philosophical Studies*, 118(1-2):153–226.

Chalmers, D. J. (2011a). Propositions and attitude ascriptions: A Fregean account. *Noûs*, 45(4):595–639.

Chalmers, D. J. (2011b). Verbal disputes. *Philosophical Review*, 120(4):515–566.

Chalmers, D. J. (2012). *Constructing the World*. Oxford University Press.

Chalmers, D. J. and Jackson, F. (2001). Conceptual analysis and reductive explanation. *Philosophical Review*, 110(3):315–61.

Charlton, W. (1970). *Aristotle's Physics, Books I and II*. Oxford: Clarendon Press.

Chastain, C. (1975). Reference and context. In Gunderson, K., editor, *Minnesota Studies in the Philosophy of Science, volume 7*, pages 194–269. University of Minnesota Press.

Chevalier, P., Kompatsiari, K., Ciardo, F., and Wykowska, A. (2019). Examining joint attention with the use of humanoid robots: A new approach to study fundamental mechanisms of social cognition. *Psychonomic Bulletin & Review*, 27.

Cho, S. H., Cushing, C. A., Patel, K., Kothari, A., Lan, R., Michel, M., Cherkaoui, M., and Lau, H. (2018). Blockchain and human episodic memory. *arXiv preprint arXiv:1811.02881*.

Chomsky, N. (1966). *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. New York and London: Cambridge University Press.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. Prager.

Chomsky, N. (1987). *Language and Problems of Knowledge: The Managua Lectures*. MIT Press.

Chomsky, N. (1995). Language and nature. *Mind*, 104(413):1–61.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge University Press.

Christensen, D. and Kornblith, H. (1997). Testimony, memory and the limits of the a priori. *Philosophical Studies*, 86(1):1–20.

Churchland, P. M. (1998). Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal of Philosophy*, 95(1):5.

Clark, H. (1992). *Arenas of Language Use*. Chicago University Press, Chicago.

Clark, H. (1996). *Using Language*. Cambridge University Press.

Clark, H. H. (2020). Social actions, social commitments. In *Roots of human sociality*, pages 126–150. Routledge.

Conee, E. and Feldman, R. (2004). *Evidentialism: Essays in Epistemology*. Oxford, England: Oxford University Press.

Contim, F. D. V. (2015). Mental files and non-transitive de jure coreference. *Review of Philosophy and Psychology*, pages 1–24.

Crimmins, M. (1992). *Talk About Beliefs*. MIT Press.

Cumming, S. (2007). *Proper Nouns*. PhD thesis, Rutgers, New Jersey.

Cumming, S. (2013a). From coordination to content. *Philosophers' Imprint*, 13.

Cumming, S. (2013b). Creatures of darkness. *Analytic Philosophy*, 54(4):379–400.

Cumming, S. (2014). Discourse content. *A. Burgess and B. Sherman*, 2014:214–230.

Cumming, S. (2020). Definite reports of indefinites. In Goodman, R., Genone, J., and Kroll, N., editors, *Singular Thought and Mental Files*, pages 207–220. Oxford University Press.

Cumming, S. (in press). Report and content. In *The Oxford Handbook to Contemporary Philosophy of Language*. OUP.

Davidson, D. (1986). A nice derangement of epitaphs. In Lepore, E., editor, *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, pages 433–446. Blackwell.

Davies, M. (1982). Individuation and the semantics of demonstratives. *Journal of Philosophical Logic*, 11(3):287–310.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P.-F., and Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, 395(6702):597–600.

Descartes, R. (1646). Lettre au marquis de Newcastle du 23 novembre 1646. In *Œuvres et lettres*, pages 1254–1257. Gallimard, coll. "Bibliothèque de la Pléiade".

Devitt, M. (1981). *Designation*. Columbia University Press.

Devitt, M. (2001). A shocking idea about meaning. *Revue Internationale de Philosophie*, 55(218):471–494.

Devitt, M. (2014). Lest auld acquaintance be forgot. *Mind and Language*, 29(4):475–484.

Devitt, M. (2015). Should proper names still seem so problematic? In Bianchi, A., editor, *On Reference*. Oxford University Press.

Dickie, I. (2011). How proper names refer. *Proceedings of the Aristotelian Society*, 111(1pt1):43–78.

Dickie, I. and Rattan, G. (2010). Sense, communication, and rational engagement. *Dialectica*, 64(2):131–151.

Donnellan, K. S. (1974). Speaking of nothing. *Philosophical Review*, 83(1):3–31.

Dretske, F. (1981a). The pragmatic dimension of knowledge. *Philosophical Studies*, 40(3):363–378.

Dretske, F. (1995). *Naturalizing the Mind*. MIT Press.

Dretske, F. I. (1969). *Seeing And Knowing*. Chicago: University Of Chicago Press.

Dretske, F. I. (1981b). *Knowledge and the Flow of Information*. MIT Press.

Dretske, F. I. (1994). If you can't make one, you don't know how it works. *Midwest Studies in Philosophy*, 19(1):468–482.

Dummett, M. (1978). *Truth and Other Enigmas*. Cambridge, MA, USA: Harvard University Press.

Edelberg, W. (1986). A new puzzle about intentional identity. *Journal of Philosophical Logic*, 15(1):1–25.

Edelberg, W. (1992). Intentional identity and the attitudes. *Linguistics and Philosophy*, 15(6):561–596.

Egan, A. (2007). Epistemic modals, relativism and assertion. *Philosophical Studies*, 133(1):1–22.

Evans, G. (1973). The causal theory of names. *Aristotelian Society Supplementary Volume*, 47(1):187–208.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.

Everett, A. (2000). Referentialism and empty names. In Hofweber, T. and Everett, A., editors, *Empty Names, Fiction, and the Puzzles of Non-Existence*, pages 37–60. CSLI Publications.

Everett, A. (2013). *The Nonexistent*. Oxford University Press.

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. Y. (1999). Common knowledge revisited. *Annals of Pure and Applied Logic*, 96(1-3):89–105.

Falvey, K. and Owens, J. (1994). Externalism, self-knowledge, and skepticism. *Philosophical Review*, 103(1):107–37.

Féry, C. and Krifka, M. (2008). Information structure: Notional distinctions, ways of expression. In van Sterkenburg, P., editor, *Unity and Diversity of Languages*, pages 123–36. Amsterdam: John Benjamins.

Field, H. (1973). Theory change and the indeterminacy of reference. *Journal of Philosophy*, 70(14):462–481.

Fiengo, R. and May, R. (1994). *Indices and identity*. MIT press.

Fiengo, R. and May, R. (2006). *De Lingua Belief*. Cambridge MA: Bradford Book/MIT Press.

Fine, K. (2007). *Semantic Relationism*. Blackwell.

Fine, K. (2010a). Comments on Scott Soames' "Coordination problems". *Philosophy and Phenomenological Research*, 81(2):475–484.

Fine, K. (2010b). Reply to lawlor's "varieties of coreference". *Philosophy and Phenomenological Research*, 81(2):496–501.

Fodor, J. A. (1994). *The Elm and the Expert: Mentalese and Its Semantics*. MIT Press.

Fodor, J. A. and Lepore, E. (1992). *Holism: A Shopper's Guide*. Blackwell.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Frances, B. and Matheson, J. (2019). Disagreement. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2019 edition.

Frank, M., Tenenbaum, J., and Fernald, A. (2012). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, 9.

Frege, G. (1892). Uber Sinn und Bedeutung. *Zeitschrift für Philosophie Und Philosophische Kritik*, 100(1):25–50.

Frege, G. (1918/1956). The thought: A logical inquiry. *Mind*, 65(259):289–311.

Frege, G. (1923/1963). Compound thoughts. *Mind*, 72(285):1–17.

Frege, G. and Beaney, M. (1997). *The Frege Reader*. Oxford, England: Blackwell.

Friend, S. (2011). The Great Beetle debate: A study in imagining with names. *Philosophical Studies*, 153(2):183–211.

Friend, S. (2014). Notions of nothing. In *Empty Representations: Reference and Non-Existence*.

García-Carpintero, M. (2000). A presuppositional account of reference fixing. *Journal of Philosophy*, 97(3):109–147.

García-Carpintero, M. (2006). Two-dimensionalism: A neo-fregean interpretation. In García-Carpintero, M. and Macià, J., editors, *Two-Dimensional Semantics*. Oxford: Clarendon Press.

García-Carpintero, M. (2020). Co-identification and fictional names. *Philosophy and Phenomenological Research*, 101(1):3–34.

García-Carpintero, M. and Macià, J. (2006). *Two-Dimensional Semantics*. Oxford: Clarendon Press.

García-Carpintero, M. and Martí, G. (2014). *Empty Representations: Reference and Non-Existence*. Oxford University Press.

García-Carpintero, M. and Torre, S. (2016). *About Oneself: De Se Thought and Communication*. Oxford University Press.

Gasparri, L. (2015). Mental files and the lexicon. *Review of Philosophy and Psychology*, 7(2):463–472.

Gasparri, L. and Murez, M. (2019). Hearing meanings: The revenge of context. *Synthese*, 198(6):5229–5252.

Gauker, C. (2003). *Words Without Meaning*. MIT Press.

Gauker, C. (2019). Against the speaker-intention theory of demonstratives. *Linguistics and Philosophy*, 42(2):109–129.

Geach, P. T. (1967). Intentional identity. *Journal of Philosophy*, 64(20):627–632.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23(6):121–123.

Glanzberg, M. (2018). Lexical meaning, concepts, and the metasemantics of predicates. *The Science of Meaning: Essays on the Metatheory of Natural Language Semantics*, page 197.

Glüer, K. and Wikforss, s. (2022). The Normativity of Meaning and Content. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2022 edition.

Goldberg, S. C. (2007). *Anti-Individualism: Mind and Language, Knowledge and Justification*. Cambridge University Press.

Goodman, R. and Gray, A. (2022). Mental filing. *Noûs*, 56(1):204–226.

Goodsell, T. (2014). Is de jure coreference non-transitive? *Philosophical Studies*, 167(2):291–312.

Gray, A. (2016). Minimal descriptivism. *Review of Philosophy and Psychology*, 7(2):343–364.

Gray, A. (2017). Relational approaches to Frege's puzzle. *Philosophy Compass*, 12(10):e12429.

Gray, A. (2022). Minimal Fregeanism. *Mind*, 131(522):429–458.

Grice, H. P. (1957). Meaning. *Philosophical Review*, 66(3):377–388.

Grice, H. P. (1969). Utterer's meaning and intentions. *Philosophical Review*, 78(2):147–177.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

Hanks, P. (2015). *Propositional Content*. Oxford University Press.

Harman, G. (1977). Review of Jonathan Bennett's Linguistic Behaviour. *Language*, 53:417–424.

Harris, D. W. (2022). Semantics without semantic content. *Mind & Language*, 37(3):304–328.

Heck, R. (1995). The sense of communication. *Mind*, 104(413):79–106.

Heck, R. (2002). Do demonstratives have senses? *Philosophers' Imprint*, 2:1–33.

Heck, R. (2012). Solving Frege's puzzle. *Journal of Philosophy*, 109(1-2):132–174.

Heck, R. G. (2014). In defense of formal relationism. *Thought: A Journal of Philosophy*, 3(3):243–250.

Hedberg, N. (2013). Applying the Givenness Hierarchy framework: Methodological issues. In *International workshop on information structure of Austronesian languages*.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, UMass Amherst.

Heim, I. (1983). On the projection problem for presuppositions. In Portner, P. and Partee, B. H., editors, *Formal Semantics - the Essential Readings*, pages 249–260. Blackwell.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Blackwell.

Heintz, C. (2018). Cultural Attraction Theory. In Callan, H., editor, *The International Encyclopedia of Anthropology*. Wiley-Blackwell.

Heintz, C., Blancke, S., and Scott-Phillips, T. (2019). Methods for studying cultural attraction. *Evolutionary Anthropology: Issues, News, and Reviews*, 28(1):18–20.

Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language, Cognition and Neuroscience*, 29(1):2–20.

Higginbotham, J. (1985). On semantics. *Linguistic Inquiry*, 16:547–593.

Higginbotham, J. (2006). Sententialism: The thesis that complement clauses refer to themselves. *Philosophical Issues*, 16(1):101–119.

Huang, C.-M. (2010). Joint attention in human-robot interaction. Master's thesis, Georgia Institute of Technology.

Huang, C.-M. and Thomaz, A. L. (2010). Joint attention in human-robot interaction. In *AAAI Publications*, AAAI Fall Symposium Series.

Hurford, J. R. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, 77(2):187–222.

Ichikawa, J. J. and Steup, M. (2018). The Analysis of Knowledge. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2018 edition.

Jackson, F. (1998). Reference and description revisited. *Philosophical Perspectives*, 12:201–218.

Jackson, F. (2010). *Language, Names and Information*. Wiley-Blackwell.

Jeshion, R. (2010). *New Essays on Singular Thought*. Oxford University Press.

Kammerer, F. (2019). *Conscience Et Matière. Une Solution Matérialiste au Problème de l'Expérience Consciente*. Paris, France: Editions Matériologiques.

Kamp, H. (2013). Prolegomena to a structural account of belief and other attitudes. In *Meaning and the Dynamics of Interpretation*, pages 513–583. Brill.

Kamp, H. (2015). Using proper names as intermediaries between labelled entity representations. *Erkenntnis*, 80(2):263–312.

Kamp, H. (2021). Sharing real and fictional reference. In Maier, E. and Stokke, A., editors, *The Language of Fiction*. Oxford University Press.

Kamp, H. (2022). The links of causal chains. *Wiley: Theoria*, 88(2):296–325.

Kamp, H. and Bende-Farkas, A. (2019). Epistemic specificity from a communication-theoretic perspective. *Journal of Semantics*, 36:1–51.

Kaplan, D. (1968). Quantifying in. *Synthese*, 19(1-2):178–214.

Kaplan, D. (1989). Demonstratives: An essay on the semantics, logic, metaphysics and epistemology of demonstratives and other indexicals. In Almog, J., Perry, J., and Wettstein, H., editors, *Themes From Kaplan*, pages 481–563. Oxford University Press.

Kaplan, D. (1990). Words. *Aristotelian Society Supplementary Volume*, 64(1):93–119.

Kaplan, D. (2011). An Idea of Donnellan. In Almog, J. and Leonardi, P., editors, *Having In Mind: The Philosophy of Keith Donnellan*, pages 122–175. Oxford.

Kim, B. (2017). Pragmatic encroachment in epistemology. *Philosophy Compass*, 12(5):e12415.

Kim, B. and McGrath, M. (2019). *Pragmatic Encroachment in Epistemology*. Routledge.

King, J. C. and Lewis, K. S. (2021). Anaphora. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition.

Knauff, M. (2013). *Space to Reason: A Spatial Theory of Human Thought*. MIT Press.

Korcz, K. A. (2021). The Epistemic Basing Relation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition.

Kriegel, U. (2011). Cognitive phenomenology as the basis of unconscious content. In Bayne, T. and Montague, M., editors, *Cognitive Phenomenology*, pages 79–102. Oxford University Press.

Kripke, S. (1982). *Wittgenstein on Rules and Private Language*. Harvard University Press.

Kripke, S. A. (1979). A puzzle about belief. In Margalit, A., editor, *Meaning and Use*, pages 239–83. Reidel.

Kripke, S. A. (1980). *Naming and Necessity*. Harvard University Press.

Kroon, F. W. (1987). Causal descriptivism. *Australasian Journal of Philosophy*, 65(1):1–17.

Lawlor, K. (2010). Varieties of coreference. *Philosophy and Phenomenological Research*, 81(2):485–495.

Lederman, H. (2017). Two paradoxes of common knowledge: Coordinated attack and electronic mail. *Noûs*, 52(4):921–945.

Lederman, H. (2018). Common knowledge. In Jankovic, M. and Ludwig, K., editors, *The Routledge Handbook of Collective Intentionality*, pages 181–195.

Lederman, H. (2022). Fregeanism, sententialism, and scope. *Linguistics and Philosophy*, pages 1–41.

Lerique, S. (2017). *Epidemiology of representations: an empirical approach*. PhD thesis, Paris, EHESS.

Levy, D. and Olson, K. (1992). Types, tokens and templates. *Technical Report, CSLI*, (169).

Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy*, 63(1):17–25.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy*, 67(13):427–446.

Lewis, D. (1986). *On the Plurality of Worlds*. Wiley-Blackwell.

Lewis, D. K. (1969). *Convention: A Philosophical Study*. Cambridge, MA, USA: Wiley-Blackwell.

Lewis, D. K. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50(3):249–258.

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, MA, USA: Blackwell.

Lewis, D. K. (1977). Possible-world semantics for counterfactual logics: A rejoinder. *Journal of Philosophical Logic*, 6(1):359–363.

Lewis, D. K. (1980). Index, context, and content. In Kanger, S. and Öhman, S., editors, *Philosophy and Grammar*, pages 79–100. Reidel.

Loar, B. (1976). The semantics of singular terms. *Philosophical Studies*, 30(6):353–377.

Loar, B. (2017). Social Content and Psychological Content. In *Consciousness and Meaning: Selected Essays*. Oxford University Press.

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N., and Hashimoto, T. (2015). Gettier across cultures. *Noûs*, pages 645–664.

Maier, E. (2016a). Attitudes and mental files in discourse representation theory. *Review of Philosophy and Psychology*, 7(2):473–490.

Maier, E. (2016b). Why my "I" is your "You": On the communication of de se attitudes. In Garcia-Carpintero, M. and Torre, S., editors, *About Oneself: De Se Thought and Communication*. Oxford University Press.

Maier, E. (2017). Fictional names in psychologistic semantics. *Theoretical Linguistics*, 43(1-2):1–46.

Maier, E. and Stokke, A. (2021). *The Language of Fiction*. Oxford University Press.

Margolis, E. and Laurence, S. (1999). *Concepts: Core Readings*. MIT Press.

Martinich, A. (1996). *The Philosophy of Language, 3rd Edition*. Oxford University Press.

Mates, B. (1952). Synonymity. In Linsky, L., editor, *Semantics and the Philosophy of Language*, pages 111–136. University of Illinois Press.

Maxfield, L. (1997). Attention and semantic priming: A review of prime task effects. *Consciousness and cognition*, 6(2-3):204–218.

May, R. (2006). The invariance of sense. *Journal of Philosophy*, 103(3):111–144.

Medina, T. N., Snedeker, J., Trueswell, J. C., and Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22):9014–9019.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

Millikan, R. G. (2000). *On Clear and Confused Ideas: An Essay About Substance Concepts*. Cambridge and New York: Cambridge University Press.

Moravcsik, J. M. (1998). *Meaning, Creativity, and the Partial Inscrutability of the Human Mind*. Center for the Study of Language and Information.

Morin, O. (2013). What does communication contribute to cultural transmission? *Social Anthropology/Anthropologie Sociale*, 21(2):230–235.

Morin, O. (2015). *How Traditions Live and Die*. Oxford University Press USA.

Moser, P. K. (1989). *Knowledge and Evidence*. Cambridge University Press.

Moss, S. (2012). Updating as communication. *Philosophy and Phenomenological Research*, 85(2):225–248.

Murez, M. (2016). *Singular concepts: from fragments to mental files*. PhD thesis, École des hautes études en sciences sociales, Paris.

Murez, M. (2019). Le Fressellianisme face au dilemme de l'accointance. *Les Etudes Philosophiques*, 3:421.

Murez, M. (2021). Belief fragments and mental files. In *The Fragmented Mind*, number 1, pages 251–278. Oxford University Press.

Murez, M. (2022). Les fichiers mentaux sont-ils transparents ? `https://www.college-de-france.fr/site/en-francois-recanati/seminar-2022-03-11-15h30.htm`. Accessed: 16 June 2022.

Murez, M. and Recanati, F. (2016). Mental files: An introduction. *Review of Philosophy and Psychology*, 7(2):265–281.

Murez, M. and Smortchkova, J. (2014). Singular thought: Object-files, person-files, and the sortal person. *Topics in Cognitive Science*, 6(4):632–646.

Murez, M., Smortchkova, J., and Strickland, B. (2020). The mental files theory of singular thought: A psychological perspective. In Goodman, R., Genone, J., and Kroll, N., editors, *Singular Thought and Mental Files*, pages 107–142. Oxford: Oxford University Press.

Neale, S. (1992). Paul Grice and the philosophy of language. *Linguistics and Philosophy*, 15(5):509–559.

Nichols, S. (2006). *The Architecture of the Imagination: New Essays on Pretence, Possibility, and Fiction*. Oxford University Press UK.

Nichols, S. and Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford University Press.

Ninan, D. (2009). *Imagination, Content, and the Self*. PhD thesis, MIT.

Ninan, D. (2012). Counterfactual attitudes and multi-centered worlds. *Semantics and Pragmatics*, 5(5):1–57.

Ninan, D. (2016). What is the problem of de se attitudes? In Torre, S. and Garcia-Carpintero, M., editors, *About Oneself: De Se Thought and Communication*. Oxford University Press.

Nyhof, M. and Barrett, J. (2001). Spreading non-natural concepts: The role of intuitive conceptual structures in memory and transmission of cultural materials. *Journal of cognition and culture*, 1(1):69–100.

O'Madagain, C. (2018). Outsourcing concepts: Deference, the extended mind, and expanding our epistemic capacity. In Carter, J. A., Clark, A., Kallestrup, J., Palermos, O., and Pritchard, D., editors, *Socially Extended Knowledge*. Oxford University Press.

O'Madagain, C. and Tomasello, M. (2019). Joint attention to mental content and the social origin of reasoning. *Synthese*, 198(5):4057–4078.

Onofri, A. (2012). *Concepts in Context*. PhD thesis, University of St. Andrews.

Onofri, A. (2016). Two constraints on a theory of concepts. *Dialectica*, 70(1):3–27.

Onofri, A. (2018). The publicity of thought. *Philosophical Quarterly*, 68(272).

Origgi, G. and Sperber, D. (2000). Evolution, communication and the proper function of language. In Carruthers, P. and Chamberlain, A., editors, *Evolution and the Human Mind: Language, Modularity and Social Cognition*, pages 140–169. Cambridge: Cambridge University Press.

Orlando, E. (2017). Files for fiction. *Acta Analytica*, 32(1):55–71.

Ostertag, G. (2007). Review of Robert Fiengo, Robert May, *De Lingua Belief*. *Notre Dame Philosophical Reviews*, 2007(9).

Pagin, P. (2003). Communication and strong compositionality. *Journal of Philosophical Logic*, 32(3):287–322.

Pal, P., Zhu, L., Golden-Lasher, A., Swaminathan, A., and Williams, T. (2020). Givenness Hierarchy theoretic cognitive status filtering. *arXiv preprint arXiv:2005.11267*.

Peacocke, C. (2005). Joint attention: Its nature, reflexivity, and relation to common knowledge. In Eilan, N., Hoerl, C., McCormack, T., and Roessler, J., editors, *Joint Attention: Communication and Other Minds*, page 298. Oxford University Press.

Peet, A. (2016). Referential intentions and communicative luck. *Australasian Journal of Philosophy*, 95(2):379–384.

Peet, A. (2019). Knowledge-yielding communication. *Philosophical Studies*, 176(12):3303–3327.

Peirce, C. S. (1931). *Collected Papers*. Cambridge: Belknap Press of Harvard University Press.

Perini-Santos, E. (2009). Does contextualism make communication a miracle? *Manuscrito*, 32(1):231–247.

Perry, J. (1997). Saying nothing? Unpublished.

Perry, J. (2003). The search for the semantic grail. *Philosophic Exchange*, 33(1).

Perry, J. (2012). *Reference and Reflexivity, 2nd Edition*. Center for the Study of Language and Information, Stanford, California.

Perry, J. (2015). The cognitive contribution of names. In Bianchi, A., editor, *On reference*, pages 189–208. Oxford University Press Oxford.

Perry, J. (2019). *Frege's Detour: An Essay on Meaning, Reference, and Truth*. Oxford, England: Oxford University Press.

Perry, J. (2020). Singular thoughts. In Goodman, R., Genone, J., and Kroll, N., editors, *Singular Thought and Mental Files*, pages 143–158. Oxford University Press.

Piantadosi, S. T., Tenenbaum, J. B., and Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, 123(4):392–424.

Pinillos, A. (2011). Coreference and meaning. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 154(2):301–324.

Pinillos, A. (2012). Knowledge, experiments, and practical interests. In Brown, J. and Gerken, M., editors, *Knowledge Ascriptions*, pages 192–219. Oxford University Press.

Pinillos, A. (2015). Millianism, relationism and attitude ascriptions. In Bianchi, A., editor, *On Reference*. Oxford University Press.

Pinillos, A. (2020). De jure anti-coreference and mental files. In Goodman, R., Genone, J., and Kroll, N., editors, *Singular Thought and Mental Files*, pages 187–206. Oxford University Press.

Pinillos, A. (unknown). Names, logical form and syntactic externalism. Unpublished.

Plunkett, D. and Sundell, T. (2013). Disagreement and the semantics of normative and evaluative terms. *Philosophers' Imprint*, 13(23):1–37.

Plunkett, D. and Sundell, T. (2021). Metalinguistic negotiation and speaker error. *Inquiry: An Interdisciplinary Journal of Philosophy*, 64(1-2):142–167.

Pollock, J. (2020). Holism, conceptual role, and conceptual similarity. *Philosophical Psychology*, 33(3):396–420.

Priest, G. (2006). *In Contradiction: A Study of the Transconsistent, 2nd Edition*. Oxford University Press.

Pritchard, D. (2005). *Epistemic Luck*. Oxford University Press UK.

Pritchard, D. (2007). Anti-luck epistemology. *Synthese*, 158(3):277–297.

Prosser, S. (2019). Shared modes of presentation. *Mind and Language*, 34(4):465–482.

Prosser, S. (2020). The metaphysics of mental files. *Philosophy and Phenomenological Research*, 100(3):657–676.

Pryor, J. (2016). Mental graphs. *Review of Philosophy and Psychology*, 7(2):309–341.

Pryor, J. (2017). *De Jure Codesignation*, chapter 41, pages 1033–1079. John Wiley & Sons, Ltd.

Pusiol, G., Soriano, L., Fei-fei, L., and Frank, M. C. (2014). Discovering the signatures of joint attention in child-caregiver interaction.

Putnam, H. (1953). Synonymity and the analysis of belief sentences. *Analysis*, 14(5):114–122.

Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7:131–193.

Quilty-Dunn, J. (2021). Polysemy and thought: Toward a generative theory of concepts. *Mind and Language*, 36(1):158–185.

Quine, W. V. O. (1956). Quantifiers and propositional attitudes. *Journal of Philosophy*, 53(5):177–187.

Quine, W. V. O. (1960). *Word and Object*. Cambridge, MA, USA: MIT Press.

Quine, W. V. O. (1970). *Philosophy of Logic*. Harvard University Press.

Ramsey, F. P. (1931/1990). Theories. In Mellor, D. H., editor, *The Foundations of Mathematics and Other Logical Essays*. Cambridge: Cambridge University Press.

Rappaport, J. (2017). Is there a meaning-intention problem? *Croatian Journal of Philosophy*, 17(3):383–397.

Recanati, F. (1993). *Direct Reference: From Language to Thought*. Blackwell.

Recanati, F. (1995). The communication of first person thoughts. In Kotatko, P. and Biro, J., editors, *Frege: Sense and Reference One Hundred Years Later*, pages 95–102. Kluwer Academic Publishers.

Recanati, F. (1996). Direct reference. *Philosophy and Phenomenological Research*, 56(4):953–956.

Recanati, F. (1997). Can we believe what we do not understand? *Mind and Language*, 12(1):84–100.

Recanati, F. (2000). *Oratio obliqua, oratio recta: An essay on metarepresentation*. MIT Press.

Recanati, F. (2001). Modes of presentation: Perceptual vs. deferential.

Recanati, F. (2002). *Literal Meaning*. Cambridge University Press.

Recanati, F. (2008). *Philosophie du Langage (Et de L'Esprit)*. Editions Gallimard.

Recanati, F. (2012). *Mental Files*. Oxford University Press.

Recanati, F. (2013). Mental files: Replies to my critics. *Disputatio*, 5(36):207–242.

Recanati, F. (2016). *Mental Files in Flux*. Oxford University Press, Oxford, UK.

Recanati, F. (2018). Fictional, metafictional, parafictional. *Proceedings of the Aristotelian Society*, 118(1):25–54.

Recanati, F. (2019). Transparent coreference. *Topoi*, 40(1):107–115.

Recanati, F. (2020a). Do Mental Files Obey Strawson's Constraint? In Cristina Borgoni, D. K. and Onofri, A., editors, *The Fragmented Mind*. Oxford University Press.

Recanati, F. (2020b). Multiple grounding. In Bianchi, A., editor, *Language and Reality From a Naturalistic Perspective: Themes From Michael Devitt*. Springer.

Recanati, F. (2021). Fictional reference as simulation. In Stokke, E. M. . A., editor, *The language of fiction*, pages 17–36. Oxford University Press Oxford.

Recanati, F. (2022). Entertaining as simulation. In *Force, Content and the Unity of the Proposition*, pages 112–135. Routledge.

Recanati, F. (forthcoming). Sameness of mode of presentation. Manuscript.

Reddy, M. J. (1979). The conduit metaphor. In Ortony, A., editor, *Metaphor and Thought*. Cambridge University Press.

Rescorla, M. (2019). The Language of Thought Hypothesis. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2019 edition.

Rescorla, M. (2020). The Computational Theory of Mind. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition.

Richard, M. (1990). *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them*. Cambridge University Press.

Richard, M. (2019). *Meanings as Species*. Oxford University Press.

Russell, B. (1903). *The Principles of Mathematics*. Cambridge, England: Allen & Unwin.

Rysiew, P. (2021). Epistemic Contextualism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2021 edition.

Sainsbury, M. (2002). *Departing From Frege: Essays in the Philosophy of Language*. Routledge.

Sainsbury, M. (2005). *Reference Without Referents*. Oxford, England and New York, NY, USA: Clarendon Press.

Sainsbury, R. M. and Tye, M. (2011). An originalist theory of concepts. *Aristotelian Society Supplementary Volume*, 85(1):101–124.

Sainsbury, R. M. and Tye, M. (2012). *Seven Puzzles of Thought and How to Solve Them: An Originalist Theory of Concepts*. Oxford, England and New York, NY, USA: Oxford University Press.

Salmon, N. (1989). Illogical belief. *Philosophical Perspectives*, 3:243–285.

Salmon, N. U. (1986). *Frege's Puzzle*. Ridgeview.

Sandgren, A. (2016). *Cruel Intensions: An Essay on Intentional Identity and Intentional Attitudes*. PhD thesis, Australian National University.

Sandgren, A. (2019). A metarepresentational theory of intentional identity. *Synthese*, 196(9):3677–3695.

Saul, J. (2009). Simple sentences, substitution, and intuitions. *Analysis*, 69(1):174–176.

Saul, J. M. (1998). The pragmatics of attitude ascription. *Philosophical Studies*, 92(3):363–389.

Schiffer, S. (1972). *Meaning*. Oxford University Press, Oxford.

Schiffer, S. (1981). Indexicals and the theory of reference. *Synthese*, 49(1):43–100.

Schiffer, S. (1990). The mode-of-presentation problem. In Anderson, C. A. and Owens, J., editors, *Propositional Attitudes: The Role of Content in Logic, Language, and Mind*, pages 249–268. CSLI.

Schiffer, S. (1992). Belief ascription. *Journal of Philosophy*, 89(10):499–521.

Schiffer, S. (2005). What reference has to tell us about meaning. In *Proceedings of the Seminar Series in Analytic Philosophy 2003-2004*, Cognição e Conteúdo (Cognition and Content), pages 138–166.

Schiffer, S. (2006). The things we mean. *Philosophy and Phenomenological Research*, 73(1):208–210.

Schiffer, S. (forthcominga). Aphonic terms and the deep problem with Gricean meaning. In Ostertag, G., editor, *Meanings and Other Things: Essays on Stephen Schiffer*. Oxford University Press.

Schiffer, S. (forthcomingb). Expression-meaning and vagueness. In Sullivan, A., editor, *Sensations, Thoughts, Language: Essays in Honor of Brian Loar*. Routledge.

Schiffer, S. R. (1987). *Remnants of Meaning*. MIT Press.

Schneider, S. (2009). The nature of symbols in the language of thought. *Mind and Language*, 24(5):523–553.

Schneider, S. (2011). *The Language of Thought: A New Philosophical Direction*. MIT Press.

Schönpflug, U. (2008). *Cultural transmission: Psychological, developmental, social, and methodological aspects*. Cambridge University Press.

Schroeter, L. (2007). Illusion of transparency. *Australasian Journal of Philosophy*, 85(4):597–618.

Schroeter, L. (2012). Bootstrapping our way to samesaying. *Synthese*, 189(1):177–197.

Schroeter, L. (2013). Are concepts creatures of darkness? *Analytic Philosophy*, 54(2):277–292.

Schroeter, L. and Bigelow, J. (2009). Jackson's classical model of meaning. In Ravenscroft, I., editor, *Minds, Ethics, and Conditionals: Themes From the Philosophy of Frank Jackson*, page 85. Oxford University Press.

Schroeter, L. and Schroeter, F. (2014). Normative concepts: A connectedness model. *Philosophers' Imprint*, 14.

Schroeter, L. and Schroeter, F. (2016). Semantic deference versus semantic coordination. *American Philosophical Quarterly*, 53(2):193–210.

Scofield, J. and Behrend, D. A. (2011). Clarifying the role of joint attention in early word learning. *First Language*, 31(3):326–341.

Searle, J. (1965). What is a speech act? In Black, M., editor, *Philosophy in America*, pages 221–239. Ithaca: Cornell University Press.

Seemann, A. (2019). *The Shared World: Perceptual Common Knowledge, Demonstrative Communication, and Social Space*. The MIT Press.

Segal, G. (2000). *A Slim Book About Narrow Content*. MIT Press.

Semeijn, M. (2021). *Fiction and Common Ground*. PhD thesis.

Shan, C.-C. (2007). Causal reference and inverse scope as mixed quotation.

Simchen, O. (2017). *Semantics, Metasemantics, Aboutness*. Oxford University Press.

Sloan, L. (2016). Using Twitter in social science research. `https://dx.doi.org/10.4135/9781473963245`. SAGE Research Methods. Accessed: 2022-06-06.

Soames, S. (1987a). Direct reference, propositional attitudes, and semantic content. *Philosophical Topics*, 15(1):47–87.

Soames, S. (1987b). Substitutivity. In Thomson, J. J., editor, *On Being and Saying: Essays for Richard Cartwright*, pages 99–132. MIT Press.

Soames, S. (2002). *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*. Oxford University Press.

Soames, S. (2008). The gap between meaning and assertion: Why what we literally say often differs from what our words literally mean. In Soames, S., editor, *Philosophical Essays, Volume 1: Natural Language: What It Means and How We Use It*, pages 278–297. Princeton University Press.

Soames, S. (2013). Cognitive propositions. *Philosophical Perspectives*, 27(1):479–501.

Speaks, J. (2013). Individuating Fregean sense. *Canadian Journal of Philosophy*, 43(5):634–654.

Sperber, D. (1994). Understanding verbal understanding. In Khalfa, J., editor, *What is Intelligence?* Cambridge University Press.

Sperber, D. (1996). *Explaining Culture: A Naturalistic Approach*. Oxford: Basil Blackwell.

Sperber, D. (1997). Intuitive and reflective beliefs. *Mind and Language*, 12(1):67–83.

Sperber, D. (2001). Conceptual tools for a natural science. In *Proceedings of the British Academy*, volume 111, pages 297–317. British Academy.

Sperber, D., Ement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., and Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, 25:359–393.

Sperber, D. and Wilson, D. (1996). *Relevance: Communication and Cognition, Second Edition*. Oxford: Blackwell.

Stalnaker, R. (1976). Propositions. In MacKay, A. F. and Merrill, D. D., editors, *Issues in the Philosophy of Language: Proceedings of the 1972 Colloquium in Philosophy*, pages 79–91. New Haven and London: Yale University Press.

Stalnaker, R. (1978). Assertion. *Syntax and Semantics (New York Academic Press)*, 9:315–332.

Stalnaker, R. (1984). *Inquiry*. Cambridge University Press.

Stalnaker, R. (1998). On the representation of context. *Journal of Logic, Language and Information*, 7(1):3–19.

Stalnaker, R. (1999). *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford University Press UK.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, 25(5-6):701–721.

Stalnaker, R. (2008). *Our Knowledge of the Internal World*. Oxford University Press.

Stalnaker, R. (2014). *Context*. Oxford University Press.

Stanley, J. (2005). *Knowledge and Practical Interests*. Oxford University Press.

Stich, S. P. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. MIT Press.

Stone, M. (2005). Communicative intentions and conversational processes in human-human and human-computer dialogue. *Approaches to studying world-situated language use*, pages 39–70.

Strawson, P. F. (1950). On referring. *Mind*, 59(235):320–344.

Strawson, P. F. (1974). *Subject and predicate in grammar and logic*. Routledge.

Taschek, W. W. (1998). On ascribing beliefs: Content in context. *Journal of Philosophy*, 95(7):323–353.

Tayebi, S. (2013). Recanati on communication of first-person thoughts. *Thought: A Journal of Philosophy*, 1(3):210–218.

Taylor, K. A. (2021). *Referring to the World: An Opinionated Introduction to the Theory of Reference*. Oxford University Press. posthumous.

Thagard, P. (2001). Internet epistemology: Contributions of new information technologies to scientific research. *Designing for science: Implications from everyday, classroom, and professional settings*, pages 465–485.

Thuns, A. (2017). Lexicalizing semantic deference. Manuscript.

Thuns, A. (2020). *Word Meanings Out There and Within: Toward a Naturalistic Account*. PhD thesis, Université libre de Bruxelles.

Tomasello, M. (1998). Reference: Intending that others jointly attend. *Pragmatics and Cognition*, 6(1):229–243.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press.

Tomasello, M. (2008). *Origins of Human Communication*. MIT Press.

Torre, S. and Weber, C. (2021). What is special about de se attitudes? In Biggs, S. and Geirsson, H., editors, *The Routledge Handbook of Linguistic Reference*, pages 464–481. Routledge.

Tulving, E. and Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of verbal learning and verbal behavior*, 5(4):381–391.

Unnsteinsson, E. (2018). Referential intentions: A response to Buchanan and Peet. *Australasian Journal of Philosophy*, 96(3):610–615.

Urquhart, A. and Lewis, A. C. (1994). *The Collected Papers of Bertrand Russell, Volume 4: Foundations of Logic, 1903-05*. Routledge.

Valente, M. and Verdejo, V. M. (2021). Relationism and the problem of publicity. *Pacific Philosophical Quarterly*.

Vallduví, E. (1992). *The Informational Component*. Taylor & Francis.

Vallduví, E. and Engdahl, E. (1996). The linguistic realisation of information packaging. *Linguistics*, 34.

Weber, C. (2013). Centered communication. *Philosophical Studies*, 166(S1):205–223.

Wegner, D. M. (1995). A computer network model of human transactive memory. *Social cognition*, 13(3):319–339.

Wetzel, L. (2018). Types and Tokens. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2018 edition.

Wikforss, s. (2015). The insignificance of transparency. In Goldberg, S. C., editor, *Externalism, Self-Knowledge, and Skepticism: New Essays*, pages 142–164. Cambridge University Press.

Wilby, M. (2010). The simplicity of mutual knowledge. *Philosophical Explorations*, 13(2):83–100.

Williams, J. (2016). *Joint Attention in Toddler Vocabulary Acquisition*. PhD thesis, Department of Psychology and Neuroscience, University of Colorado Boulder.

Williams, T., Schreitter, S., and Scheutz, M. (2019). Situated open world reference resolution for human-robot dialogue.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press.

Williamson, T. (2007). *The Philosophy of Philosophy*. Wiley-Blackwell.

Williamson, T. (2009). Reply to Goldman. In Pritchard, D. and Greenough, P., editors, *Williamson on Knowledge*, pages 305–312. Oxford: Oxford University Press.

Witek, M. (2022). An Austinian alternative to the Gricean perspective on meaning and communication. *Journal of Pragmatics*, 201:60–75.

Wittgenstein, L. J. J. (1953). *Philosophical Investigations*. New York, NY, USA: Wiley-Blackwell.

Woodfield, A. (1991). Conceptions. *Mind*, 100(399):547–72.

Yablo, S. (2006). Non-catastrophic presupposition failure. In Thomson, J. J. and Byrne, A., editors, *Content and Modality: Themes From the Philosophy of Robert Stalnaker*. Oxford University Press.

Yalcin, S. (2015). Quantifying in from a Fregean perspective. *Philosophical Review*, 124(2):207–253.

Zavaleta, M. A. (2019). Communication and Variance. *Topoi*, 40(1):147–169.

Zollman, K. J. S. (2007). The communication structure of epistemic communities. *Philosophy of Science*, 74(5):574–587.

Zollman, K. J. S. (2013). Network epistemology: Communication in epistemic communities. *Philosophy Compass*, 8(1):15–27.

RÉSUMÉ

Cette thèse étudie la nature de la relation entre les représentations mentales dans la communication verbale réussie, l'attribution des pensées, l'accord et le désaccord -- relation que j'appelle "samethinking". La nature du samethinking soulève plusieurs questions fondamentales sur la nature de la signification (non naturelle) et les fondements cognitifs de l'émergence de la culture. Elle concerne également des énigmes de longue date en philosophie de l'esprit et du langage (telles que le problème de Frege et le problème de Kripke sur la croyance). Le samethinking ne se résume pas au partage de la référence (par "partage" je fais référence au fait pour deux ou plusieurs penseurs d'avoir quelque chose en commun) : il est plus exigeant. Comment pouvons-nous expliquer et caractériser cette relation, plus exigeante que la coréférence, qui est instanciée par une paire de pensées lorsque le samethinking a lieu ? On suppose souvent que cette relation implique le partage d'un contenu plus fin que la référence. Dans cette thèse, je soutiens que la question est plus complexe que ce qui a été communément supposé, et je propose un modèle alternatif dans lequel le partage du contenu plus fin que la référence n'est pas nécessaire.

MOTS CLÉS

Philosophie du langage & de l'esprit ; Communication ; Contenu ; Problème de Frege ; Relationnisme

ABSTRACT

This thesis investigates the nature of the relation between mental representations in successful verbal communication, thought attribution, agreement, and disagreement -- a relation which I call “samethinking”. The nature of samethinking raises several foundational questions about the nature of (non-natural) meaning, and the cognitive underpinnings of the emergence of culture. It bears on long-lasting puzzles in the philosophy of mind and language (such as Frege's puzzle and Kripke's puzzle about belief). Samethinking does not amount to sharing a reference (with “sharing" I refer to two or more thinkers having something in common): it is more demanding. How can we explain and characterize this relation, more stringent than coreference, that is instantiated by a pair of thoughts when samethinking takes place? It is often assumed that this relation involves sharing a thought content more fine-grained than reference. In this thesis, I argue that the issue is more complex than what has been commonly assumed, and I suggest an alternative model in which sharing thought content is not necessary.

KEYWORDS

Philosophy of mind & language; Communication; Content; Frege's puzzle; Relationism