Charles University
Faculty of Arts
Institute of Phonetics

Master thesis

Zuzana Oceláková

Spoken communication in context: Integrating language as a discriminative code into enactive cognition

Mluvená komunikace v kontextu:
Začlenění jazyka jakožto diskriminativního kódu do enaktivní kognice

Prague, 2020 Supervisor: Kateřina Chládková, M.A., Ph.D.

In a roughly chronological order, I would like to thank Jelle Bruineberg for the pleasant chat in *het Amsterdamse Bos*, which determined the following two years of my academic life; Kateřina Chládková for being adventurous enough to immerse in this topic, for hundreds of apt comments and for her encouragement; Radek Ocelák for sharpening my thoughts by his uncompromising scepticism about academic science; and all the dancers for constantly reminding me what *communication* really means.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 21. 7. 2020

Zuzana Oceláková

Abstract

Spoken communication is only one of many types of human interaction with the environment. The aim of this thesis is to propose a theory of spoken communication based on basic principles of cognition, which govern all our behaviour. To this end, two established theoretical positions are integrated: 1) the skilled intentionality framework (an enactive view of cognition) and 2) a discriminative approach to human communication. According to the resulting theory, communication is a skilful shaping of an interlocutor's environment which serves to fulfil the agent's positively biased expectations about her own situation. Language is presented as an assemblage of sociomaterial regularities that make this skilful behaviour possible. The suggested perspective is radically action-oriented, in contrast with traditional representational, content-based approaches. The proposed view is then applied to two specific phenomena widely studied within speech sciences (namely categoricality of speech and turn taking) and is confronted with selected empirical findings. Possibilities of empirical testing of the suggested theory are discussed.

Keywords: spoken communication, enactive cognition, embodied cognition, skilled intentionality framework, discriminative codes, anticipation

Abstrakt

Mluvená komunikace je pouze jedním z druhů interakce mezi člověkem a jeho prostředím. Cílem této práce je navrhnout takovou teorii mluvené komunikace, jež by vycházela ze základních principů kognice řídících veškeré naše jednání. Za tímto účelem jsou integrovány dvě již existující teoretické pozice: 1) enaktivní pojetí kognice zvané *skilled intentionality framework*; a 2) teorie lidské komunikace založená na diskriminaci. Podle výsledného návrhu spočívá komunikace v tom, že mluvčí dovedně ovlivňuje prostředí svého komunikačního partnera ve snaze naplnit svá pozitivní očekávání ohledně vlastní situace. Jazyk je představen jako soubor sociomateriálních pravidelností, díky nimž je získání takovéto dovednosti vůbec možné. Primární roli hraje v tomto přístupu jednání, což jej odlišuje od teorií orientovaných na sdílení obsahu a na mentální jazykové reprezentace. Navržená teorie je pak aplikována na dvě témata hojně studovaná v řečových vědách (kategoriálnost řeči a turn taking, tj. střídání replik v konverzaci) a konfrontována s vybranými empirickými poznatky. Nakonec jsou diskutovány možnosti empirického testování její adekvátnosti.

Klíčová slova: mluvená komunikace, enaktivní kognice, vtělená kognice, skilled intentionality framework, diskriminativní kódy, anticipace

Contents

Introduction

Motivation

The way people communicate is, in my view, their most fascinating trait. Years of phonetic studies have nothing but deepened this astonishment, which, moreover, has broadened its scope during that time. Initially, I was interested in language, focusing on its both onto- and phylogenetically primary instantiation, i.e. speech. Then I gradually became fascinated by communication in general, in all its forms and flavours. Spoken communication took its place beside communication by signs and gestures, communication between other than human animals, or physical communication in partner dancing. After all, one often encounters the parallel "dancing is like a conversation". Practically useful as it may be for learning how to dance, this parallel is not very satisfying in theoretical terms – the principles and mechanisms underlying successful conversation are just as mysterious as the ones underlying a good dance. A lot is still unknown about how speakers and listeners process speech and what enables them to communicate with the smoothness and readiness they do. What are the basic units of speech processing? What cues do language users use to recognize them? How are they learned and in what form are they stored? What mechanisms do interlocutors use to organise their conversations? Vast number of studies address these and other related questions every year by carefully designed experiments, each of them adding a small piece to the mosaic of our knowledge.

This thesis is not going to be one of them. Its aim is to zoom out from

these endeavours and make full use of the fact that spoken communication is only one of many types of interactions that animals have with their environment. Language behaviour must surely fit into a broader picture of our general relationship with the world around us.[1] Following this idea, this thesis is an attempt to approach the subject of my long-standing interest, i.e. spoken communication, from an unusual direction. Instead of starting from speech and then searching for its place among the rest of our activities, I will start from the characteristics that humans share with all living creatures and let my view of speech be governed by the most general principles of interaction between organisms and their environment. After all, human language is a novelty – for millions of years, organisms were interacting with the world without it (and most of them still are). Our own non-linguistic behaviour, connecting us to other animals as well as to our own past, might therefore be a useful source of insights for linguistic research. Perhaps rather than dancing being like a conversation, conversation is like dancing.

Outline

In both cognitive science and linguistics, a great amount of work has been done that I can build on. I have neither the space nor the ambition to cover the research in those fields exhaustively – I will focus on the directions of

[1] This relationship is studied by cognitive scientists, and, in a different manner, philosophers. I am convinced that speech science should be open to insights from both these fields, because they aim to create a frame into which our findings about spoken communication must fit if our view of us humans is to be coherent. Assumptions concerning the nature of cognition are anyway necessarily present (more or less explicitly) in communication research. Often these are conceptual. For instance, when linguists talk about "levels of representation" (distinguishing for instance phonological and syntactic level), what is assumed is a representation-based view of cognition: humans are supposed to create internal structures that somehow reflect the outside world. These assumptions necessarily govern the research, influencing not only interpretation of experimental results and selection of methods, but crucially also the *questions* that are asked. I therefore consider it worthwhile to turn explicitly to the underlying concepts and ideas at some point.

thought that I consider the most convincing and promising, always accounting for my choice. The first two chapters of this thesis will therefore be dedicated to setting the cognitive and linguistic background, respectively:

- **Chapter 1** introduces the main tenets of a particular variety of enactivism, called the skilled intentionality framework (Bruineberg, 2018). This is the theory of cognition I adhere to and it will serve as the frame which the subsequently developed theory of language should fit into. The skilled intentionality framework seems particularly suitable for my endeavour, because it claims that there is no substantial difference between our basal vital functions and highly complex activities (including intellectual ones).

- **Chapter 2** introduces a theory of language (and human communication in general) based on insights from information theory (Ramscar, 2019). Language is presented as a discriminative code that serves to identify the intended message by eliminating all other possibilities. I consider this theory appealing on its own right, and moreover see certain affinities between this approach to language and the skilled intentionality framework, which I will carefully explore.

While the first two chapters serve merely to establish the necessary conceptual ground, presenting two conceptions that already exist, the remaining two bring my contribution:

- **Chapter 3**, points out the affinities that I found between the views presented in chapters 1 and 2 and tries to exploit their potential. By adjusting the discriminative theory of language to the demands of the skilled intentionality framework, a novel perspective on human communication arises.

- **Chapter 4** explores the applicability of the integrated theory of language on two phenomena that are extensively studied within speech

sciences: categoricality of speech and turn taking. Crucial concepts related to these topics are reinterpreted in terms of the novel perspective. The suggested theory is also confronted with some relevant empirical findings, and possibilities of further empirical research inspired by the integrated perspective are considered.

Chapter 1

Cognitive background

1.1 Enactive cognition

When we look up the term "cognition" on Wikipedia, the first we see is the following definition: "The mental action or process of acquiring knowledge and understanding through thought, experience, and the senses." (*Wikipedia: Cognition*, n.d.) Let us compare this definition with another one, found in Di Paolo and Thompson (2017, p. 76): "Cognition, in its most general form, is sense-making – the adaptive regulation of states and interactions by an agent with respect to the consequences for the agent's own viability."; and one more, from Thompson (2007), page 13: "(...) cognition is the exercise of skilful know-how in situated and embodied action." It is the approach expressed in the last two definitions (i.e. what is called the *enactive approach*[1]) that I

[1]Enactivism is not a completely homogeneous theory. Bruineberg (2018) explicitly distinguishes two enactivistic positions: autopoietic enactivism (introduced by Varela, Thompson, & Rosch, 1991, as the first version of enactivism), and radical enactivism (Hutto & Myin, 2013, 2017), calling both of them his close philosophical neighbours (Bruineberg, 2018, pp. 23–25). Listing and comparing projects falling under enactivism is not an objective of this thesis; an interested reader may find a lot of relevant information and resources in The Oxford Handbook of 4E Cognition (Newen, De Bruin, & Gallagher, 2018). I follow the path that enactivism took in Bruineberg's (2018) work, which to some extend guides the way I characterize this approach in the following sections.

advocate in this thesis, but given how Wikipedia works, we might assume that it is the first definition that more or less captures the common understanding of cognition. Below I explain how the common and the enactive approach to cognition relate to each other.

1.1.1 Embodiment

The first difference one might notice is that whereas the first definition starts with characterizing cognition as "mental" (i.e. related to the mind), there is no such term in the enactivist definitions, indeed, no reference to mind whatsoever. Does it mean that according to enactivists, cognition is *not* a mental issue? Yes and no: It is a mental issue in the sense that it relates to the mind; but it is not a mental issue if we take "mental" as the opposite to "physical/bodily". Enactivism has developed from (and builds upon) the crucial idea that cognition (and the mind itself) is *embodied*.[2] As the term suggests, this approach holds that the body needs to be taken into account when thinking about cognition. This general idea can take various forms, differing among other things in the degree to which embodiment is thought to determine the overall theory of cognition (see the abstract of Di Paolo and Thompson's 2017 preprint for a short overview). The enactive approach draws upon the *radical embodiment* theory, which "(...) is the view that cognition ought to be understood *primarily* in terms of the embodied agent–environment dynamics. Neural dynamics can only be studied while taking into account the larger brain-body-environment dynamics," (Bruineberg, 2018, p. 40, with reference to Chemero, 2009). To fully appreciate this requirement, I will now take a step back and sketch what is the position of the cognizing subject from the perspective of this type of embodied and enactive approaches.

We can see the whole world as a complex dynamic system the organiza-

[2] A nice illustration of this is the fact that the book by Varela et al. (1991), where enactive approach to mind has been introduced, is called *The Embodied Mind*.

tion of which is constantly changing. Elements and processes of this system are related to other elements and processes in various ways and influence each other to different degrees. Some parts of the system are so tightly interrelated that they can be considered units – i.e. subsystems that have their own distinguishable internal dynamics, but are nevertheless still parts of the whole system, constantly influencing it and being influenced by it. We can find such subsystems at various levels: rainforest might be one; tree another; stone; bacteria; human; liver; brain (what one defines as a unit depends largely on conventions, goals of the observer, etc., as Di Paolo & Thompson, 2017 point out). But since the organization of the whole system is in a constant state of flux, all the subsystems are subject to changes as well. Their organization can change to such a degree that they cease to be the kind of subsystems they were: a tree falls down, rots and dissolves into the soil, at which point we would no longer call it a tree.

This description of the world might seem self-evident when illustrated by the above examples. It is however not so trivial to retain this perspective while theorizing about human cognition. If we think of cognition as "a mental action or process" (remember our first definition), the general world-view I just sketch might even seem irrelevant. One might simply want to focus on what is going on in human heads in certain situations (e.g. while listening to speech), without thinking about how the world is working as a whole. From the enactivist perspective, however, the big picture is crucial, because cognition is defined in terms of the dynamics going on between the (living and therefore cognizing)[3] subsystem and the larger system it is a part of. As Thompson (2007, p. ix) puts it: "(...) mental life is also bodily life and is situated in the world. (...) Our mental lives involve our body and the world beyond the surface membrane of our organism, and therefore cannot be reduced simply to brain processes inside the head." In what follows I will describe the agent-environment dynamics that constitutes cognition in

[3]This implication is not at all self-evident; I will elaborate on it in the following subsection.

detail.

1.1.2 Mind-life continuity

Let us get back to the description of the world as a large system in a constant state of flux, which makes the existing subsystems dissolve and new ones emerge. Some of the subsystems actively resist this fatal change of organization – their internal dynamics and interactions with the environment counteract the natural tendency of (sub)systems to disorganize.[4] Such systems may be called *autonomous*[5]. According to some scientists (Bruineberg, 2018, p. 28; Friston & Stephan, 2007, pp. 4–5), this property defines systems that are *alive*: what differentiates a living creature from a dead creature is that the latter does not do anything that could prevent it from disintegration, whereas the former does (it eats, seeks shelter, regenerates if hurt etc.).[6]

For enactivists, the definition of life is of great importance, because they adhere to what is called the *mind-life continuity thesis*. It states that there is no qualitative difference between the principles of what we call "mind" (or "mental capacities" etc.) and the principles of life, i.e. of what makes a piece of matter organic. Or as Thompson (2007, p. ix) puts it: "Where there is life there is mind, and mind in its most articulated forms belongs to life," and that is because "(...) life and mind share a set of basic organizational properties, and the organizational properties distinctive of mind are an enriched version of those fundamental to life," (p. 128). This is not to assign all living

[4]By this "natural tendency to disorganize" I mean the fact that entropy, in other words disorder, of a closed system inevitably increases with time – i.e. the second law of thermodynamics.

[5]Di Paolo and Thompson (2017) explain in detail how an autonomous system is defined. One of the two constitutive properties (called *operational closure*) is that the processes that constitute the system are in the relation of mutual dependence – if some of them stops, the others stop as well. The second constitutive property, *precariousness*, concerns the natural tendency of any system to disintegrate, against which autonomous systems act – see the main text.

[6]For a nice formulation of this position (and an explanation of the conceptual connection between life and the second law or thermodynamics) see Friston and Stephan (2007).

organisms the same level or type of mental life or even of (self)consciousness – the mind-life continuity thesis is not blind to the huge difference between a person contemplating a philosophical book and a bacteria navigating itself through the environment. But neither is it to claim just that life is a necessary condition for the presence of mind, so that where is mind, there has to be life, but not vice versa. Bruineberg (2018, p. 25) cites an illuminating formulation from Colombetti: "(...) the idea here is that the autonomous and adaptive organization of living systems sets up an asymmetry between them and the rest of the world, such that living systems realize a perspective or point of view from which the world acquires meaning for them, and not vice versa (Colombetti, 2014, pp. 19–20)." What I understand under this "realizing a perspective from which the world acquires meaning" is nothing else but cognition. What the mind-life continuity thesis says, then, is that cognition arises inevitably with life; and this should be taken into account if cognition is to be fruitfully studied.

I adhere to the mind-life continuity thesis for the sake of parsimony – if the mind can be understood on the same basis as life, I see no reason for not doing so and treating it as something qualitatively different, governed by its own principles. Another, admittedly less scientific, reason is that I intuitively deny to see any substantial difference or gap between people and other animals. I believe that a too strong emphasis on our intellectual capacities, which are moreover typically seen as human-specific, can hide from us important and informative correspondences and similarities between us and other species and create an impression of our disconnectedness form the rest of the living nature. Letting the ethical aspect of this issue aside, I consider contrasting humans with other species by default also scientifically inconvenient, because it might constrain our research on humans within unnecessarily narrow bounds, instead of encouraging its cooperation with and inspiration by research concerning other living creatures.

1.1.3 The purpose of cognition

I hope to have clarified what the enactivist interpretation of the first part of the dictionary definition – i.e. "cognition is a mental action or process" – would be; and that if the word "mental" is understood from the perspective of radical embodiment, it might in fact be left as it is. The next part – "cognition is a process of acquiring knowledge and understanding" – concerns the goal or purpose of cognition, and here enactivists would need more than a careful (re)interpretation of the included words to accept it. What was said earlier by itself makes such formulation problematic – we might not want to attribute the activity of "acquiring knowledge and understanding" to the most basic forms of life, such as simple cells, but as explained above, we do want to ascribe to them some form of cognition. But there is something even more suspicious about the formulation, because one might ask: Why on Earth should the organism strive to acquire knowledge and understanding (of the world, presumably)? This question might sound naive because there is an obvious answer: An animal (to make it less abstract) needs to have knowledge about the world around itself in order to survive, i.e. in order to make the right decisions about where to get food, what animals to fear etc. But then the knowledge is not a goal in itself, it is instrumental; and it is legitimate to ask whether the ultimate goal, i.e. survival, really requires precisely the instrument of acquiring knowledge – and if yes, then what kind of knowledge is needed for that.

The only thing we have said about living systems so far is that what makes them living is their autonomy, i.e. their ability to actively resist disintegration (for some time). The goal of their behaviour is therefore encoded in the definition of life itself – to be alive *means to* exhibit self-maintaining behaviour. And this behaviour in turn founds a specific relationship between the system and its environment – and this relationship is cognition. Concepts like "gaining knowledge" imply much narrower and more anthropocentric view of cognition than the one just sketched. Crucially, "gaining knowledge" also makes cognition seem as a unidirectional process: the world

remains intact while a cognizing subject gains information about it – the flow of change goes, so to speak, only from the world to the subject. In contrast, the enactive approach highlights the interactive nature of cognition, as is expressed in the two enactivistic definitions of cognition cited above: one defines cognition as "adaptive regulation of states and interactions", the other as "exercise of skilful know-how in situated and embodied action". How the organism influences its environment is no less a part of cognition than how the environment influences the organism (with a caveat, one might say that action is no less a part of cognition than perception)[7]. This is because the essence of cognition is the system's autonomy, i.e. its *acting against its own disintegration*, and not a deterministic need to gain information about the world or an internal representation of it.

However, it would be wrong to think that from the enactivist perspective, the cognizing organism does not learn or know anything. Recall the second definition, according to which cognition is the "exercise of skilful know-how". The point here is that the organism does not learn facts *about* the world, because it does not need to know *that* (something in the world is this or that way), it needs to know *how* (to act in order to stay alive). Its "knowledge" has the form of a skill, not of a storing or representation of facts. Knowledge being a skill is one of the most crucial ideas for this thesis and will be elaborated on extensively in the following chapters. For now we only need to bear in mind that perhaps the most important difference between the common view of cognition (as found e.g. on Wikipedia) and the enactive approach is in the task or purpose they ascribe to cognition. If we define cognition as "acquiring knowledge", the task is to somehow take what is *outside* and get it *inside* (into our mind/brain, presumably in a different form, e.g. as a representation).

[7]The caveat is that, at least in the branch of enactivism that I will primarily work with (Bruineberg, 2018), to separate action and perception as two distinct processes running in opposite directions is inadequate. "There is no separation between perception and action because perception on our account just is the organism's preparing itself to act (...)." (Bruineberg, 2018, p. 79) I will return to the relationship between action and perception later (section 1.2.3).

However, the task for *enactive* cognition is to steer the actions of the organism "with respect to the consequences for the agent's own viability", or in other words, to ensure skilful interactions with the environment.

1.2 Skilled intentionality framework

The main source of my knowledge about enactive cognition (and also the original and most important inspiration for this thesis) is the work done by Jelle Bruineberg and his colleagues roughly between the years 2013 and 2018, as reported in Bruineberg's (2018) dissertation "Anticipating Affordances. Intentionality in self-organizing brain-body-environment systems". In this work, Bruineberg develops what he calls the *skilled intentionality framework* (SIF), a conceptual framework meant to provide a common ground for all scientific fields concerned with cognition (from neuroscience to philosophy of mind).

First of all I want to make clear what exactly SIF is and what we can and cannot expect from it. Bruineberg is very clear on this point: Conceptual framework (such as SIF) is "a tool to make explicit and intelligible the concepts and structures that constitute, underlie or in some other sense relate to a particular phenomenon" (Bruineberg, 2018, p. 33). He highlights the difference between a conceptual framework and a theory: While a theory should make predictions that can be empirically tested, a framework is one step before that – it is the conceptual space within which theories are articulated. Whether a framework is good or not depends on whether it is internally consistent and whether it helps us to think about the subject matter in a productive and useful way. Works like the present one are therefore desirable follow-ups to Bruineberg's endeavour; in fact, they are necessary if the framework should fulfil its goal. Within Bruineberg's thesis itself, the application of the framework to concrete phenomena is limited to brief examples serving mostly to illustrate or clarify some theoretical points [8]. However, there have

[8] An exception is section 1.7 where applications of SIF (on boxing, ice-climbing and

been several attempts to take a further step with SIF, mainly from the co-authors of the framework (to my best knowledge). Of particular interest for us are the works that are concerned with so-called higher cognition[9], because that is the category that language falls into. Rietveld and Kiverstein (2014) and Bruineberg (2018, chapter 4) sketch the general lines along which applications of SIF to higher cognition might go. They propose that one should take our sociocultural practices simply as another source of regularities in our environment, with the same functionality as regularities in our physical surroundings (van Dijk & Rietveld, 2017). Rietveld and Brouwers (2017) and van Dijk and Rietveld (2018) apply SIF to architectural design practices and long term planning in this field, which are examples of "higher cognition" par excellence. Van Westen, Rietveld and Denys (2019) used a SIF-like approach to analyze the way how clinicians assess the effect of deep brain stimulation on patients with an obsessive-compulsive disorder. I am not aware of any attempt to apply SIF directly to language, although Bruineberg considers it possible and encourages the idea (personal communication, June 2018). Applying the skilled intentionality framework to language is the aim of the present thesis.

One of the main pillars of SIF, enactivism, has been introduced in the previous section; the reader should therefore have an idea about what is meant by the "self-organizing brain-body-environments systems" which are at the core of Bruineberg's book and will underlie my approach to language in the chapters that follow. In the remainder of this chapter, I will cover a few more of those pillars – namely the ones I consider important for the attempt to integrate language behaviour into enactive cognition as instantiated in SIF.

mood disorders) are addressed slightly more extensively; it is however still rather a taster.

[9]In fact, SIF questions the division between "lower" and "higher" cognition itself, showing that they might very well work on the same principles (Bruineberg, 2018, chapter 4).

1.2.1 Free energy principle

In section 1.1, I adopted the view that living organisms are defined as autonomous dynamic systems, i.e. systems that interact with their environment in a way that ensures the maintenance of their organization. Moreover, I presented the idea that the same principle that drives this specific dynamics (making a system alive) underlies also cognition in all its forms. A reader to whom this sounds rather abstract and vague will be pleased that the principle has a name and a mathematical expression; and it has already found its place in (neuro)biology. I will try to explain the principle in a rather intuitive manner, without going deep into technicalities (an interested reader may find more detailed and technical explanation in Bruineberg, 2018, chapter 5).

Imagine a set of values, say (to adopt Bruineberg's illustrative example[10]) temperatures of my body, measured repeatedly over a long time period. The data have a probability distribution peaking sharply somewhere around 36.5 degrees; values around 35.5 or 39 are much less probable, values such as 20 do not occur at all. Now I can take a new measurement and compare it to the distribution to see how surprising the new value is – a value of 30.1 would be very surprising, a value of 36.6 not at all. The distribution would be different for, say, a rabbit, but the basic pattern is the same – there will be a rather small set of very common values surrounded by a wide range of uncommon ones.

Let us move from the concrete example of body temperature to the overall state of the organism. We can treat all parameters that characterize the system and its relationship with its environment (such as temperature, blood pressure, saturation with water, etc.) as dimensions of a multidimensional space – the *state space* of the organism. We can imagine the organism as moving through this space as the values of the parameters change; "to move" in this space means to get from one possible state to another. We can apply the same procedure here as with the temperatures: If we keep track of the states a particular organism goes through during a longer period of time,

[10]Bruineberg, 2018, p. 139.

we find a distribution where a few states are very common and many others are rare or unattested; in other words, we find that the organism usually occupies a rather limited set of states. In information-theoretic terms, we say that the distribution of states of the organism has low entropy, or, to get closer to Bruineberg's terminology, that the states of the organism have low average surprisal[11]. This is intuitive, because if we sample such distribution, we most of the time get one of a the few common values, so the outcomes will not be surprising for us.[12]

The shape of the distribution is, according to Friston (2010), not accidental. As explained in section 1.1.2, a living system strives to maintain its organization in an ever-changing environment – it tries to avoid all changes of state that would irretrievably disrupt its integrity. To be on the safe side, the organism needs to occupy the states that are unsurprising – if they are common for the organism, then they must also be viable for it, unlike all the others. To illustrate it on the body temperature again: The unsurprising temperatures around 37 degrees are the ones that are viable and most convenient for the organism, whereas highly surprising values such as 32 or 42 are dangerous. In order to survive, the organism therefore needs to keep the average surprisal of its states low, and this is only possible if the distribution of all its possible states has low entropy – a requirement which the sharply peaking distribution described above meets. The shape of the state distribution is therefore not an accidental property that also happened to be convenient for the organism – if living systems are to maintain their organization as they do, the distribution of the states they occupy *has* to be shaped this way, i.e. a small set of states has to be very common and the others very rare or unattested.

The insight that living systems need to minimize surprisal in order to

[11]The term "surprisal" is often used in this context to distinguish the measure from the non-technical, emotionally charged "surprise".

[12]An example of a distribution with high average surprisal would be a uniform distribution, where all values have the same probability of occurrence. We cannot predict what value is about to come, so we will be "surprised" by the outcome all the time.

stay alive is the core idea of the free energy principle; and for the purpose of this thesis, this non-technical level of understanding is sufficient. I will just add a brief remark to make clear why the principle contains the term "free energy", and not the term "surprisal".[13] The organism cannot minimize surprisal directly (for details, see Bruineberg, 2018, p. 143); there is, however another quantity, free energy[14], that the system does have a direct access to. Crucially, free energy always has a higher value than surprisal. The system can therefore implicitly minimize surprisal by minimizing free energy.

1.2.2 Modelling the world?

In the previous section, we saw that the organism needs to minimize free energy in order to stay alive. The crucial question then is: How?

To understand how an organism can influence its free energy, we have to take a step back and look at the factors that influence the states of living systems. It is definitely the internal dynamics of the system, that means, what is happening inside of it. But the system is also not independent from its environment, it is constantly interacting with it – the external dynamics influences the system's states, and the system in turn influences its surroundings. These interactions, too, need to be taken into account by the system if it should have control over its states, and not only its internal states as such. However, while an omniscient observer knows what the state of the world is and what consequences will a change in the environment have for the system and vice versa, the system itself must make do with estimates based on its own previous experience (and presumably also the experience of the whole species, transmitted genetically).

[13] I again refer an interested reader to Bruineberg (2018, chapter 5), for a mathematical explanation.

[14] Free energy is a concept employed in various scientific fields (e.g. chemistry, thermodynamics, statistics or information theory); its precise definition therefore understandably varies according to the field. For us, the information-theoretic quantity is relevant, being defined as a "measure that is an upper bound on the surprisal of some data, given a generative model" (Bruineberg, 2018, p. 52).

From here, I will take a short digression to describe an influential view of human cognition which makes use of Friston's free energy principle and shares some key features with SIF.

A famous theory of how human cognition works is the theory of predictive coding/processing[15]. Proceeding from the assumption that the brain has no direct access to the world, but nevertheless needs to take it into account, this approach sees the brain as an inferential and predictive machine. It is inferring the causal structure of the world from sensory input, creating an internal model of the world and trying to predict sensory input that will come next – if the prediction is wrong, the brain updates the model, trying to minimize prediction error as much as possible. The brain then in a sense overcomes its radical separation from the environment – by modelling it.

This approach is important to us for two reasons: First, it is commonly thought to form a natural tandem with Friston's free energy principle (Bruineberg, 2018, pp. 66–67); and second, it can seem so close to SIF that it is worthwhile to explicitly point out the differences between SIF and the version of the predictive coding theory described above. The two points are connected: In chapter 2, Bruineberg (2018) argues that the free energy principle fits also (and in fact better) with the enactive and more holistic approach to cognition offered by SIF than with the standard theory of predictive coding. I will now therefore present the relevant specifics of SIF.

Predictive coding is clearly a theory of the workings of the *brain* (Clark, 2013). The brain is treated as the one that perceives, that predicts, that cognizes. SIF, on the other hand, emphasises that it is the whole organism

[15]For an accessible overview of this approach (more specifically one of its offshoots: the prediction error minimization framework), see Clark (2013). It should be noted that Clark takes into account also a view of cognition that is in line with the skilled intentionality framework, calling it "action-oriented predictive processing". He finds it appealing in some respects, but is not convinced by its most radical form, which is the one that is advocated in this thesis.

that does these things. Our body is not just a vehicle that carries the brain and provides it with sensory input; it is an equal part of the system. As Bruineberg (2018, p. 57) points out, we can sometimes treat the brain separately to some extent and it can be a useful perspective. It cannot, however, be our starting point for the research of cognition – an obvious reason is that many living organisms, all of which are also cognizing (see section 1.1.2), do not even *have* a brain. This is an important adjustment to the theory of predictive coding: If the internal dynamics that is said to model the environment comprises the whole organism, than we ought to say, together with Bruineberg and Friston, that "an agent does not *have* a model of its world, [but rather] it *is* a model" (Friston, 2013, cited in Bruineberg, 2018, p. 74). When there is a need for adjustment of the model because some very surprising input arrived, what changes is not some picture of the world that the organism draws in its "mind" – it is the whole system that changes, because it does not create a representation of what it has learnt about the world, it *embodies* it.

This view has one encouraging theoretical consequence: The alleged separation of the cognizing subject from its environment seems much less severe. When introducing the theory of predictive coding, Clark (2013, p. 183) describes the alleged unenviable situation of the brain: "[the brain] must discover information about the likely causes of impinging signals without any form of direct access to their source. (...) The world itself is (...) off-limits (...)". But under SIF, the cognizing subject is not a brain secluded within the skull, having to infer causes of the input it gets from the senses, but a system that interacts directly with its environment all the time. As shown in section 1.1.1, the system can even be approached as a part of a bigger system comprising itself *and* its environment, so deeply is it entangled in the world. This takes a great burden off the shoulders of cognition and makes us frame our questions differently.

Another point concerns *what* exactly is being modelled (by the brain or by the whole organism, depending on the theory). In Clark's account, the

organism models "the nature of the signal source" (Clark, 2013, p. 183), or in other words: "how the world is" (p. 182). This would be in line with the idea that the purpose of cognition is gaining knowledge about the world, which is the traditional view of cognition that I tried to argue against in section 1.1.3, and instead adopted the view that cognition should enable the organism to act efficiently against its disintegration. To make this happen, the model embodied by the organism must contain a crucial component: the organism itself in a viable state. If the organism only modelled the world as it is, regardless of the organism itself, then even if it was successful and modelled the world accurately, the model itself would do nothing for the organism's survival. It could definitely be *used* to this end, but there would have to be another force at work – will, intention, goal-directedness. In combination with such a force, a model of the world could perhaps be sufficient to ensure survival. But this would be quite a costly solution – it is not at all clear where such a force should be coming from and how it should biomechanically operate in the organism, so we would be trapped in the ancient philosophical puzzle of the relationship between body and soul. The beauty of Friston's and Bruineberg's view is that they do not need this extra element. If the organism models not just the environment, but the environment *including* the living organism itself, then the *intentionality* of the organism's behaviour is already hard-wired in the model it embodies, i.e. the kind of system it is. In the following section, I will elaborate on how this works.

1.2.3 Action and perception

The need for modelling the world arose because we saw that the organism must take it into account if it wants to have control over its states. The fundamental need is still to keep the organism within the limited set of its viable (= unsurprising) states, which the organism can achieve by minimizing its free energy.

If we see free energy as a measure of the misfit between the organism's internal model (which defines its expectations) and the reality, there seem

to be two options how to minimize it. First – the organism can change its internal model so that it fits better the input it gets from the environment. To give a prosaic example: I see a tap with a blue mark on one side and a red mark on the other. I turn it to the blue mark and expect, guided by my experience, that cold water will flow out. However, hot water starts to flow – so I change my internal model of this part of my environment and I will now correctly expect hot water to be flowing. This is what is traditionally called *perception*. Then there is the second option: The organism can take *action* and change the environment so that it fits better its internal model. In the above example, this would correspond to turning the tap to the other side – cold water is now flowing out, in accordance with my original expectation. It is the environment, not my expectations, that has changed.

Described in this way, perception and action may seem as two distinct strategies of free energy minimization, two processes going in opposite directions (Howhy, 2016). Bruineberg (2018), however, takes up a different perspective that is in accordance with his emphasis on the deep union of an organism and its environment. He prefers to speak about a *brain-body-environment system* and interprets the interactions between an organism and its environment as "reducing a disequilibrium" in this system as a whole (Bruineberg, 2018, p. 63). Both perception and action are then just two aspects of the same continuously ongoing process, there is no clear separation between them. If "perception" means a change of the organism's internal state, then we have to say in the same breath that the "internal" state (and nothing else) defines what the organism is currently doing, i.e. its action; and that its interaction with the environment is *constant*, even if it seems rather passive from the outside (e.g. when an animal is just standing still). This is not to say that distinguishing between perception and action is not useful at some levels of analysis; but once we have built a certain picture of what a living system is and are trying to use that as a basis for further analyses, it is important to realize the implications this approach has for the concepts we commonly make use of. If SIF should bring us anything useful,

we have to let it inform our view of precisely the concepts that lie right at the heart of our field; and "perception" is no doubt an important one in cognitive science, including the sciences of speech and language. The shift is then from the view of perception and action as separate mechanisms to the view that they are just two aspects of the constantly ongoing process of mutual tuning of the organism and the world. Perception does not somehow extract information from the environment and transfer it to the agent's brain, so that the brain can then plan what to do and eventually launch a command to do it. Perception brings about a state of the living system that we can interpret directly in terms of action; in Bruineberg's words, a state of action-readiness (Bruineberg, 2018, p. 69).

This perspective implicitly prioritizes action over perception (if we are still treating them separately). That is correct, because in the end, it is action that effectively leads to minimization of surprisal. We have to bear in mind that the organism embodies a model of the world *with itself, alive, in it*. Surprising situations are therefore the ones that bring the organism in danger, and these need to be actively avoided, and not just truthfully perceived. Let us use the example with the tap once more: If the agent just adapts its expectations, she will get burnt by the hot water. What she needs instead is to actively change the situation so that it is not surprising (= dangerous) for her anymore. This is how being a model of myself within my environment enables me to survive – it makes dangerous situations surprising and the tendency to minimize free energy therefore drives action that keeps me away from them.

1.2.4 Skilfulness

In the previous sections, we unpacked the term "intentionality" from the skilled intentionality framework and saw that action is crucial for an organism's survival. Obviously, it cannot be *any* action – the organism needs to act in a way that helps it to maintain its organization. From all the actions that are possible in a certain situation, it should select the right one. How

can the organism do that? We should not content with the metaphor of "selection", as if possible actions were laid in front of the agent and she would then contemplate about which one to choose and why. Because all we have at our disposal is a living system with a certain state space and with the ability to move (and to be moved) from one state to another; and where it moves depends on the situation it finds itself in and on how the state space is orgnanized. It is the second point we will focus on in this section.

In theories of dynamic systems, we find a useful concept of *attractors*. An attractor is a state into which the system tends to evolve whenever it finds itself near it in the state space. If we imagine the state space as a landscape the system is travelling through, we can compare an attractor to a valley – if the system gets to the slanting area, it is going to slip down to it. In other words, the system cannot do anything at any point in time – it is attracted to certain states in certain situations. That shows the idea of it "choosing" the action it will take in a different light: The action is actually already pre-selected by where the system currently is in the state space and by what attractor is nearby. The key to an appropriate action selection lies therefore in previous shaping of the state space. This might sound a bit obscure, but it is nothing else than building a skill. All of us can certainly recall the experience of becoming skilful at something, say, driving a car, playing a ball game, or dancing. At the beginning, our reactions are often wrong or too slow (or both), but with practice, we become able to react adequately to what is happening – we change gear automatically at the right moment; we no longer step on our dance partner's feet. We do not have to think about it, quite the opposite, thinking would slow us down – we perform these actions automatically, which is what makes us skilful. The space of our possible actions is shaped in such a way that a certain kind of input (such as a specific sound of the car engine) immediately triggers a certain reaction (changing gear).

We are now getting to what is meant by the "skilled intentionality". Skilfulness was the missing piece of the puzzle: We saw that the living system

has to stay within a certain set of states and that it needs to act in an appropriate way to achieve that, and skilfulness is the answer to the question *how* the system can do that. It acts appropriately if among all possible actions, the right one is attracting it. This is what Bruineberg (2018) calls "selective openness to affordances" (where "affordance" is an action possibility, see Bruineberg, 2018, p. 42) – a skilful organism automatically responds to the right action possibility because of how its state space is organized. The right action possibility stands out as the most relevant, because it corresponds to the largest valley in the surrounding space.

What is implicitly present in the view of organisms I am presenting here (and should not be missed) is the anticipatory nature of its existence. The agent never simply reacts on what has come to her from the outside – she has a strong opinion (to put it metaphorically) on what *is going to* come, namely, she assumes that she is going to survive and thrive, and it is this anticipation that shapes her own organization in such a way that she is constantly working towards its fulfilment.

1.3 Takeaway from chapter 1

In this section, we went through the tenets of the skilled intentionality framework that are important for this thesis. To summarize it briefly:

Friston's free energy principle states that a living system needs to stay within a limited set of unsurprising states, i.e. to keep its free energy low. In order to do this, the system has to embody a model of its environment with itself included in it. This introduces intentionality in its behaviour and drives a process of mutual influence that keeps the whole organism-environment system in a setting that is viable for the organism. Perception and action, rather than being two distinct (opposing) processes, are two aspect of this one process of constant organism-environment attunement – perception naturally brings about action-readiness. What action the organism will perform depends on what situation it finds itself in and on what are her (= the dy-

namic system's) attractors, i.e. which action it inclines to in the particular context. A thriving organism is therefore the skilful one – the one that is automatically attracted to the actions that help her stay within the set of her viable states.

Chapter 2

Linguistic background

So far, I intentionally stayed in a respectful distance from language. I treated humans just as one species among many others and made sure that the principles and ideas presented in the first chapter were applicable to any living system, no matter how simple it may be. In this chapter, I will proceed from the other end and will present a linguistic theory that I consider compatible with the above theory of cognition; the way they could be synthesized will be explored in chapter 3.

The linguistic theory that will be presented here is currently being developed by Michael Ramscar. His joint paper with Robert F. Port "How spoken languages work in the absence of an inventory of discrete units" (Ramscar & Port, 2016) can be recommended as an accessible presentation of the main idea and some of the crucial arguments; the most current and comprehensive version of the theory can then be found in Ramscar's (not yet published) paper "Source codes in human communication" (Ramscar, 2019). I should note that my choice of this theory does not depend solely on its compatibility with SIF. I find Ramscar's view of language as a discriminative code the most convincing and promising theory of language I have encountered so far, and the fact that it seems to be compatible with SIF is just one of the properties that make me think so. In this chapter, I will present the main tenets of Ramscar's approach with an emphasis on where it is at odds with

the view of language I have mostly encountered during my language-related studies.

2.1 Discrimination vs. composition

There is a parallel between Bruineberg and his colleagues' shift in the view of cognition and Ramscar's shift in the view of language, in that both of them start from the basic *function* of the phenomenon of interest, and both of them see the function in a non-mainstream way. For Ramscar, the nature of human communication is discriminative – we understand what someone said because we discriminate it from all other things she might have said, and this is done by a process of continuous prediction and elimination (Ramscar & Port, 2016; Ramscar, 2019).

The idea that discrimination plays a crucial role in communication is not in itself new. At least within the study of spoken communication, i.e. phonetics and phonology, the importance of discrimination is ubiquitous. To give just three examples: 1) Phonemes, i.e. the traditional invariant units of the sound layer of language, are notoriously defined through their potential to differentiate meaning bearing units (Trubetzkoy, 1969; Skarnitzl, Šturm, & Volín, 2016). 2) Listeners' (in)ability to discriminate between speech sounds is seen as indicative of phoneme boundary placement (Liberman, Harris, Hoffman, & Griffith, 1957). 3) Troubles with foreign speech learning are to some extent ascribed to the learners' inability to discriminate sounds in the target language (Escudero, 2005, pp. 1–2). However, acknowledging the importance of discrimination in a certain sub-area is not the same thing as considering all communication a fundamentally discriminative process and treating language as tailored to meet the needs of such process. The usual approach that most students of linguistics encounter and language scientists typically work with (and that is deeply rooted also in my own way of thinking about language) is *compositional* rather than discriminative: It states that language users are trained to recognize meaningful units such as words

or morphemes (this is where discrimination plays its part) and then compose the meaning of the whole utterance by adding up the meanings of the units according some (syntactic, pragmatic...) rules. So when they hear for instance "I bought a pan", they identify the meaningful elements, i.e. for example discriminate "pan" from "pen" or "bought" from "boat", and because they are familiar with the SVO[1] structure of English sentences and know that "bought" is a past tense etc., they combine the elements and gain the information that the speaker bought a pan.

In the first three sections of their paper, Ramscar and Port (2016) address a weak spot of the compositional approach to communication: its need for discrete units. If language users get the meaning of an utterance from combining smaller elements, such elements need to be sufficiently invariant for us to identify them repeatedly and in some way attached to the meanings they are supposed to contribute to the final message. Ramscar and Port (2016) summarize problems that we encounter when we search for units such as phonemes, morphemes and words in spoken communication – it turns out impossible to work with these units in a consistent manner. Phoneticians know that well from their everyday practice: in a continuous stream of speech, clear boundaries (be it between phoneme-sized units or words) are rarely found (Machač & Skarnitzl, 2010); contextual variability of all potential candidates for such units is enormous; context and expectations can change profoundly which "unit" is perceived; subphonemic details are demonstrably perceived and exploited by listeners[2]. We, as linguists, can usually find our way around and determine which units were supposed to be uttered and what happened to them (which ones where omitted, which ones blended, which of them listener reconstructed from context etc.). It, however raises the question whether human communication indeed rests upon such units, or whether linguists are just trying to use their deep-rooted linguistic

[1]Subject–verb–object.

[2]For example, Dutch listeners seem to use duration of word stems as an additional cue for the difference between singular and plural forms of nouns, which is marked also morphologically (Kemps, Ernestus, Schreuder, & Baayen, 2005).

ideas to deal with something that in fact works in a different way. My personal impression is that a lot of us linguists are somewhere in between: We would not claim that speech perception actually resides in identifying and combining phoneme-sized units or that speech signal can be unambiguously segmented into words, but we nevertheless make heavy use of the common types of discrete units, and more importantly, we keep on thinking about language and communication in compositional terms. The issues mentioned above are treated as mere complications that need to be explained away or dealt with by adjusting the common view of language processing a little bit, and not as reasons to take a radically different perspective. This is why I find Ramscar's (2019) paper important: it presents also evidence for the discriminative view of language independent from the problems with discrete units (namely certain statistical patterns observed in language corpora). When this evidence is combined with the fact that there *are* obvious problems with the compositional view of language which the discriminative perspective does not run into, the discriminative approach appears quite appealing. I will now proceed to explain how discriminative communication works and what it would mean to see language within this frame.

2.2 Source codes in discriminative communication

An essential inspiration for Ramscar's view of language is Shannon's (1948) theory of communication. Shannon was interested in artificial codes and efficient transmission of messages through noisy channels. In this context, "communication" looks like this: A sender chooses a message from a given set, familiar to all participants, and encodes it (ideally with some level of compression), then sends it through a channel (where noise might intervene); and a receiver identifies which one from the set it was – i.e. she discriminates the actual message from all other candidates. The question for information theory is how to, on the one hand, compress the message in order to make

the transmission as efficient as possible, and on the other hand maximize the probability that the receiver will identify the message correctly. Note that information theory is not concerned with what the messages *mean*, they might very well not "mean" anything – the source of messages can be just a sequence of random numbers and the question, as well as the answer, remains the same.

In the first part of his paper, Shannon (1948) shows how statistical properties of the source code (i.e. the message generator) influence the informativity of the message. If, for instance, we have two possible messages A and B, then if the probability of A being sent is 0.9 and the probability of B is 0.1, the message is not very informative – we could have guessed the outcome even before we received the message and we would be probably right. If, however, the probability of A and B being sent is the same (0.5), the message is highly informative – without it, about 50 % of our guesses would be wrong. Informativity, seen from another angle, is the receiver's *uncertainty* – in the first case, her uncertainty about the message is very low, whereas in the second case it is maximal. The information-theoretic measure of informativity/uncertainty is nothing else than the famous Shannon's entropy – the reader already knows it as the measure of the average surprisal from section 1.2.1. Even from the primitive example with just two candidate messages A and B, it is obvious that the statistical structure of the source has a profound impact on communication – no matter how messages are encoded and what noise is added during transmission, the statistical properties of the source themselves can increase the chance of correct identification consirerably.[3]

Several linguists got inspired by information theory and applied some of its principles to language (see Ramscar, 2019, p. 3, for references). However, if discriminative communication is supposed to work and be successful, an important condition has to be met: The probabilistic structure of the source

[3]Encoding of course also matters – it determines the degree of compression and, if chosen well with respect to the properties of the intervening noise, can further increase the chance of correct identification of the message.

has to be known a priori not only to the sender, but also to the receiver. If the receiver does not know that message A has the probability of occurrence 0.9 and message B only 0.1, the mere fact that it is the case does not help the receiver with her task at all. The question is whether this condition of a shared source code is met in spoken communication. Ramscar (2019), instead of approaching language from information-theroretic perspective right away, pursued the question whether human communication *could* possibly work discriminatively. Another way to put this question is: Is language, the source code for human communication, *shared enough* to approximate the communication situation addressed by information theory? By means of analysis of mainly (but not exclusively) proper names in various languages and various time periods, Ramscar shows that language has statistical properties that not only can make it "shared enough" in the relevant sense, but which are also hard to explain if we do not consider language a discriminative code and stick with the compositional account instead.[4] The information-theoretic perspective therefore seems not only justified, but even appropriate.

Let us therefore take seriously the idea that spoken communication works, at least to a substantial degree, according to the scheme sketched at the beginning of this section, and that language is a source code for such discriminative communication. A natural question to ask is: Where is meaning in this picture? As Shannon (1948, p. 379) famously stated, "semantic aspects of communication are irrelevant to the engineering problem", and he was thus not concerned with meanings at all. Ramscar and Port (2016), however, naturally were, because human communication obviously requires that the code we use has some relation to the world. In their theory, meaning directly participates in the process of message transmission and identification rather than being its result. Meanings are not *encoded in* messages, instead, semantic contrasts guide the discriminative process in parallel with acous-

[4]The crucial property are geometric frequency distributions found in various functionally relevant domains.

tic (and also visual and contextual) ones and are related to them by mere correlation (Shannon, 1948, p. 379; Ramscar, 2019, pp. 32–33; Ramscar & Port, 2016, p. 72). The important point here is that the communicative code is not structured by the need to encode and transfer meanings, but by the needs of a real-time process of discrimination.

An example might help illustrate the point. If one supposes that meanings are encoded by units of language (say, words) and that communication serves to transfer them from speakers to listeners, it seems reasonable to expect that the frequency of occurrence of a particular unit will be governed by how frequently we need to communicate its meaning. So if we take for example colour terms in English, their frequencies should probably reflect how often we need to speak about something that is red, white, blue etc. (of course, also metaphorical uses of the words would have to be taken into account, but let us disregard them for now). It is of course not easy to guess the frequency of our need to speak about differently coloured things, but the distribution should probably be a function of the number of so-and-so-coloured objects in our environment and the significance of these objects in our everyday life. Since the colours of objects that surround us are rather unrelated to each other, there seems to be no reason to expect any specific relation between the frequency of one colour term and the frequencies of the others. But when Ramscar (2019, pp. 59–60) collected English colour terms in the Corpus of Contemporary American English (COCA), what he found was a highly skewed distribution where the frequencies of individual colour names decrease geometrically: White has the highest frequency; when divided by a constant, we get the frequency of black, which we can divide by the same constant to get the frequency of red; the same division gets us to green etc. (in other words, the frequencies build a geometric progression). What does that mean? Do we really need to refer to white object more often than to black ones, and to black ones by the same factor more often than to red ones, etc.? Does this reflect the distribution of coloured objects in our environment? Or perhaps some kind of cognitive bias influencing how

we treat coloured objects? Or is this a result of the metaphorical and other specific uses of colour terms that we dared to disregard? None of these is impossible, and Ramscar (2019, p. 60) himself notes that skewed distributions do not directly contradict the more traditional (i.e. compositional) view of language. But when one sees very similar pattern in other semantic groups (first names, kinship terms, "throw" verbs... see Ramscar, 2019, for more), the question arises whether there could be an underlying principle that shapes the frequency distributions. One such principle is offered by information theory. Recall how in our example of a very primitive source code the probabilities of the messages A and B could heavily influence the receiver's success in identifying them. This is a general (and fairly intuitive) feature of discriminative communication: Shannon (1948) proves that the more uniform the probability distribution of possible messages is, the bigger the entropy of the signal, i.e. the uncertainty of the receiver. And because the point of communication is to *reduce* the uncertainty about the incoming message, it follows that optimally structured source codes will contain highly skewed distributions – such as the ones found by Ramscar (2019).[5]

2.3 Eliminating alternatives

We will now look more closely at how communication evolves in real time, from the point of view of the receiver. The receiver needs to identify which message from the set of all possible ones is the intended one, and she can

[5]This of course does not make the information theoretic explanation of the distributions the only possible one. But an additional argument in its favour is that while there are many types of skewed distributions, the one that Ramscar keeps on finding in his corpus studies (i.e. geometric) appears to be optimal in terms of learnability – any other skewed distribution would be more difficult for language users to discover in their partial samples of language (Ramscar, 2019, pp. 30–31). Geometric distribution is therefore also crucial for the question of whether a language can be "shared enough" to work as a source code for discriminative communication – it can, because thanks to this distribution, people end up with sufficiently similar probability estimates.

do this because "[t]he conventionalized, *systematic* relations that hold probabilistically between all the linguistic signals as well as between the signals and the world enable listeners to incrementally reduce uncertainty about the messages speakers send in context" (Ramscar & Port, 2015, p. 92, emphasis in original). It might seem that the set of possible messages is hopelessly large[6], but we need to bear in mind how rich the situations in which people communicate usually are, i.e. how much information the receiver already has *before* the speaker even makes a sound. The context of when, where and why the people are meeting, what relationship is between them, what they discussed the last time they met, what their facial expression and body language are signalling, what social conventions are connected to the particular situation – these and other things narrow the set of possible messages right away. The process of eliminating alternatives simply does not start with the first word or sound.

Then, in case of spoken communication, the acoustic signal starts to unfold. As explained in the previous section, the listener does not receive a sequence of variously sized units with meanings encoded in them. What reaches her is a sequence of acoustic events that she uses to gradually reduce her uncertainty about what the speaker's intention is. A toy example (structurally identical to Ramscar's, 2019, p. 32) might again be helpful. Imagine a code that allows for four possible messages. Two of them are pronouns – 0011 is "me" and 0000 is "you", the other two denote activities – 1111 is "eating", 1100 is "sleeping". As the message unfolds, the listener has until some point received e.g. 00. At this point, although she does not know yet which of the four messages is being transmitted, she already reduced her uncertainty considerably – half of the possible messages are eliminated, only two are left. One might be tempted to say: "All right, but this just means that 00 is a codeword for 'pronoun' and 11 for 'activity' – it still seems that the listener might be composing the total meaning from smaller parts that encode partial

[6]Moreover, the set of messages people might wish to communicate does not appear to be closed. I will get back to this point in section 2.4.

meanings on their own." But this is not the case – when we look at the four codewords, we see that 00 is not only the "prefix" for pronouns, it is also the second half of the codeword for "sleeping", and 11 has analogously dual function. 00 is not a unit that points to a certain concept, it is just a means for creating a *contrast* – a cue that enables the listener to eliminate some alternatives. Which alternatives are eliminated is relative to the position of the contrast in the signal – at the beginning, 00 eliminates activities, and at the end, it eliminates "me".[7] This might also shed more light on what is meant by the "correlation" between acoustic and semantic contrasts that Ramscar and Port (2016) and Ramscar (2019) are talking about. We saw that although 00 does not *mean* "pronoun" in any traditional sense, it nevertheless can let the listener know that a pronoun is going to be communicated – the semantic contrast (pronouns versus activities) is correlated with the contrast of forms (00 versus 11). We can imagine a listener that does not know the code so well and thinks that there might be another codeword, say 0001, that denotes an activity, e.g. "running". From this listener's perspective, the initial 00 does not *mean* "pronoun" in any sense – she has to wait until the rest of the signal arrives to eliminate all alternatives. In fact, there might be a code like that, with five messages instead of four; and in that code, not only that 00 does not in any sense mean "pronoun", but also 11 cannot be said to mean "activity" – it could only mean something like "not a pronoun", which gets us back to the idea that the code simply serves to make contrasts.

Various types of contrasts (acoustic, semantic, visual, and others) all together lead the listener through the discriminative process and put constraints on each other. Maybe the preceding semantic information has been such that now it makes no sense to communicate an activity – in that case,

[7]This also explains how words (or other supposed units of language) "change meanings" with context. It is not that they have more meanings attached to them and the listener chooses the appropriate one according to context – the particular stretch of sound just eliminates different messages at different positions in the code (notably because in different contexts, there is usually not even the same set of messages to start with).

half of the possible messages from our toy language has been eliminated even before the first two digits arrived. And maybe the speaker pointed her finger to herself while sending the first two digits (00), by which the message "you" has been eliminated and the last two digits only confirm what the listeners already knows – that the intended message is "me". This also shows that such discriminative communication can be quite robust, because the co-occurring contrasts introduce some amount of redundancy into the signal, like when the speaker points her finger to herself and then "repeats" this information by the digits. In fact, our toy language is redundant by itself – the four messages could as well be encoded in two digits instead of four. But when the transmission channel is noisy (as is often quite literally the case in spoken communication), parts of the signal can be lost and redundancy can then save the situation. If for instance two digits are masked by noise, in case of the most efficient, non-redundant code, the message is lost, whereas in case of our code, the remaining digits and/or gestures might compensate for the loss. In spoken communication (and listening to speech in general), such contextual compensation clearly takes place. We know it from our own experience (e.g. from being able to have a conversation in noisy environment or via poor telephone connection) and also from laboratory experiments: Studies have shown e.g. that listeners can restore segments that are replaced by noise and are often not even aware of doing so (Warren & Obusek, 1971; Samuel, 1981) and that sentences in which every other interval is replaced by noise can still be quite intelligible, even when the intervals are as long as e.g. 333 ms (Powers & Wilcox, 1977).

2.4 Prediction and learning

There is one crucial feature of discriminative human communication that remains to be addressed explicitly: prediction. We already touched upon this topic at two points: First, we saw in section 2.2 that predictability of possible messages can have a profound effect on receiver's success (recall the

example with various probabilities of messages A and B). And in section 2.3 we saw that prediction based on redundant cues can help to successfully deal with a signal that was distorted during transmission.

The role of prediction in human communication, however, goes far beyond the above mentioned supportive one. Now the differences between human and artificial communication become relevant: namely that 1) people do not share the source code of their communication as immediately and completely as artificial agents do, and 2) the set of messages people might wish to communicate does not seem to be closed in any straightforward sense. It thus seems that substantial level of uncertainty will always be present in human communication. Human listeners therefore cannot be thought of as simply discriminating messages using the knowledge of the source code they already have at their disposal – they also need to be constantly learning. Notably, a great deal of human (and generally animal) learning also appears to be a discriminative and predictive process.[8] The listener is trying to predict what will come next from the currently available information, and she reinforces or suppresses connections between the cues she used and the predicted outcome according to whether the event occurred or not. In this way, she gradually learns to predict possible outcomes according to the cues that are most predictive of them (we can see the cues as features based on which the learner is now able to discriminate the possible outcomes).[9]

[8]Baayen (2011) shows that the outcome of an artificial discriminative learning model, based on Rescorla and Wagner's (1972) influential theory of learning, accord well with linguistic corpus data; and Baayen, Milin, Đurđević, Hendrix, and Marelli (2011) successfully use discriminative learning to model empirical findings concerning morphological processing.

[9]It should be noted here that this is clearly not the only type of learning that plays a role in language acquisition. Learning what cues are most predictive of certain outcomes presupposes that the learner already has an idea about what the possible outcomes are – for instance, what the vowel categories that she should discriminate by formant values are. A mechanism of category creation that I consider compatible with the view of language advocated in this thesis can be found e.g. in Boersma, Benders, and Seinhorst (2018) – in their model, categories are an emergent feature of a system that is fed by input

The process of learning through prediction is always on – it does not stop when the communicative code is learnt, because it never fully is.[10] In every communication situation, the listener learns about what contrasts the speaker used to encode her message, and the speaker sees whether the contrasts she used were sufficiently discriminative for the listener or not, and next time she can adjust her behaviour. Users of the code are therefore constantly learning from each other and all of them together are both learning the code and forming it. This is possible not only because of the very rich assemblage of contrasts allowing for redundancy, but also because human communication is *interactive* – the listener actively participates in the process by signalling understanding or the lack thereof, by explicitly asking for clarification, by finishing the speaker's utterances etc. This at the same time allows for immediate corrections and gives the speaker feedback about her communication strategy. Besides, language itself seems to be structured in a way that aids communication in which many forms that the listener hears may be new for her. Ramscar (2019, pp. 12–14) explains how the existence and distribution of both irregular (= highly discriminative) and regular (= highly predictive) forms in natural languages enable successful communication even though each user has inevitably only partial knowledge of the whole system.

---

with multimodally distributed acoustic properties. The topic of linguistic categories and phoneme emergence will be discussed in more detail in section 4.2.

[10] Ramscar and Port (2016, p. 70) put a strong emphasis on the inseparability of language and learning, seeing "language processing as a function of learning". The inevitable uncertainty present in human communication is in fact its point: "(...) because speakers and listeners embody learning systems, natural communication is 'meaningful' to the degree that listeners are not able to predict speakers perfectly. The goal of almost all speech acts is to effect a reduction in a listeners' uncertainty and to see that a listener learns something from them."

2.5 Takeaway from chapter 2

In this chapter, I presented human communication as a discriminative and predictive process, largely analogous to artificial communication dealt with by information theory. The transmitted signal is a sequence of contrasts (acoustic, visual, and semantic) that the listener uses to gradually eliminate possible messages and thereby reduce her uncertainty about the speakers' intention. The message is not obtained by combining meanings encoded in some discrete units, but by discriminating the intended message from other messages that the speaker might have sent.

Language then should be seen as a source code for discriminative communication. Its observable properties, e.g. skewed frequency distributions of items within semantic or functional groups and the interplay between regular and irregular patterns, encourage this view – language seems to be structured in a way that supports discriminative rather than compositional view of communication.

Unlike artificial agents, humans have only partial knowledge of the source code (because they acquire their knowledge from limited samples and because the source itself is evolving). We are therefore constantly learning during communication and the learning mechanism are predictive: The listener tries to predict what will come next in the signal and adjusts her model of the source code according to whether the prediction was correct or not.

Chapter 3

Integration: Discriminative communication within the skilled intentionality framework

Language processing is an integral part of cognition. My goal is to adopt the skilled intentionality framework in the field of cognitive science and the discriminative view of communication in the field of linguistics. Therefore I am bound to show that the two perspectives can work together. In this chapter, I will explain why I consider SIF and discriminative communication a happy marriage of ideas and how I envisage their integration.

In section 3.1, I present what I think are the demands SIF places on any theory of language (be it the discriminative approach or any other) if the two are to be compatible. Section 3.2 contains my idea of how to interpret and adjust Ramscar's discriminative approach to communication in order to meet those demands; and in section 3.3, I will explicitly turn back to them to assess whether this endeavour was successful.

3.1 General demands

Let us first examine what demands SIF places on any theory of language if the two are to be compatible. One demand follows from the mind-life continuity thesis (see section 1.1.2). SIF builds on the idea that **cognition resides in the basic interaction between a system and its environment** that makes the system a living one. The inevitable asymmetry of this interaction, caused by the system's need to maintain viable conditions, gets more pronounced as we proceed from simple living systems to more complex ones. In case of humans, this asymmetry is so conspicuous that we tend to think of our cognition in rather intellectual terms and disregard its continuity with the simple but powerful process that characterizes even the most simple organisms (such as cells). Although for many practical and research uses, this might be a justifiable and efficient stance, it should still be possible to track even our most sophisticated cognitive processes down to their fundamentals. This demand is worthwhile not only because it is generally desirable to strive for coherence and interconnectivity of the theories we build within different fields. It can also make us reassess some of our assumptions or views that seem unproblematic, but might in fact be preventing us from exploring new and potentially fruitful directions of research. A linguistic theory consistent with this view should therefore approach language processing as essentially the same process as the very basic forms of cognition, i.e. as an interaction between a dynamic system and its environment that is constantly heading towards the kind of mutual attunement that suits the living system.

This view might be difficult to reconcile with theories of language that rest upon stored representations of various linguistic units. Linguists often work with concepts like "inventory of phonemes" or "lexicon", invoking an image of boxes that people fill with items suitable for their language and use them to recognize corresponding portions of speech signal and to create their own utterances.[1] SIF does not go well with such accounts because, as

[1]This is common to theories which might otherwise be very distinct, e.g. for modular generative models (Chomsky & Halle, 1968) as well as exemplar theories (Bybee, 2010).

explained in sections 1.1.3 and 1.2.4, it considers knowledge a *skill* rather than a stored representation of a fact or a rule. SIF requires us to think about communication very much like about riding a bike, and not like about consulting inner dictionaries and grammar books.

Another demand SIF has on linguistic theories concerns the *purpose* of language use. Reasons why we communicate are enormously varied (chatting with my mother over a cup of coffee is a very different activity from asking my boss for a pay rise), but as long as we want to embed communication in SIF, all the different situations and goals should eventually boil down to the purpose of everything that a living system does. In chapter 1, I presented maintenance of the system's organization, i.e. survival, as the ultimate purpose of cognition. Of course, it does not seem very apt to interpret a chat with my mother in terms of survival. People have developed an amazingly rich and complex sociocultural layer with many practices that do not appear to be sheer survival skills. However, to be honest, thinking of many of them in these terms does not seem so far-fetched to me: Many people who experience existential problems today are in this situation not because they are not good enough in hunting animals or escaping from predators, but because they did not prove sufficiently skilful in the social practices needed e.g. to get a job. Things like self-presentation and social intelligence in my opinion *are* survival skills in today's world.[2] It is nevertheless clear that there are practices that cannot be reconciled with this perspective so straightforwardly

[2]I should note here that one of Bruineberg's closest collaborators on SIF, Erik Rietveld, would probably disapprove of the fact that I took survival as the starting point and now I am adjusting the perspective to deal with human sociocultural practices. According to Rietveld (June 2018, personal communication), one should take the opposite route: Theorizing about human cognition should start with the most sophisticated things we are capable of. This is probably why his work is concerned with, for instance, architectural practice (e.g. Rietveld & Brouwers, 2017) or clinicians' choices about deep brain stimulation treatment of OCD patients (van Westen, Rietveld, & Denys, 2019). I took the "bottom-up" route because I am deeply interested in the biological basis of cognition and I felt unable to present its relevance for communication clearly enough if proceeding the other way.

(and that some even seem to work, if anything, *against* survival). I think that this is why Bruineberg (2018, e.g. chapter 1) does not in fact usually speak about survival and works with a more general idea of "the tendency toward an optimal grip on a situation". This is what Bruineberg calls the basic concern of an organism rather than raw survival (which is naturally covered by it). This formulation in fact captures the goal that is more immediate for the organism than survival – i.e. the attunement between the organism and its environment. That is, the organism strives for such interactions with the environment that are the most convenient for it, which effectively ensures survival.

With that in mind, it may no longer seem so unrealistic to see communication and language as a part of an agent's effort to optimize her situation. This claim might seem trivial; however, seeing language as a means to optimize the agent-environment system has crucial consequences for what one considers to be a necessary part of language analyses. The role of *context* becomes pivotal. It is not enough to acknowledge context as a factor that can modify meaning of language units or can help to disambiguate unclear utterances. Context does not complement or modify communication, it provokes it, co-defines it and steers it all along – we speak and attend to someone else speaking because our skilfulness tells us that it is the way to resolve the divergence from the optimal model of ourselves in the particular situation. We do not do it because we need to transfer some content from one person's mind to another; we don't even always need to care about what the other person is saying (however sad this might sound)[3]. Theories of language that seek for ecological validity should be ready to deal with this deep situatedness of their subject.

After exploring whether there is in general a common ground for SIF and

[3]For illustrations of just how insensitive listeners can be to what their conversation partner is saying, see e.g. G. Roberts, Langstein, and Galantucci (2016) or Galantucci, Roberts, and Langstein (2018). In those studies, a substantial number of participants failed to detect clear inconsistencies artificially introduced in their conversations.

the discriminative approach to communication, we will come back to these demands to see whether (and how) the discriminative approach to communication can meet them (section 3.3).

3.2 Towards integration

3.2.1 Common ground

The skilled intentionality framework and the discriminative approach to communication share two salient features: the important role that low-entropy distributions play within them and their emphasis on prediction. Both of them were addressed in more detail in the preceding two chapters – what follows is only a brief reminder, serving to facilitate comparison.

Low-entropy distributions In both SIF and the discriminative approach to communication, we at some point come across information-theoretic entropy and the need to keep it low.

In case of SIF, we saw that the entropy of an organism's state distribution needs to be low in order to keep low the organism's average surprisal about its states (see section 1.2.1). This is ensured e.g. when the distribution peaks sharply around some values, i.e. when a relatively small set of states is highly probable and the others are highly improbable. Low entropy of the state distribution is not a property found by observation, it follows from the need of living systems to maintain their organization, i.e. to survive. To put is in somewhat odd terms: An organism needs its states to be highly predictable, i.e. absolutely "uninformative" – so much so that the organism actively influences its environment to make the situation meet its predictions (see 1.2.3).

In case of the discriminative approach to communication, we also saw the need for low-entropy distributions – this time distributions of possible messages, e.g. colour terms (see 2.2). Highly skewed distributions make the signal more predictable for the receiver, and also (namely geometric distri-

butions) make the source code of communication shareable among language users despite their limited experience with the whole system. In short, low-entropy language distributions help to reduce listeners' uncertainty about what speakers are saying.

The parallelism between the two accounts is obvious, no matter that in one case, we interpret entropy as "surprisal" and in the other as "uncertainty". In both cases, we work with the idea of a limited set of things (states or language contrasts) that needs to have certain statistical properties to enable the agents to work with it efficiently.

Prediction As already implied above, some kind of prediction (or anticipation, to get closer to SIF terminology) plays an important role both in cognition under SIF and in discriminative communication.

According to SIF, the agent embodies a model of herself flourishing within her environment, and this gives her existence a fundamentally anticipatory structure: She is not a model of what *is* the case but rather what *should be* the case if she is to survive (see 1.2.2). Any divergence from this model, i.e. any prediction error, drives action that serves to make the (thoroughly biased) predictions true. Effectively, the agent is thus the model of what *will be* the case (if her actions work as expected).

In discriminative communication, receivers' ever-present attempts to predict the following portion of the signal firstly help to keep communication seamless despite many inevitable distortions of the signal, and secondly they drive (again, via prediction error) constant process of learning that is an inseparable part of communication in natural languages.

Neither cognition nor communication are thus, according to the views adhered to in this thesis, "unidirectional" processes of simply receiving data from the outside and interpreting it afterwards.

It is this affinity what made me think that SIF and Ramscar's discriminative approach to communication might be a good match. However, a proper

integration of the two systems requires more explicit links between the two. Some of the obvious questions are:

- What does the distribution of states of an organism have to do with the distribution of language messages?
- How (if at all) do predictions serving to optimize the organism's situation relate to predictions concerning the upcoming speech signal?
- How does the bias that, according to SIF, characterizes all our cognition manifest in speech processing?

I will return to these questions explicitly in section 3.2.3, after I build the basis for answering them.

Perhaps even more important than the commonalities are the points where SIF and the discriminative approach to communication seem to clash. There seems to be a major obstacle in putting the two systems together, pertaining to the question of *what it is* that we communicate, i.e. what is the nature of the "messages" we share. The discriminative approach to communication, using the information-theoretic scheme, assumes messages that are being transported from one person to another. Although Ramscar's main point is that the structure of language is not shaped by the semantic content we communicate (i.e. the distribution of colour terms does not straightforwardly depend on coloured things), he does not deny that there *are* such contents we intend to share – after all, we do not select our messages randomly. This idea of exchange of content, being analogous to the view of cognition as gaining knowledge about the world, is exactly what SIF is very suspicious about, asking: What exactly is this content, how is it stored and where do you find something equivalent to it in the simplest living systems? We may try to answer the first part of the question by consulting Ramscar (2019) and Ramscar and Port (2016), to see what they say about what is actually being communicated. A precise definition of the content is, as far as I can see, not given, but there are certain formulations that might help. We repeatedly

read about speaker's *intention* or *wish* to communicate something (Ramscar & Port, 2016, pp. 2, 10, 13; Ramscar, 2019, pp. 32, 33). For this *something*, the words "message" and "meaning" are often used, but these have also other uses in the texts and are only substitutes for what we are searching for now. Sometimes the *something* is characterized as an *experience* (Ramscar, 2019, pp. 4, 32, 42) or a (semantic) state (Ramscar, 2019, pp. 32, 33); and there is also a concrete example where the content are *identities* or *experiences related to identities*, communicated by proper names (Ramscar, 2019, pp. 32, 33, 34). To me, this seems to imply that our experience with the world has a structure of discriminable experiential states, which are correlated with the messages that we are selecting in communication.

The idea that we communicate experiential states has some intuitive appeal, but it is not unproblematic. Firstly, it does not resolve the question of how these states should be stored and represented in our minds. It is certainly not the case that we actually directly experience the particular state whenever we want to communicate about it (otherwise we would probably not be so willing to tell all our friends about the last terrible visit at the dentist's), and thus we are facing the same troubles as with any other type of knowledge that needs to be stored in our minds. Secondly, even if we put the question of storage aside, there seem to be a problem with the degree to which experiential states are shared among people. Recall that in discriminative communication, the content is not packed in the code and transported from one person to another so that the listener might unpack it and discover an experiential state that she knew nothing about before; the content is discriminated by listener from other contents that she already had at her disposal (chapter 2). Sure, we belong to the same species and live in the same world, which means that we might have a lot of experience in common,[4] but there is obviously a lot of experiences we do not share, and quite

[4]In fact, there is no way of telling whether we indeed have the "same" experiences. We can see whether people react on particular stimuli in the same way by observing their behaviour, brain responses etc., but a direct comparison of *experiences* is strictly speaking not possible (if by *experiences* we mean the "first-person sensations" we have).

often these are the ones we communicate about. When a friend is telling me about herself giving birth, successful communication is possible even though I don't share this experience with her. So not only do we not share many of the experiential states we might wish to communicate about, but we also do not seem to *need* to share them in order to do that. This underscores the difference between human communication and Shannon's artificial communication scheme, where the goal is to "reconstruct the source message from the received message by discriminating the source message from other possible messages that might have been selected" (Ramscar, 2019).

In short, if the content of communication should be our experiential states, we have to resolve the question of their storage and representation, and we need to deal with the fact that in many cases, the experiential states are not shared. Both might be possible; I would, however, like to suggest a different route (also due to my reluctance to base a theory on something we cannot in principle access, such as humans' experiential states).

3.2.2 Communicating action

Speaker

Let me first take the perspective of a speaker (the sender of a message). In section 1.2.4, I introduced the concept of attractors, i.e. states into which a system tends to evolve. These states define the system's actions, *all* of them – and this holds also if the system happens to be a human being and the action happens to be a speech act (or generally an act of communication). Instead of the situation where the agent selects an experiential state (or any other type of semantic content) from a set that is somehow present in her mind, what we see is the agent being attracted to a certain action that involves producing a stream of acoustic and visual contrasts. To put it somewhat provocatively: the agent does not deliberately choose what she will talk about and then talk about it; the agent is steered by her environment and by the structure of her state space towards a state in which she produces

certain kind of noise (that we call "speech"). What sounds she produces is fully determined by the attractor she arrives at; and which attractor she arrives at is determined by how her previous experience has shaped her state space.[5] If she is skilled, her action will be adequate for the situation and the communication will successfully unfold. This is not to say that what the speaker is saying is not *intended*; I believe that Ramscar is correct to include intention (or "wish", as he sometimes calls it) in the communication scheme. We just have to bear in mind how intention is dealt with in SIF (section 1.2.2): That the agent "behaves intentionally" is just another expression for the claim that she includes herself in the model of the world she embodies. Being skilled does not just mean to be able to do what I want; perhaps even more importantly, it means that (thanks to the self-including model) I

[5] I am well aware of how counter-intuitive this may sound. We have the experience of choice – the reader may now be thinking of a tough oral exam or a conflict in a relationship, where the need for choosing *what* to say and *how* to say it is painfully present. One would be very glad to be simply attracted to an adequate (or in fact perhaps any) speech act in such situations. I admit straight away that I do not have an explanation for the experience of having and making a choice (just as I do not have it for self-awareness in general), and it can understandably be seen as a weak spot of my endeavour, perhaps even the weakest one. I dare to proceed despite this shortcoming because I am convinced that self-awareness and deliberate choice are not the determinative features of human behaviour. They certainly accompany (some of) it, but I do not believe that they are the basis – I see them rather as a superstructure that probably somehow supports certain types of highly complex behaviour that humans have developed. When I think of our everyday life, what I find truly fundamental (and fascinating) is quite the opposite of the highly conscious, thought-out choices we sometimes make. It is the impressively smooth sequence of thousands of activities which we do without really thinking about them, from making our morning tea to a casual chat with a friend. It may seem strange to include communication in that kind of activities, because we are used to thinking of language and its use in highly intellectual terms, but the degree of automatism that is present in the overwhelming majority of our daily talks in my opinion fully justifies it. The situations when we need to stop and make a thought-out decision are conspicuous *precisely* because they are rare and they do not align with how we do things most of the time; and although these situations and the mechanisms we use in them are definitely worth investigating, they are not the focus of my work.

want the right things – i.e. that in a particular situation, I am attracted to the appropriate reaction. I am not suggesting here that the agent does not in fact want to say anything or even that she is biologically forced to say something else than what she wants to say. Quite the opposite: the agent says *exactly* what she wants to say, because she simply always does precisely what attracts her as the best possible action in the situation.

To me, Ramscar's tendency to connect our communicative intentions with our experiential states in fact seems like a step in the direction I am pursuing now; I only take the idea further, led by SIF. I agree that what an agent communicates is determined by the her states, but the question is how. According to my view, we do not communicate our experiential states to others; rather, the act of communication *is* a state – but it is not directly bound to the experience that is the topic of the conversation. The connection between the state I experienced when I hurt my knee and my utterance concerning the event is not straightforward enough to say that I am *communicating the state* I was in back then. "But," the reader might object, "there surely *must* be some connection between the state I was in when I hurt my knee and my utterance about this event – otherwise it would hardly be an utterance *about the event*!" I partly agree. There is a connection between the state I was in back then and the utterance I am producing, just as there is a connection between *any* state I have ever been in and *any* utterance I produce. My experiences, i.e. the states I am going through, are shaping my state space and thereby influencing which states will be the attractors for me in the future. The experience of pain in my knee might very well be a crucial element in the chain of state transitions that got me to speak about how I hurt my knee. But not necessarily – I can very well speak about how I hurt my knee without actually ever hurting it, precisely because my utterance is *not determined* by the experiential state that seems to be its topic.

It might feel disturbing that the connection between an utterance and its topic is so loose and subsidiary according to this account – it might seem that a fundamental pillar of communication, i.e. that it *about something*, is

being pushed aside. The impression is partly correct, because I do consider this a side issue, and not a fundamental pillar. Even though the question of how an utterance happens to be *about* something might be an interesting one, I believe that we have to resist the temptation to think that it is the basic or central question of linguistics. What is "hello" or "go to hell" *about*? Language use is a very varied activity containing all sorts of speech acts. In acknowledging this variety and not taking the referential function of language as the primary one, I follow the line of thought famously pursued e.g. by Wittgenstein (2009), putting the *use* of language expressions, instead of their reference, to the centre of interest.

Listener

Let me now move to the perspective of a listener. Ramscar describes the process of discrimination during which the listener progressively eliminates possible messages to get to the one that was intended by the speaker (how it is with the speaker's intentions was clarified above). Once again, the question of storage and representation arises: Where should the set of possible messages be in the listener's mind and what actually *are* these messages? My proposal is analogous to the one concerning the speaker, only here it is perhaps more intuitive. SIF itself portrays the agent as moving through a set of possibilities – her state space – and arriving at one of them – the attractor. We can therefore try to substitute the set of the listener's possible states for the set of possible messages. The agent's movement through her state space can be seen as a process of elimination as well: At the hypothetical time zero, the agent might move anywhere in the state space, all possibilities are open, and the impulses from the environment, interacting with the structure of the state space, gradually eliminate the possibilities by navigating the agent in a certain direction until she reaches an attractor. The impulses can be of any nature, including an acoustic signal produced by the speaker. As the sequence of acoustic contrasts unfolds, they provoke changes in the listener's state – what changes exactly, depends on her previous experience that shaped

her state space in a particular way.

The process of communication thus starts with an attractor of the speaker and ends with an attractor of the listener. It might seem that attractors took the place of the content that is being transferred from one agent to the other and we might be tempted to claim attractors "the new meanings". These are, however, rather unfortunate formulations for a simple reason: the attractor that determined the speaker's utterance and the attractor that the listener arrived at *are not the same*. They cannot be – if the listener arrived at the same attractor that the speaker started with, she would repeat exactly the same utterance, because the attractor brings about this particular action. As we saw in section 1.2.3, there is nothing like a purely "perceptual state" (because any state determines an action), so it can neither be the case that the listener would somehow only "perceive" the attractor that caused the utterance but would not act upon it. The states of the speaker and the listener are simply completely different, even though they are somehow connected to the same acoustic signal.

There is nothing disturbing about this picture once we abandon the idea that communication serves to get something from the speaker's mind into the listener's mind. We can compare communication to a ball game: the action of the person that catches the ball has to be very different from the action of the person that throws it if the interaction should work. The game is not about the *ball* that is travelling from one person to another, any object could take its place – the ball is just a material tool that participates in a process of provoking desired actions of the players, sometimes in a cooperative, sometimes in a competitive setting. Analogously, what is being "communicated" in a conversation is not a content, but an action, be it a verbal response or any other type of behaviour.

3.2.3 From affinities to explicit links

I would like to return to the questions I promised to address in section 3.2.1, just after I presented the apparent affinities between SIF and the discrimina-

tive approach to communication. Answers to all of them should be at hand by now.

What does the distribution of states of an organism have to do with the distribution of language messages? My proposal is that these two distributions are the same thing. I suggested to substitute the agents' attractors, i.e. states they are inclining to under particular circumstances, for the messages "selected" by the speaker and arrived at by the listener. The fact that both the state space and the structure of communicative code are characterized by low-entropy distributions then seems to be expectable, and not an accidental affinity. However, we must bear in mind that while distributions of language elements were discovered empirically (Ramscar, 2019; Linke & Ramscar, 2020), the claim that the state distribution of living organisms has low entropy is Friston's logical deduction from the premise that the systems need to stay alive within changeable environment (see section 1.2.1). I am in no way claiming that I have *proven* the identity of the two distributions, neither in the logical nor in the empirical sense. What I hope to have shown is that although the two accounts arrive at low-entropy distributions each in its own way and for its own reasons, they are conceptually compatible and the reasons might in fact not be independent.

How (if at all) do the predictions serving to optimize the organism's situation relate to the predictions concerning the upcoming speech signal? This is a tricky bit, because we seem to be dealing with two different kinds of prediction. When it comes to the states of the organism, we saw in section 1.2.2 that the predictions concerning them are in fact also *plans* – the agent predicts what she *needs* to be the case and act so that the predictions come true. But predicting of the upcoming speech signal seem to be a different task – given that the listener cannot influence what the speaker will utter in the following (milli)seconds, it seems pointless to "plan" on what the speaker should say. We seem to have found a domain where cognition is indeed just

gaining information from the environment (the speaker in this case) and the agent tries to simply accurately predict the "objective reality", independent of herself – just as the classic version of predictive coding theory sees it (section 1.2.2).

Two things, however, shed a different light on this issue. 1) It is not really true that the listener cannot influence what the speaker will say next. Communication is a highly interactive process and we adjust our utterances to our partner's reactions in very short time windows, too. Sometimes a mere expression on my face can make the speaker stop speaking immediately (and I do not consider this my superpower). Gestures, backchannels, interruptions – we do have means to adjust the speaker's production to our needs and we are making use of them all the time, although usually subconsciously. 2) It is worthwhile to ask *how we actually know* that listeners are predicting the upcoming speech signals. The evidence falls into two groups: First, listeners are often able to restore missing or distorted portions of signal on the basis of preceding context. The second type of evidence is based on speed. In conversations, turns usually follow one another very tightly or even overlap, which seems to require quite precise predicting of the end of the partner's utterance (Levinson, 2016); and in monitoring experiments, context is known to influence reaction times, making them shorter for predictable targets.

The bottom line of these proving techniques is that they monitor our reactions. As obvious and trivial as it may sound, it means that all we know from the experiments is that people *act as if* they were predicting what the speaker will say. They say nothing about the *mechanism*. What I suggest is that the reactions are merely reflecting the state into which the signal, together with my skilfulness, got me at a certain point in time. When I hear "hello, how are –", it gets me to a state from where I will readily move to saying "fine" if a friend whom I just met said it; or I will readily move to saying "you" if I am in a lab, asked to finish the sentence; or I will readily move to pressing a button if my task is to monitor the word *you*. I might perform all of these acts even if the word "you" actually never comes (in the

last case, that would a mistake – the kind of mistake we can easily imagine to happen in monitoring experiments). Does it mean that I predicted it? Well, in many contexts I will not hesitate to say "yes", because this is simply how we use the term *prediction*; but let me be more precise here. What happened is that I got to a state which we are used to connect with the final "you", and I got there earlier than a person who is less skilled in English perhaps would. Contextual cues were sufficient for me to get to this state, whereas e.g. a beginning learner might need also the final "you" to get there. To stress again: I have nothing against calling this phenomenon "prediction" or "anticipation" and I will continue doing so, but for my endeavour it is crucial to acknowledge that it makes no sense to speak about prediction unless it is manifested in action.

"Prediction of the agent's future state" and "prediction of the upcoming speech signal" are therefore simply different expressions of the fact that a skilled agent has certain attractors in her state space. That the agent predicts her optimal state means that she is, thanks to her skilfulness, attracted to the actions that will help the optimization; and that the agent predicts the upcoming speech signal means that she is, thank to her skilfulness, attracted to an appropriate reaction to an utterance even before it is finished. We are not dealing with two different types of prediction here, where the second one would be more "objective" than the first one; we are dealing with two equivalent manifestations of the agent's skilfulness.

It was further stressed in section 2.4 that in communication, prediction error drives the never-ending process of learning. But what is a "prediction error" here? I just refused the idea that prediction means making a guess of what signal is going to come and then comparing this guess with reality. The only error I can make resides in taking an action that does not lead to the expected optimization of my situation, i.e. taking an action after which the disattunement between me and my environment will not be reduced. Based on this experience my state space will be rearranged so that the next time, my reaction to an equivalent situation will be different. Learning is noth-

ing else than rearranging of the state space in a way that makes my actions more efficient with respect to my needs, i.e. developing a skill. This aligns well with the fact that language learning (especially at the sound level) very often leads us to *ignore* some features of the signal, in fact many of them – I will get back to this in more detail in section 4.2.2. At this point, I just wanted to point out that the "prediction error" that drives language learning is ultimately not a mismatch between my guess of the following signal and the signal itself, but rather a mismatch between the effect that my action had on my situation, and the effect I expected it to have.[6]

How does the bias that, according to SIF, characterizes all our cognition manifest in speech processing? This question is tightly connected to the previous one. The bias (i.e. that we predict what is convenient for us) determines what the attractors in our state space are, so that we incline to actions that will adjust the situation to our needs. Speech processing is just one of the things that navigate us through our state space – but where exactly we end up, depends as much on the speech signal as on how our state space is organized. Speech signal will steer me towards an attractor based on my previous history – if I developed a skill of prompt witty comments, my friend's utterance will move me towards saying one; if not, it will perhaps move me towards a silent nod. The utterance was the same and the processing mechanism also does not change, it is still elimination of action possibilities – but since the bias forces us to set our attractors in the most convenient way and our personal histories that determine this setting are different, our reactions may differ as well. In addition, the bias determines what we will learn through communication, as explained in the previous paragraph.

I hope that by addressing these questions explicitly, I have shown that what

[6]To remind again: Talking about the listener as "predicting the upcoming signal" can be a convenient and completely justified shortcut in many contexts, including linguistics. Here I am offering a hypothesis of the underlying mechanism; its practical relevance will be explored in chapter 4.

might have looked like accidental affinities is in fact a solid basis for explicit links between the two systems I am trying to integrate.

3.3 Meeting the demands

Let us now get back to the demands placed on linguistic theories by SIF, presented in section 3.1. The first one followed from the mind-life continuity thesis and required us to interpret language in terms of the agent's constant effort to optimize her situation. In the previous section, we made a big step in this direction by reframing communication as indirect mutual influencing of the participants' actions, but it remains state explicitly where and what exactly *language* is in this picture and what it means to "know" it. In what follows, I draw upon the general outline of how SIF could be applied to the so-called higher cognition described by Bruineberg (2018, chapter 4).

Ramscar (2019) sees natural languages as to some extent analogous to source codes for discriminative communication, i.e. as systems that "allow messages (that relate to discriminable human experiences) to be mapped onto discriminable speech contrasts" (p. 4, footnote). I diverted from this view when I substituted attractors for the messages related to human experience, but I agree with the core idea: language appears to me as a system that allows us to use acoustic and visual contrasts to deal with our social environment in a skilful way. There has to be some system in how the contrasts that we produce affect our environment (mainly other people, but not exclusively) – otherwise these contrasts would be useless. This can be illustrated by me asking my husband to do the dishes in a language that he does not know – I produced some acoustic contrasts, all of them perhaps present also in our native language, but I used them in a way that does not fit in the system that my husband knows and they will therefore not affect his actions in the way I might have wished.

What any system is based on and what makes it possible for an organism to become skilful is *regularity*. If objects in our environment or our own

bodies would not conform to many regularities (e.g. the laws of physics) and would behave at random, there would be no chance for us to develop the skills we depend on in our everyday lives – it would be impossible to survive in such a world. The same holds for the effects that the sounds we produce have on other people. Sociolinguistic regularities are apparently not as stable as the laws of physics, but it does not matter. A living system does not need to discover unchanging laws and use them ever since, because its knowledge has the form of a skill – the regularities just have to be reliable enough to allow for developing a skill that will work efficiently. In currently spoken languages, the regularities are even constantly changing (word meanings shift, new grammatical structures occur while some of the old ones go out of use etc.), but there is enough stability for communication skills to be gained and, if needed, adapted. Dictionaries and grammar books do nothing else than try to capture these regularities, although it is often not obvious from the way they are written. The experience of learning a foreign language can give us a sense of this. We all probably started with learning lists of words with translations; and we all probably came to a point where we found out that this is not the point. To learn a new word does not mean to learn its translation, not even its definition in the target language (although that is already much better) – it means to discover how it is *used* in the language, in which contexts (both linguistic and situational) it occurs and in which not. To some extent, one can learn this from the definitions, but that still is not enough – until the learner is able to use the word properly (where "properly" means that the whole utterance will have the desired effect in communication), we can be rightfully reluctant to say that she has learnt it. The process of learning is not a matter of remembering something *about* the word, it resides in gradual mastering of its usage – it is a process of skill development. In terms of SIF, learning a new language corresponds to rearranging of our state space, rather than adding something new. We were always able to hear the sounds of the target language – what we are learning is to be attracted to the appropriate actions by the acoustic (or

visual) elements, and to use them to elicit desirable actions in others.

From what I just said it must be obvious that I am not fond of definitions (to put it mildly). But if I were forced to offer a definition of language, based on the previous paragraphs, it would go along these lines:

Language is an assemblage of sociomaterial regularities that enable us to skilfully engage with our environment, primarily other humans.

Important additions would be these:

1. It is not an unchanging assemblage and it is constantly co-created by people as they are communicating.

2. The "material" in "sociomaterial" refers to anything people use (or might ever use) to communicate: acoustic signal, gestures and signs, written text, etc.

3. I felt obliged to add "primarily" because it is increasingly common to use language to control electronic devices (not to mention verbal communication with pets).

This is a view of language that, as far as I can see, satisfies the first demand of SIF. Knowledge of language is a skill just as any other (dancing, carpentry or cellular chemotaxis) and does not require storage of discrete units mysteriously connected to their meanings.

The second demand of SIF on language theories was that they have to respect the fundamental purpose of cognition, which is the attunement between the organism and its environment (ultimately ensuring survival). I said in section 3.1 that this manifests mainly in how the theory deals with context. We took a major step towards meeting this demand in the preceding paragraphs, where I presented communication as mutual influencing of human actions and language as an assemblage of regularities thanks to which the influence can be systematic and predictable. Regularities are by definition

contextual – they arise when a phenomenon co-occurs repeatedly with some other phenomena. In other words, there is no regularity without context.

Another step was taken already in section 2.3 where we explored the role of both linguistic and situational context in comprehension. Firstly: In our toy language with four messages consisting of zeros and ones, interpretation of the digits was heavily dependent on where they occurred in the message and what preceded/followed them, i.e. on "linguistic" context. And secondly: I noted that situational context plays exactly the same role as the linguistic one in that it narrows down the set of possibilities that the listener has to eliminate and thus determines the interpretation of linguistic contrasts. This, however, is not by itself enough – one could acknowledge this great contribution of context and still see it rather as a supplement, a source of additional cues. Recall that in SIF, the connection between an organism's acts and the environment (i.e. context) is much tighter, because cognition is a process of optimizing the brain-body-environment system as a whole (section 1.2.3). It is a disattunement in this system that elicits the organism's action, which means that the environment is just as "responsible" for the action as the organism itself. The view of communication as mutual influencing of the partners' actions fits well with this perspective. When I am speaking, I am shaping the environment of my partner to create a disattunement in her brain-body-environment system, which makes her react in a way that will, hopefully, help me resolve the disattunement in *my* brain-body-environment system. that made me speak. In fact, I am a part of my partner's environment and she is a part of mine – communication is therefore a process that leads to a certain kind of equilibrium in a system that encompasses us both. Context is then the *cause* of an utterance in a very strong sense – the discrepancy between the situation and the agent's model of the optimum elicits the particular stream of sound, and the stream of sound is uttered only because the agent's skill tells her that it will lead to the desired change of the situation.

The discriminative view of language, with the adjustments made in sec-

tion 3.2.2, therefore in my opinion meets also the second demand of SIF. It offers a straightforward connection between language and the optimizing tendency of an organism: It considers language an assemblage of regularities that allow us to develop skills that serve us to maintain the right attunement with our environment in a particular domain – the social one.

3.4 Takeaway from chapter 3

In this chapter, I tried to integrate the skilled intentionality framework and Ramscar's discriminative approach to communication. To remind quickly the two conceptions: 1) SIF considers cognition a specific type of interaction between an agent and her environment – one that maintains the brain-body-environment system sustainable. Disattunement between the agent's model of the optimum and her actual situation will navigate a skilful agent towards the state that brings about a remedying action, because these states are the attractors in her state space. 2) Ramscar's discriminative approach to communication portrays speech as a stream of contrasts that enable the listener to gradually narrow down a set of possible messages until she ends up with the final one, ideally the one intended by the speaker. Since the set of possible messages is not closed and the system of contrast used for signalling is changeable, an inherent component of communication is learning.

The decisive step in my attempt at integration was centering communication around the agents' attractors instead of a content that should be transmitted. The idea of content that is shared when communication is successful (be it messages, experiential states or anything else) was abandoned completely. Speech acts are elicited by disattunement between the speaker's optimal model and her actual situation, and serve to reduce it by causing disattunement on the side of the listener, which in turn makes him act (verbally or in any other way) in order to optimize his situation. Communication is thus a way to skilfully affect an environment where other acting subjects are present. The substitution of attractors for content then enables us to

fully benefit from the affinities between SIF and the discriminative approach to communication, i.e. the role of low-entropy distributions and prediction in both of them.

Under this view, language is as an assemblage of sociomaterial regularities that make this skilful engaging with other people possible. Although the system is very complex and invites descriptions in abstract terms (typically found in grammar books), its essence are regularities in how people react on certain type of audiovisual stimuli, no matter how involved both the stimulus and the response might be.

Chapter 4

Phonetic topics from the integrated perspective

I am grateful to the phonetically oriented reader for bearing with me through the first three chapters, where mentions of phonetics and phonology were rather rare. In the following chapter, I will finally try to apply the above developed view of communication and language (which I will call the "integrated perspective" for brevity) to concrete phenomena studied within the field of phonetics and phonology. Since this field is concerned with the sound level of languages, I will start with a general remark on how linguistic levels are treated within the integrated perspective (section 4.1). Section 4.2 will be dedicated to the notorious phenomenon of categoricality of speech perception and production, and in section 4.3 I will explore speech in its most natural environment, i.e. spoken conversation, focusing on turn taking.

4.1 Linguistic levels

So far, I was quite agnostic as to which level of linguistic structure[1] my suggestions should be applied to – i.e. whether we should associate attractors

[1]Quite common term is "levels of representation". I avoid it here because, as should be clear from the preceding chapters, I do not see linguistic knowledge as representational.

with phrases, or words, or phonemes etc. This was intentional, because they should be applicable to all of them. I believe that linguistic levels are a (useful) construct that helps us describe languages in a neat way; however, language users are not limited or bound by them. What appears to be a "word" when linguists analyse an utterance (or when we simply write it down) might work in many different ways for language users. Let us take the sound [pʰɔːɬ] (corresponding to the name "Paul") as an example and consider three different situations where it could be used. 1) Someone says: "I am Paul." Here the sound operates on what we would probably call the *word level* – it contrasts with other words like "John", "tired" or "here". 2) Someone knocks, I open the door, see a friend and exclaim: "Paul!" Here the sound works at the *sentence* or *phrase level* – it contrasts with reactions like "What a nice surprise!" or "What are you doing here?" 3) "Who do you like more – Paul Newman or Paul Walker?" "Paul Newman." Here the sound [pʰɔːɬ] in the response is not informative at all – from the semantic perspective, it is entirely superfluous. It behaves rather as a sublexical unit – very much like e.g. the first syllable of the word "because" that can often be heavily reduced or left out without the message being harmed by that. According to the integrated perspective, language users process the sound [pʰɔːɬ] differently in those situations and do not make use of the same abstract unit – the word "Paul" – in all of them. We can recall our toy language from section 2.3, consisting of four digits. It was shown that s stretch of signal (e.g. "11") cannot be considered a unit with a stable "meaning" – it serves completely different purposes in different contexts, because in each case it eliminates a different set of possible messages. The exact same logic holds for [pʰɔːɬ] or any other stretch of speech, and the variation of functions a stretch of speech can serve is not limited to one linguistic level.

This is not to say that traditional linguistic levels have no support in language use, they certainly do.[2] However, "corresponding to a word" or "cor-

[2]Linke and Ramscar (2020) argue that they are crucial for successful language learning and use, which is the crucial difference between artificial and human communication.

responding to a phoneme" are not stable characteristics of certain stretches of speech, and these concepts might mislead our ideas about speech processing if we treat them as such. It is also not to say that speech processing does not have hierarchical structure. Hierarchical organization of the organism's dynamics is explicitly assumed by the skilled intentionality framework (Bruineberg, 2018, e.g. pp. 18–19, 59). The organism's state evolves simultaneously at various time-scales. The slower dynamics defines higher-level states that serve as parameters for the faster dynamics; but at the same time, perturbations coming from the faster level determine how the slower level evolves.[3] This interaction between levels is necessarily taking place also during spoken communication, and it might even be possible to associate some of the levels with production/perception of phrases and some other with production/perception of speech sounds in terms of phonemes. But this association cannot be absolute with respect to the portions of the speech signal – instead, I believe it should be functional, as illustrated by the above example with the sound [pʰɔːɬ].

Crucially, we could not get around this by constructing a detailed "contextual grammar" where we would describe in what linguistic contexts [pʰɔːɬ] plays the role of a word, where the role of a phrase etc. First, such grammar would have to include all sorts of situational factors – purely linguistic tools would not be enough to describe even the three simple examples above. And second, even more importantly: Different people might process the same sound differently even in the same situational and linguistic context, due to

[3]To give a simple example: When I am standing, I have a certain range of possibilities how to move my body – I can walk, swing my arms, jump etc. When I lay down, I change the higher-level state of my body and the set of lower-level possibilities that are open to me changes drastically – jumping or walking is no more possible, but I can e.g. hold both legs at a right angle to my body, which was not possible in the previous state. Also impulses from the outside will result in different reactions – if something pushes my leg to the side, my body will respond very differently in the two positions. And at the same time, it is the lower-level movements (such as bending knees or putting arms in front of me) that makes me either stand up or lay down – I can only change the higher-level state through changes at the lower levels.

the differences in their experience. We might illustrate this on second language acquisition: While for a native speaker, the sound [kʰuːɬ] often works at the phrasal level (contrasting e.g. with "I don't like that"), a beginning learner might be unable to process it this way and will end up with a word-level interpretation (being aware only of the contrast with e.g. "warm"). Although in such cases the differences in processing are most prominent, I am convinced that they apply to speakers of the same language too. The experience people have with their native language is very diverse and the integrated perspective predicts that the processing strategies will reflect this diversity.[4]

Assigning a stretch of speech to a certain linguistic level therefore requires careful consideration of the linguistic and situational context and of the agent's experience, which makes it a highly non-trivial (if not outright unfeasible) task.[5] The integrated perspective does not offer a solution to this problem, on the contrary – it allows us to leave it unsolved and requires us to operate with linguistic levels very flexibly (if at all). In what follows, I will make use of the standard concepts of words, phonemes, etc. since they are descriptively convenient; however, I ask the reader to bear in mind their

[4]One might get a sense of it when entering an unfamiliar community – for me, it could be e.g. a senior management meeting. Lost in the corporate newspeak, I would be able to process most of their "COO"s and "BAU"s (standing for "chief operating officer" and "business as usual") at best at the level of syllables and write the abbreviations down to look them up later, whereas the managers swiftly use these stretches of sounds at the word or phrasal level and are navigated by them to attractors that are not available to me.

[5]It is also difficult to free oneself from our writing system, and sometimes also from well-established linguistic tools of analysis. We tend to consider "how are you" a phrase consisting of three words because we write it that way, although when it comes to real-time processing, those three syllables might be as cohesive as e.g. "computer". On the other hand, when we encounter a single word functioning at the phrase level, e.g. when someone says "I cannot go with you." and I reply "Pity!", linguists analyse it as an ellipsis, as if the speaker at some point worked with the phrase "It is a pity!" and left a part of it out, assuming that the listener can reconstruct the full phrase. From the integrated perspective, the utterance "Pity!" simply leads the listener toward the same attractor as the full sentence would, without the need of reconstruction.

mere auxiliary and flexible status.

4.2 Categoricality of speech

In the previous section, I looked from the integrated perspective at what could be called the "vertical" structure of language. This section will be dedicated to a topic concerning the "horizontal" structure; specifically, the topic of categories at the sound level of language.

In phonetic and phonological research, the word "category" can be understood in at least three different ways. I will first present the distinction and make clear what type of categories the integrated perspective needs to deal with.

- First, "category" might refer to a phoneme. That is an abstract functional unit, defined with respect to the system of language (Trubetzkoy, 1969). Phonemes of a language are defined as sound in mutual contrastive distribution, which means that replacing one phoneme with another changes the meaning of a word. Phonemic inventories of languages are often (although not always) obtained through searching for minimal pairs, i.e. pairs of meaningful words that differ only in one segment.

- Second, "category" might refer to a cluster in a (psycho)acoustic space that emerges when we record a sufficient amount of production data (e.g. when we plot vowels produced by Czech speakers in a F1–F2 graph[6]).

- Third, "category" can be a perception-based notion: In identification and/or discrimination experiments, clusters of sounds might emerge that get the same label from listeners and/or are difficult to distinguish (compared to other, psychoacoustically equidistant sounds).

[6]I.e. a graph where the values of the first vowel formant are on one axis and the values of the second formant on the other, which represents human vocalic space.

The three meanings of "category" are naturally not completely independent; however, their relationships and also their role in online speech processing are not self-evident.

For decades, phonetic research has been interested in the relationship between phonemes (i.e. categories in the first sense) and either perception-based or production-based categories (Casserly & Pisoni, 2010). In the perception-oriented research, a recurrent notion is that of "mapping" – the key question has been how listeners map from segments of speech signal to phonemes. Years of research made it clear that this task is anything but trivial due to the so-called *lack of invariance problem*. There does not seem to be a stable set of psychoacoustic cues that would enable us to unequivocally determine the mapping, irrespective of context, speaker, task etc. (Holt & Lotto, 2010). And yet, the lack of invariance seems to be a "problem" rather for linguists and speech technologists, while the majority of language users communicate effortlessly in their native language despite the high variability of speech signal (stemming from multiple sources).

From the integrated perspective, the question of mapping between phonemes and the speech signal is subsidiary, because it does not directly concern speech processing. There are no abstract representations such as phonemes in language users' minds to map the signal to, and speech processing therefore cannot reside in such mapping. This is because the integrated perspective denies representations stored in language users' minds in general, and sees cognition (including speech perception) as a continuous process of mutual attuning between the agent and its environment (see e.g. sections 1.1.3, 1.2.3, 3.1). In their paper "Categorization (without categories)", Ramscar and Port (2015) deal with categories associated with lexical units (such as "dog"). They argue that "category" in linguistics would be appropriately used "to describe a set of items with a common label" (p. 78). To put it in more casual terms, for example "*games* are whatever speakers of English call *games*" (Ramscar & Port, 2015, p. 77, italics in original). In the cited paper, the authors show why it is problematic to consider lexical categories

something more than that, and that the discriminative approach to language avoids it. I am convinced that the situation of categories based on phonetic features, i.e. phonemes, is analogical: the phoneme /ɛ/ is whatever language users treat as /ɛ/, be it in a lab or in everyday communication. This is because phonemes are defined functionally and are abstract – they do not themselves have any inherent acoustic (or articulatory) properties. For descriptive purposes, it might be useful to search for reasonably reliable connections between speech signal and phonemes as descriptive constructs, but when it comes to understanding human everyday communication and online speech processing within the integrated perspective, categories in the first sense are irrelevant and the lack of invariance problem can remain unsolved.

However, that does not mean that the integrated perspective does not need to account for categoricality of speech. There clearly *are* categories of some sort in the speech signal (as is clear from production data) and speech perception clearly *can* and often *does* evince categoricality. I will now show how this property of speech can be treated within the integrated perspective.

4.2.1 Categorical behaviour

Categoricality in general is an inevitable consequence of how dynamic systems with non-uniform state spaces work. Attractors, as the states that a system tends to evolve into whenever it is nearby in its state space, naturally bring about categoricality of the system's behaviour – it is *attracted* to the same reaction in a certain range of different situations. What situations fall within the range and what reaction serves as the attractor is given by the experience of the agent. The existence of regularities in her environment helps to determine the range – for instance, the regular concurrence of sharp objects with pain teaches the agents that in situations when a sharp object is approaching her hand, she withdraws it. That in a sense creates a *category* from these situations. In the same fashion, speakers of Czech learn to respond e.g. to written symbols like "e" or "E" with the sound [ɛ]. The range of situations that make Czech speakers produce the sound [ɛ] is larger than

the range of situations that make them produce e.g. [æ], and therefore we will find a cluster of various [ɛ]-sounds in Czech production data, and not a cluster of [æ]-sounds. This categorical behaviour in turn creates and maintains the sociomaterial regularities that constitute our language and can be exploited by listeners. By regularly encountering [ɛ]-sounds and [a]-sounds in different situations, language users can use these clustering patterns to determine two ranges of situations that call for different responses, for example pressing a different button during an identification task. The two attractors, i.e. pressing the button A or pressing the button B, can then be used by linguists to establish two categories. Categoricality of speech (as a sociomaterial phenomenon) therefore maintains itself through language use. From the integrated perspective, then, psychoacoustic cues do not serve language users to *identify* a certain linguistic unit. They only bring her to the vicinity of an attractor towards which she then inclines. In other words, the cues (just as any other input from the environment) participate on the process of elimination of action possibilities.

It might be tempting to think about attractors as representing phonetic categories. Listeners' state space would seem like a perceptual map warped by experience with the native language, where valleys correspond to individual phonemes. The integrated perspective would then appear equivalent to the models of speech processing based on such "warping", such as the one proposed by Guenther and Gjaja (1996) to deal with what is called the perceptual magnet effect[7] described by Kuhl (1991). In their model, uneven distribution of formant values that are fed to a neural map leads to uneven firing preferences of the neural cells. If we for example train the model with many instances from the [ɛ] production cluster, more cells will respond to the most common values than to the uncommon ones. And because the final percept is a result of averaging over all the cells' activations, the perceptual

[7]This effect concerns internal structure of phonetic categories, which, according to Kuhl's (1991) results, is prototypical. Listeners perceive some category members as better representatives of the category than others, and sounds that are psychoacoustically closer to the prototype are less reliably discriminated than sounds on the periphery.

map becomes skewed. If we then feed a relatively uncommon [ɛ] to the network, the activation pattern will yield a percept that is closer to the most common [ɛ] than the input actually was. The resulting percept is therefore determined by both the input and the organization of the map, shaped by training; and the percept corresponding to the most typical sound from the set thus in a sense acts as an attractor.

The result of this warping is that certain sounds are grouped together and drawn apart from others – in other words, categorization. Holt and Lotto (2010) argue that speech perception is an instance of categorization as a more general cognitive mechanism. They contrast their suggestion with the view that categoricality of speech perception is language- or human-specific phenomenon based on identification of phonemes. The integrated perspective shares with Holt and Lotto (2010) the intention to explain speech perception from general cognitive mechanisms, and also the conviction that listeners do not use psychoacoustic cues to identify phonemes. It might therefore seem that Holt and Lotto's (2010) shift from seeing speech perception as based on identification to seeing it as categorization, combined with a "warping" model such as the one proposed by Guenther and Gjaja (1996), captures the idea of speech processing offered by the integrated perspective. Under the simplifying assumption mentioned above, i.e. that we interpret attractors as representing phonetic categories, the integrated perspective and Holt and Lotto's (2010) account seem to be practically interchangeable.

However, as explained in sections 1.2.3 and 3.2.2, attractors correspond to *actions*, not to stored abstract representations of any kind. From the integrated perspective, segments of speech and their psychoacoustic properties neither serve to *identify* phonemes (or other units), nor are they *categorized* – they only serve as cues that eliminate action possibilities of the agent. What the initial set of action possibilities is, depends heavily on context. The same psychoacoustic cues that lead to pressing the "e" button during an experiment do entirely different job when they occur in a different situation, for instance when a friend says "hello" to me. The [ɛ] in the laboratory experiment

and the one in "hello" would probably fall within the same production-based category (i.e. the same cluster in the F1–F2 graph) and would perhaps also end up in the same perception-based category (determined by identification and/or discrimination task), but from the *processing* point of view, they have nothing to do with each other. These two sounds, just as the digits in the toy language in section 2.3, share some characteristics, but they have completely different jobs in the process of eliminating action possibilities.

Holt and Lotto (2010, p. 6) acknowledge that "phonetic categorization is extremely context sensitive", i.e. that in different (linguistic and/or non-linguistic) context, the same sound might be categorized differently. My claim is even stronger – the role that a particular segment plays in speech processing is fully context *dependent*, because context determines the set of action possibilities that the segment helps to eliminate. Context dependency follows from the hierarchical structure of dynamic systems described in section 4.1. Contextual (and other) factors serve as higher-level parameters that shape the listener's state space so that a particular sound is treated differently than we would expect on the basis of laboratory experiments. If we once again adopt the simplified view according to which attractors correspond to phonetic categories, we can describe the situation as follows: In laboratory setting, an [e] sound might navigate the listener to the attractor corresponding to /e/. In another context, however, the valley in the state space that corresponds to /e/ might be outstripped by a much larger valley corresponding to e.g. /i/, as illustrated in figure 4.1. As a result, the same [e] sound will be categorized differently. It is not that the /e/ attractor is not there anymore, but it is irrelevant in the particular situation. The dynamics that gave rise to this /i/-favouring setting might correspond e.g. to a lexical bias (Ganong, 1980) or a visual interference of the kind described by MacDonald and McGurk (1978).[8]

[8]This is of course not the only way how to account for contextual effects. Models that presume phoneme identification or speech categorization as the basis for speech perception can be complemented by various adaptation and normalization mechanisms. What I find noteworthy about the integrated perspective, however, is that it accounts for contextual

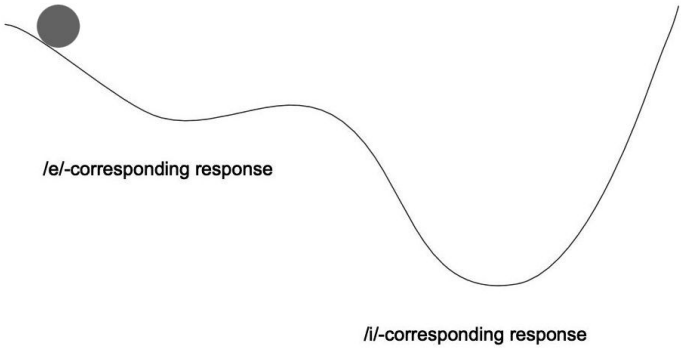

Figure 4.1: A scheme of an agent's state space. If we imagine the agent as the grey ball, context can be thought of as determining its speed. In laboratory setting, the ball goes slowly ("pushed" only by a context-free stimulus) and simply settles in the local optimum. In a conversation, additional cues give the ball more energy, making the smaller valley irrelevant – the ball will go over it and settle in the global optimum.

In laboratory conditions, where context is strictly controlled and listeners' expectations are supressed or manipulated, the difference between speech processing as categorization on the one hand and as elimination of action possibilities on the other may be rather cosmetic. In case of natural everyday communication, however, it becomes relevant. Holt and Lotto (2010,

effects on speech perception automatically, with no need for an additional mechanisms. It is because context does not *modify* the way we communicate, but fully *determines* it, as should be clear from sections 2.3, 3.1 and 3.3. Furthermore, research of normalization typically deals with *systematic* variation of speech, e.g. between-speaker differences (Sjerps, Fox, Johnson, & Chang, 2019), speech rate effects (Diehl, Souther, & Convis, 1980) etc., and normalization mechanisms are therefore conceived of as certain *shifts* in how acoustic cues are interpreted. The integrated perspective has a more general approach – the same mechanism that results in "normalization" of systematic differences accounts also for completely unpredictable variation, introduced e.g. by noise. Context (both situational and linguistic) sets the listener's expectations and these can then "shift" perception in any direction that is convenient, no matter what the source of the variation is. Normalization observed in laboratory settings then shows that psychoacoustic characteristics of speech not only serve as cues that set our expectations, but are themselves subject to these expectations, too.

p. 1223) point out themselves that the mechanism that listeners use in perception experiments might be different from the one that underlies speech processing in everyday communication, i.e. that we perhaps do not categorize speech signal in order to comprehend fluent speech. They do not suggest what the mechanism of everyday speech perception is, and I believe that the integrated perspective offers a good candidate. If we do not reduce attractors to categories and we let (psycho)acoustic cues navigate us through the same extremely rich and structured space of action possibilities as any other environmental input, we can account for everyday communication and at the same time capture speech categorization observed in laboratory settings as a special case.

4.2.2 The role of cues

It is important to point out that the listener will incline to the attractor determined by context and her experience-based expectations, unless something changes her position in the state space substantially. Instead of the question "How does a listener recognize a unit (say, the phoneme /ɛ/), what cues does she use to this end?", we might ask rather the contrary: "What could make the listener *abandon* the behaviour associated with the unit that is expected in this particular context?" Laboratory experiments are deliberately poor in context, and can therefore to some extent inform us about what acoustic properties are relevant for listeners (as Holt & Lotto, 2010, put it, they "highlight phonetic-level processing"). But instead of considering them as cues that help listeners categorize segments of the speech stream, these properties serve as parameters which, if set for a value outside of a (rather generous and flexible) range, can push the listener away from the attractor she would otherwise incline to in the particular situation.

To give an example: In the lab, it was found that bilabial plosives with voice onset time (VOT) below roughly 20 ms are interpreted as lenis stops by native speakers of English and values above 40 ms as fortis sounds (Pisoni,

1973).[9] The interpretation I disagree with would be that listeners use this cue in everyday communication to recognize phonemes (or other units), i.e. that when a listener hears for instance "I accept the [b̥]et", she uses the VOT of, say, 15 ms to determine that the last word is "bet" rather than "pet". Instead, the integrated perspective would emphasize that context enables prediction of the last word (it is difficult to imagine a real situation where both options are equally probable). Therefore the listener would be presumably much more tolerant of higher VOT than the laboratory results suggest – to divert her from the attractor associated with the word "bet" would require a rather extreme value. We can imagine that if the plosive was realized with the VOT of, say, 70 ms, the listener could still guess what the intended message was, but she would perhaps initiate a repair to make sure, because such an extreme divergence from the usual value can hardly be assigned to any of the usual sources of variation. How extreme the divergence from the usual value must be to disrupt a conversation is an interesting empirical question. Testing it would require a careful design which would approximate standard conversational setting as closely as possible, but at the same time would enable including utterances prepared in advance. For example, participants could lead a conversation with the experimenter on a predefined topic on the telephone, and the experimenter could at certain points play pre-recorded (and acoustically manipulated) utterances. The stimuli would differ in the degree to which the selected acoustic parameter would diverge from the natural (i.e. presumably the expected) value. With a sufficient amount of participants, a pattern could emerge in their reactions that would indicate what the "level of tolerance" for the manipulated parameter is. Due to the big amount of factors that could not be controlled for in this unrestricted setting (and also due to individual differences), I would expect to find rather a certain tendency than a clear-cut result. I would however still find it interesting to confront it with the results of a more traditional per-

[9]The specific values are not important for the point I want to make here – I ask the reader not to focus on the absolute numbers but rather to take them relatively.

ception test that would map participants' sensitivity to the changes in the particular parameter in a contextually poor situation. That is not to say that the results of traditional perception tests are not relevant. They might tell us what the acoustic properties worth investigation are and what the critical values in a specific setting are. It is rather the role ascribed to the discovered acoustic cues in real-life communication that is reassessed by the integrated perspective.

This attitude towards acoustic cues might resemble constraint-based formalizations of grammar such as the Optimality Theory (Prince & Smolensky, 1993/2004) in that the role of acoustic properties in linguistic processing is, so to say, negative. In the above example, the important thing is that English /b/ usually does *not* have VOT of 70 ms, and not so much that it usually has VOT of ca 20 ms. Boersma (2011) developed a model of the phonetic-phonological interface based on this OT mechanism (called BiPhon-OT). Acoustic cues to phonological elements are formalized as constraints, for instance "the first formant of 349 Hz does not correspond to the phonological vowel /i/", and we can establish as many constraints of this sort as we need, so that we cover the acoustic dimension in question with the desired level of granularity and across the desired range. Their ranking with respect to all other constraints then captures how important the particular constraint is for listeners and speakers[10]. Crucially, although BiPhon-OT distinguishes several levels of representation (auditory level, phonological surface level, lexical level etc.), all constraints can be evaluated *in parallel*, no matter which level they belong to. Cue constraints concerning mapping from the auditory form to the phonological surface form, such as the one mentioned above, can therefore directly interact with e.g. structural constraints (for instance "no complex onset") or lexical constraints concerning mapping from an underlying phonological form to meaning.[11] This parallelism allows for the top-down

[10]Boersma's (2011) model is bidirectional, i.e. the same constraint ranking is used to model both production and perception.

[11]An interested reader can find more on cue constraints and their interactions with other types of constraints in Boersma (2009).

effect described above – a higher-level (we may say "contextual") constraint that favours the word "bet" might be so strong that it makes many cue constraints irrelevant in this situation (e.g. the one that says: "VOT of 40 ms does not correspond to /b/"). Although the modular character of BiPhon-OT, i.e. that it postulates several rigidly hierarchically organized levels of representation, does not align well with the integrated perspective, the way linguistic *processing* is modelled seems to match it quite well – linguistic levels become largely irrelevant, because all constraints are interacting. In line with the integrated perspective, then, a cue can eliminate[12] some possible outcomes at any point, regardless of which level it theoretically belongs to. Furthermore, individual differences in processing are easily accounted for by BiPhon-OT, because the ranking of constraints is given by the language user's experience (during learning, constraints move higher or lower in the hierarchy according to which candidates violate them).

This is another reason why lack of invariance is not a problem from the integrated perspective. Due to the predictive and eliminative nature of communication, functionally equivalent units do not need to share any specific property to be recognized – they rather need to *lack* properties that would diverge listeners from the optimal attractor.

To summarize: Just as any other input from the environment, acoustic properties of speech signal help to eliminate the set of action possibilities that is open to the agent by a particular situation. They serve neither to identify abstract linguistic units, nor to categorize segments of speech, although in some situations, categorization can be the apparent result of the process. Perceptual categorization effects observed in laboratory settings reflect how a relatively low level of listeners' state space is organized. However, speech processing usually takes place in a rich linguistic and situational context that

[12]The eliminative nature of linguistic processing in BiPhon-OT, inherent to any optimality-theoretical account, by itself makes the model appealing from the integrated perspective.

might shape the state space on higher levels so that a particular lower-level setting is largely irrelevant. The role of psychoacoustic cues associated with individual speech sounds then turns out to be rather specific – a cue becomes relevant if it is able to divert the listener from the attractor she is inclining to. Whether this happens presumably depends on how strongly the cue is at odds with the listener's expectations and how robust the expectations are.

4.3 Turn taking

There is a domain of communication science that appears particularly suitable for application of the integrated perspective, namely turn taking. At the very core of turn taking research is *(re)action*, and the tight link between perception and action (a central feature of the skilled intentionality framework, see section 1.2.3) is therefore extremely pronounced there. Furthermore, it is clear that interactive oral communication is the homeland of language, and turn taking seems to be not only its important aspect, but also its (onto- and perhaps also phylogenetically prior) precondition (Levinson, 2016). It therefore fits well with Bruineberg's (2018) effort to keep continuity between "higher" and "lower" cognition (1.2), between humans and other organisms, between mind and life (1.1.2). Symptomatically, although there seem to be two partly competing approaches to turn taking (the "signalling" and the "anticipating" approach, see ?, ?, for a concise characterization of the positions), the need for a tight perception-action coupling is acknowledged in both of them – sometimes explicitly as a part of the proposed model (Pickering & Garrod, 2013), but always inevitably as a part of the *question*. Indeed, it is the ability to respond readily to others' turns, the astounding swiftness with which we act in our conversations (in other words, our skilfulness) that is the main explanandum.

4.3.1 The challenge (?) of turn taking

From the latencies obtained in various psycholinguistic experiments and analyses of natural conversations, it is clear that we cannot first receive our partner's whole utterance and then plan and implement our response – we react faster than this sequence would make possible (Levinson & Torreira, 2015). It seems necessary that we start preparing our response already during the interlocutor's utterance, and also that we anticipate quite precisely the end of the interlocutor's turn so that we can start speaking almost immediately after that (de Ruiter, Mitterer, & Enfield, 2006). Levinson calls this the "cognitive challenge" of turn taking (Levinson, 2016). The challenge is that interlocutors seem to be, at certain points in conversation, simultaneously comprehending their partner's utterance, predicting how it is going to proceed, planning their response and working on its implementation – quite impressive multitasking indeed. The situation seems to be further complicated by the fact that perception and production of speech partly draw upon the same neural resources; Levinson and Torreira (2015) suggest that this might be solved by rapid task switching.

Pickering and Garrod (2013) developed an influential model of language production and comprehension that explicitly addresses this matter. There are two ideas that the authors consider crucial for explaining rapid turn taking (and also other phenomena): 1) Prediction (of both my own and my interlocutor's actions) plays a vital role in communication. 2) Production and comprehension of speech (being just special cases of action and perception in general) are tightly interwoven – more than previous models known to the authors suggest. And it is this tight link that enables the predictive process. In Pickering and Garrod's (2013) account, the currently listening interlocutor is simultaneously perceiving the utterance, covertly imitating it and, on the basis of her own experience with production, running a forward simulation, thereby predicting what will come next. This prediction is then compared to the interlocutor's real production; but crucially, it enables the listener to start planning her response in advance, because she already knows

(from her forward model) what her interlocutor will say and how long it will take. The currently speaking party is doing something similar: apart from producing her utterance, she is also running a forward model of her own speech (which runs faster than the actual production), modelling comprehension of the planned utterance and then comparing the actual production to this prediction. Production is therefore not "clear" production and comprehension is not "clear" comprehension – each of the processes involves also forward modelling of the other one. This is how Pickering and Garrod (2013) see the tight link between them.

This model is interesting for me because it seems to aim for something similar as the integrated perspective and even draws upon similar core ideas, but employs a different strategy. The integrated perspective agrees with both claims stated above (i.e. that prediction is crucial in communication and that action and perception are deeply interwoven), but would not subscribe to the architecture proposed by Pickering and Garrod (2013), because 1) it understand something rather different under the term "prediction", and 2) it views action and perception as even more radically unified than Pickering and Garrod (2013) suggest. In what follows, I will try to clarify these points.

In section 3.3, I described communication as a set of interactions where my utterance creates a disattunement in my partner's brain-body-environment system that she resolves e.g. by her own utterance, which in turn causes a disattunement in *my* system etc. In a conversation, my interlocutor's utterance (together with other stimuli from the environment) navigates me through my state space, and an action corresponds to any point that I find myself in at any moment. As my interlocutor's turn is unfolding, I might be attracted e.g. to a silent nod, to a backchannel, or to my own full utterance. For simplicity, I will now focus on the last case, assuming that a verbal response is the appropriate reaction that my interlocutor's utterance should lead me to.[13] Where is comprehension and where is planning of my production in this

[13]Note that "appropriateness" of the reaction might be evaluated from two perspectives. From the interlocutor's point of view, my verbal response might be what she wanted to

scheme? The answer is that it is the very same process (as highlighted in section 1.2.3). Just as comprehension does not lead to an internal representation of the perceived utterance (since it is a process of elimination of action possibilities), production does not start with representation of my response that the agent would subsequently implement. Production and comprehension start at the very same point (or more precisely, they are both running continuously all the time), because they both are the same process of moving through the agent's state space.

The difference between the integrated perspective and Pickering and Garrod's (2013) model should be clear. There are, strictly speaking, not several processes going on at the same time, there is just our skilful interaction with our environment. Comprehension and production are not only tightly connected, each containing also elements of the other – they are unified. I consider this a further step in the same direction that Pickering and Garrod (2013) seem to promote, and I think it is an advantageous step to take because it considerably simplifies the architecture of communication. Not only does it eliminate the need for internal forward modelling of production and comprehension, which required simplified duplicates of the production and comprehension systems, but it also truly removes the separation between production and comprehension as such (which seems to be at least partly also Pickering and Garrod's, 2013, ambition).

The question may arise, when and how prediction takes place if we get rid of the forward modelling. Does the picture of an agent being navigated through her state space allow for any predictive process? The answer lies in section 3.2.3 where I explained what prediction means within the integrated perspective. To summarize it briefly: That an agent predicts what will be

achieve by her utterance; in terms of the skilled intentionality framework, the disattunement in her brain-body-environment system is reduced by my response. From my point of view, my verbal response is what reduces disattunement in *my* brain-body-environment system (otherwise I would not produce it). More often than not, these two perspectives probably meet. If they do not, the result is be either an undesirable pause or interruption – in both cases a lapsus that needs to be resolved.

said (and when) means that she arrives at the attractor that we tend to associate with the unfolding utterance even before the utterance is finished. Whether she can do this depends on her level of skilfulness – if she is so skilled that the cues that come early in the utterance are sufficient to get her to the appropriate attractor, "predictive" behaviour will occur. Skilfulness takes care of both predicting the content and reacting with the right timing (i.e. of both *what* and *when*). That the agent could in theory respond already in the half of her interlocutor's utterance (because she can predict the content) does not mean that she will do it. Her skilfulness will usually probably make the outset of her articulation dependent on additional impulses (e.g. turn end signalling intonation). This is because the state in which the agent would be able to finish the interlocutor's utterance (i.e. in which she predicts the content) is not identical with the state of launching a response. For instance, if the interlocutor suddenly starts coughing, then although the agent could already predict the rest of the utterance, she might ask "Are you all right?" instead of responding to the turn, simply because the impulse that came navigated her to another attractor than she would arrive at if nothing unusual happened. Taken together, we do not need explicit representative modelling of the interlocutor's utterances to account for predictive behaviour.

This account questions the view of turn taking as a "challenge" or a "puzzle" (Levinson, 2016; Levinson & Torreira, 2015), firstly because it takes a different view on the relationship between production and comprehension, and secondly because it questions some of the premises of the cited views. I will illustrate the second point with an example. Levinson and Torreira (2015) point out that "in production single words have to be plucked from a word lexicon consisting of over 20,000 entries", and since reaction times in general increase with the number of choices that the agent has, this is allegedly one of the reasons why producing an immediate verbal response is a challenge. The integrated perspective, however, would not have the agent choose from the whole lexicon at any point. The situation itself eliminates a vast majority of options, and as the interlocutor's utterance unfolds, the set

of options further narrows down. As pointed out in section 1.2.4, the action that will be taken is in fact already pre-selected by the shape of the agent's state space and by the agent's current position in this space. This is why the latencies obtained e.g. in picture naming tasks cannot be straightforwardly adverted to in theories of turn taking – such tasks make relevant an entirely different set of action possibilities than an interlocutor's turn in a natural conversation. The processes that results in a verbal response (even if it is the same response in both cases) are therefore hardly comparable. My interlocutor's utterance is a bit like a ball flying my way – I am continuously optimizing my position and reacting to any change in the trajectory of the ball in order to catch it. I am not choosing from the countless number of possible body positions and hand grips when the ball is already almost touching me – at that point, I am ready to catch it in the most efficient way I am capable of. When a stream of speech is coming, I am also continuously adjusting my actions so that the conversation unfolds in an advantageous way, which results in nodding, backchannels, silence, and at some point also producing my own utterance. The point here is not to see turn taking as something too special – it is no bigger wonder that we react to each other verbally in such a smooth succession than that we can dance with each other or play tennis.[14]

4.3.2 Incremental processing

Besides the unifying view of production and comprehension, another feature makes the integrated perspective a suitable frame for turn taking, namely the incremental nature of speech processing that it implies. Gregoromichelaki,

[14]It should be noted that even though the parallel between communication and physical activities (just as ball games or dancing) can in my opinion be illuminating, it does not by itself unequivocally support the integrated perspective. Pickering and Garrod (2013) draw this parallel as well, but they use it to make a different point – in their account, joint action such as dancing also requires prediction of one's own and the partner's movements, achieved via the same complex architecture that they suggest for verbal communication.

Kempson, Howes, and Eshghi (2013) illustrate by analysing compound utterances[15] that incrementality of both production and comprehension might help substantially to account for certain conversational phenomena, for example when my interlocutor finishes my utterance in a different way than I would finish it myself (in terms of both form and meaning). According to Gregoromichelaki et al. (2013), the "high-level" predictions based on the interlocutors' (assumed) intentions that Pickering and Garrod (2013) suggest are not needed and would not be efficient in these cases – I do not need to predict the rest of my interlocutor's utterance if I am determined to finish it myself. Gregoromichelaki et al. (2013) therefore suggest that we are usually dealing with speech signal in an incremental fashion, neither having full propositions planned as speakers, neither predicting them as listeners. The integrated perspective naturally supports such view, because if we see communication as based on progressive elimination of action possibilities, incremental processing of speech signal is implied. Also Gregoromichelaki's (2013) claim that what is constitutive for grammar is *procedural* knowledge is highly compatible with the integrated perspective – it is just another way of saying that language knowledge is a skill, which is one of the main ideas advocated in thesis.

Incrementality of language processing can also offer an answer to the question raised by de Ruiter et al. (2006), namely how listeners can predict turn ends from lexicosyntactic information. Discovering that lexicosyntax is much more important for anticipation of turn ends than intonation, the authors speculate on the precise mechanism that allows for such accurate turn end predictions and point out that we need to focus on the inherently *temporal* nature of lexicosyntax (instead of its atemporal structure, represented e.g. by

[15]Compound utterances are sequences of turns that are produced by different speakers, but form a linguistic unit (typically both syntactically and semantically). An example would be the last two turns of the following conversation:

A: I met John yesterday.

B: Oh, and did you –

A: Ask him about the message? No.

syntactic trees). The view of language as a discriminative code does precisely that, claiming that language processing is governed solely by the succession of discriminative elements, and not by an abstract structure characterizing the whole utterance.

4.3.3 Latencies in turn taking

Turn taking is a phenomenon that has been extensively experimentally studied, with a substantial attention dedicated to factors that systematically influence latencies between turns (e.g. Sacks, Schegloff, & Jefferson, 1974; Stivers et al., 2009; ?, ?; S. G. Roberts, Torreira, & Levinson, 2015). In order to see how the integrated perspective can deal with experimental results, I will briefly confront it with Roberts, Torreira and Levinson's (2015) study. I have chosen this one mainly because the research was based on a corpus of real conversations rather than on laboratory experiments. Given how crucial role the integrated perspective ascribes to context, the relatively high ecological validity of the data makes it suitable for the purposes of this thesis.

The aim of S. G. Roberts et al. (2015) was to evaluate the relationship and relative importance of two types of factors. One type were factors related to cognitive processing (such as turn length, frequency of the words used, concreteness, surprisal[16], or syntactic complexity). The other type were factors related to sequential structure of the conversation (whether turns constituted an adjacency pair, whether a response was preferred or dispreferred, whether a turn was a backchannel, or whether a turn involved laughter). The authors' prediction, confirmed by the results, was that it is not possible to state simply which type of factors is more determining or to find clear overall trends in how each of the factors influences the latencies. The factors stand in com-

[16]A caveat: This linguistic measure should not be confused with the key quantity of the free energy principle (see 1.2.1). The two quantities have the same information-theoretic basis, but linguistic surprisal has its own rigorous definition (see S. G. Roberts et al., 2015, p. 7).

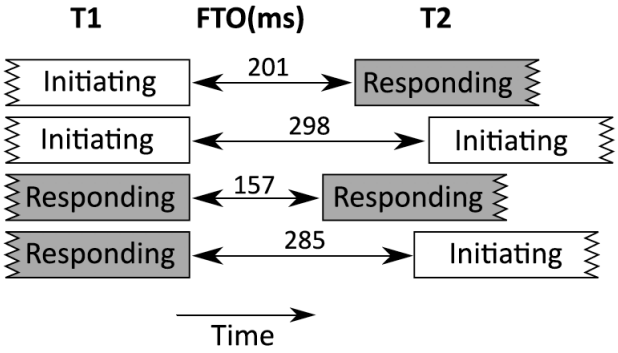

Figure 4.2: Mean time intervals between turns with different kind of sequential actions (responding and initiating). FTO = floor transfer offset; T1 = the first turn; T2 = the second turn. Reprinted with permission from Roberts et al. (2015, p. 12).

plex relationships to each other (such that a particular value of a processing factor may make one of the sequential factors highly relevant, or the other way around) and factors of other types (e.g. sex of the speakers) are also involved. I am neither going to present the results in detail nor will I go through all thirty-two factors included in the study. Instead, I will focus on several factors that turned out to be highly important. I will suggest how they could be interpreted within the integrated perspective and comment on the relevant results.

Responding vs. initiating

One of the most important factors affecting turn latencies was whether some of the adjacent turns included a responding action or not. S. G. Roberts et al. (2015) analyse this factor together with the complementary one, i.e. whether a turn included an initiating action or not. Figure 4.2 shows how these factors interact. As can be seen, the second turn comes with the shortest latency when both the turns are responding acts, and with the longest latency when both are initiating acts.

This factor seems to include two components, influencing the latencies in different ways (as is acknowledged also in the original study). First, there

is the normative aspect. Silence (be it within a turn or between turns) indisputably belongs to means of communication, and therefore its use can be expected to be to some extent systematic. In other words, longer latency of a turn might be a part of the most appropriate action possibility (as the works cited by S. G. Roberts et al., 2015, p. 4, suggest). For instance, introducing a new topic to a conversation might normatively require longer latency of the topic-changing turn, simply to signal the change. This might be the case of some of the pair where the first turn is a responding act and the second turn is an initiating act – these pairs indeed tend to have longer latencies on average than when a responding act follows an initiating act.

The second component would then be of processing nature. In my opinion, initiating and responding utterances might systematically differ in the degree to which they aim to change the interlocutor's internal dynamics. The structure of dynamic systems is hierarchical (see section 4.1) and the levels run on different timescales. If arriving at a certain attractor requires changes of higher-level dynamics, the process will take longer than if only low-level changes are needed. When I ask a question, I am creating a situation (all the more so if I am introducing a new topic). I am setting a frame within which the next couple of turns will probably take place. This might require my partner to adjust her higher-level dynamics. When she then replies, it surely brings about some changes in my system, but presumably not that severe ones – the higher-level setting stays the same and I therefore only need a relatively short time to arrive at the attractor corresponding to my next turn. This fits nicely with Roberts, Torreira and Levinson's (2015) results. We can take the latencies in a classic question–answer situation (i.e. the first row in figure 4.2) as a baseline and compare the other situations to it. In the baseline situation, speaker A sets the ground by producing the first turn and speaker B adjusts her higher-level dynamics and produces the second turn. If both turns are responding acts, the picture is different, because speaker B has already set the situation by an utterance that preceded the turn pair we are interested in. Speaker A got her higher-level dynamics adjusted and re-

sponds by the first turn; and speaker B then reacts by the second responsive turn within the same frame she *herself* has previously set. There is no need for speaker B to adjust her higher-level dynamics before the second turn and therefore it makes sense that the latency of that turn is shorter than in the baseline situation.

These two components, normativity and processing demands, interact with each other and it also seems reasonable to assume that the normative constraints on communication did not develop independently from the constraints posed by processing. Setting a higher-level frame by an initiating turn presumably requires more time-demanding changes on the side of the speaker, so the latencies get longer – and as this becomes a regular pattern, it gains a normative status that enforces longer latencies even in cases where the speaker might in fact be able to start the turn immediately.

Turn duration

Admittedly, not all factors from Roberts, Torreira and Levinson's (2015) study are easy to interpret from the integrated perspective. This holds e.g. for turn length which turned out to be highly important for turn latencies. Especially the fact that the length of the first turn was so highly ranked poses a puzzle for the integrated perspective, as I will explain below. But first I will comment on how the length of the second turn can influence its latency, because there my suggestion is more straightforward.

After a certain threshold, there was a tendency for longer turns to come with longer latencies. I suggest that the relation between the length of a turn and its latency again follows from the hierarchical structure of dynamic systems. If we consider a turn one action, i.e. corresponding to one attractor, then an attractor that includes more elements must be at a higher level of the hierarchy than an attractor corresponding to a shorter turn.[17] Longer

[17]To draw an analogy: If I am lying on the floor, standing up is a relatively high-level action (the whole body position is changing), whereas raising a hand is a low-level one. In accordance with this, standing up involves more elements (muscle movements in this

latencies preceding longer turns might then reflect the amount of time that is needed to arrive at the corresponding attractors. If the situation navigates the speaker to an attractor that only requires changes in the lower-level dynamics, evolving on a faster timescale, the attractor will be reached sooner than if the situation requires changes on a slower timescale.[18]

What is more puzzling is the influence of the duration of the *first* turn on the latency of the *second*. The question is how (if at all) the number of elements in a message (approximated by its length) can affect the listener's processing of that message. Recall the process of comprehension desrcibed in section 3.2.2 – the elements of a message serve the listener to successively eliminate action possibilities. The message is supposedly created by the speaker in such a way that it is sufficient for provoking the desired reaction, i.e. it should, in cooperation with context of course, navigate the listener to the appropriate attractor. Since elimination of action possibilities is a progressive process, it should be finished around the time that the listener receives the last discriminative cue (or sooner, thanks to anticipation), regardless of how many elements the message contains. We can compare the situation with navigation through a maze, where the message is a set of instructions and the goal is the attractor. If the instructions were received by the listener *before* she entered the maze and she would subsequently use them to go through it, then the number of instructions (i.e. the length of the message) would matter – the longer the message, the more time it would take to reach the goal. But if the listener is receiving the message *as she is walking through the maze*, making use of each element at the moment when

case), than raising a hand – i.e. standing up is a "long turn" and raising a hand is a "short turn".

[18]Note that according to this account, the latency is not a matter of *implementation* of a planned utterance. The latencies are longer for longer turns not because there is e.g. a more complicated structure in the speaker's mind that needs to find its way to the articulators; they are longer because it takes longer to arrive at the structure that is going to be said. To put it simplistically: The speaker does not need more time to say something more complicated, she needs more time to find out *that* she wants to say a more complicated thing.

it arrives, then no matter how long the journey is, the listener will reach the goal shortly after the last instruction. It is the second parallel that fits speech comprehension under the integrated perspective.

From this it seems that turn duration should not affect the latency of the following turn. However, it *did* in Roberts, Torreira and Levinson's (2015) study. Except for very short utterances, the overall trend was that longer turns tended to be followed by longer latencies of the following turns.[19] While admitting that this is a puzzle for the integrated perspective, I will suggest two directions where an explanation could perhaps be found.

The first rather vague suggestion draws upon the important property of human communication described in 2.4 – i.e. that an inherent part of communication is learning. While being navigated towards an attractor, the listener is also learning about what cues the speaker is using to get her there. In other words, she is rearranging her state space so that the next time, the cues that did not help this time will be more useful. In longer messages, there are more elements that might need to be dealt with in this manner, and that could perhaps result in longer latencies.

The second suggestion (not excluding the first one) is once again related the hierarchical structure of dynamic systems. Comprehending a long turn might require adjustments of higher-level dynamics, resulting in longer latency of the following turn. The first turn might be long precisely *because* it aims to change the listener's state space profoundly – by producing more cues, the speaker can affect the listener's state to a bigger extent.[20] If the adjustments of the higher-level parameters take longer than the duration of the first turn, a gap occurs before the following turn and the gap might be proportional to the severity of the adjustments.

[19]The relationship between turn durations and turn latencies is more complicated than the correlation I just sketched, which is also why the authors analysed the data with random forests, and not linear regression – see the original study.

[20]This is a where the parallel with a maze is inadequate, because it does not capture the hierarchical dynamic structure where changes on the faster timescales bring about changes on the slower ones.

These are both mere suggestions and would require significant empirical support to appear more convincing. There is an aspect of Roberts, Torreira and Levinson's (2015) results that seems to support the proposed connection between turn length and processing demands and it might inspire a follow-up empirical study. It is the fact that turn duration ended up higher in the importance ranking than the measures of turn complexity. There were two such measures: the height of the highest syntactic tree included in the turn and the number of clauses. Turn length and turn complexity are naturally correlated, as both Roberts, Torreira and Levinson's (2015) approach and the integrated perspective acknowledge; the nature of their relationship is, however, seen differently from those perspectives. When stating hypotheses for their study, S. G. Roberts et al. (2015) write that "longer utterances are likely to be more complex than shorter utterances, requiring more processing", and by "complex" they obviously mean the type of complexity captured by the measures mentioned above. This seems to imply that complexity is the decisive factor, length being rather an accompanying phenomenon. In theory, then, if two utterances were of the same length but differed in complexity (e.g. in syntactic depth), the more complex one should still bring about longer turn latencies. One might then expect that measures of complexity would turn out to be more important than turn duration in Roberts, Torreira and Levinson's (2015) analysis. The integrated perspective would make a different prediction. Since no structured representation of an utterance is needed for either production or comprehension (syntactic trees being only a descriptive construct, irrelevant for processing), the two turns should be processed equally quickly. Complexity is seen as the epiphenomenon here – what is decisive is the number of discriminative elements contained in the turn (approximated by turn duration). That turn duration ended up higher than the complexity measures then perhaps indirectly supports the discriminative view of language held by the integrated perspective.

Due to the correlation between turn length and turn complexity, it would be rather difficult to test those hypotheses directly. Pairs of utterances would

have to be found that differ in complexity but not in length, and vice versa. We can take two fabricated sentences as an example – sentence A: "Infuriated elephants ran towards the homestead.", and sentence B: "The blue bird that I saw there sang, but could not fly I think." Due to the same number of syllables, their duration could be very similar, but while A only contains one clause, B contains four, and is therefore more complex. If we found a sufficient number of such pairs in a corpus of spoken conversations, we might see whether it is primarily turn length or turn complexity (as captured by common linguistic measures) that correlates systematically with longer latencies. The fact that the corpus used by S. G. Roberts et al. (2015) is already annotated for the relevant parameters would facilitate the search for these utterance pairs. However, it might turn out that there is not a sufficient amount of them in natural conversation data (as the awkwardness of the made-up sentence A, full of multisyllabic words, suggests).[21]

To summarize: In turn taking research, crucial role of prediction and of a tight link between perception and action is widely acknowledged. The integrated perspective captures both of these features: Action and perception are seen as two aspect of the same process and predictive behaviour is a result of agents' skilfulness (for prediction is the capacity of reaching the appropriate attractor on the basis of early cues). This enables us to explain the remarkable smoothness of turn transitions without the assumption of demanding multitasking (present e.g. in the model by Pickering & Garrod, 2013). The incremental nature of language processing implied by

[21]One might try to circumvent this problem by inventing the sentences and testing participants in laboratory settings. However, this would require a very careful design that would approximate common communication as closely as possible. If the participants were asked to react to the end of the utterances in any predefined way, the experiment would have no point – if the attractor that should be arrived at was given in advance, we could not expect that processing of the incoming utterance would work in the same way as during a normal conversation. To be honest, I cannot think of a feasible design that would not run into this trouble.

the integrated perspective then accounts for other phenomena described in literature, namely compound utterances and the role of lexicosyntax in turn end prediction.

Factors that were proven to systematically influence latencies in turn taking can be interpreted from the integrated perspective with a varying degree of success. Certain sequence-organizational factors (such as initiating and responding nature of turns) can be explained well by normativity combined with hierarchical structure of dynamic systems. Certain processing factors, namely turn duration, are more problematic and further investigation would be required to decide whether they can be fully reconciled with the integrated perspective.

Conclusion

In this thesis, I strived to embed spoken communication into a general cognitive framework. In chapter 1, I adopted a view of cognition called the skilled intentionality framework, which claims a radical continuity between basic principles of life and human cognitive capacities. In chapter 2, I presented a theory of language based on discriminative coding, according to which language processing is a process of progressive elimination of possible messages, fully determined by context and relying heavily on continuous prediction. Chapter 3 is an attempt to integrate these two conceptions, exploring the relationships between their constitutive features and adjusting the discriminative theory of communication to the demands of the radically action-oriented theory of cognition. This was achieved mainly by substitution of attractors (i.e. action-related states of the agent) for messages defined by (linguistic or experiential) content. As a result, communication appears as a skilful interaction with other humans, during which the interlocutors mutually shape each other's environment in order to improve their own situation. They do so by producing signals that should navigate their partner to the desirable action, relying on a shared set of sociomaterial regularities that ensures high level of predictability in these interactions. In the last chapter, I tried to show how this perspective of communication can be applied to concrete phenomena investigated within the field of speech sciences.

I hope to have successfully shown that, in theory, spoken communication could be straightforwardly linked to the same principles that underlie cognition in its most basic form. I would like to believe the proposed integrated

account is viable; however, whether this is truly so, can only be shown with empirical work. Admittedly, direct empirical testing of the integrated perspective is difficult, above all because it claims that knowledge of language is a set of skills that are developed in a full dependency on context. Controlled laboratory experiments, for understandable reasons, require participants to deal with language in situations that are very different from everyday communication, and they therefore cannot straightforwardly tap into the skills that people usually rely on. In other words, the problem of ecological validity arises here in its most pressing form.

At least some aspects of the integrated perspective might, however, after all be tested (with a carefully selected method). For instance, the discriminative approach to language generates testable predictions concerning the distribution of discriminative elements. Ramscar (2019) presents several distributions drawn from various corpora to support his claims, concerning the word level. Recently, Linke and Ramscar (2020) used this analysed also at a lower level of language, examining the distributions of word-initial phones. This is a logical step: If language works the way that Ramscar (2019) suggests, sounds that occur at the same functional position (e.g. word-initially) should be distributed in the same manner as the groups of words analysed by Ramscar (2019) – e.g. proper names or colour terms. In functionally defined subsets, Linke and Ramscar (2020) indeed found the expected pattern. Similar analyses could be carried out on Czech to assess crosslinguistic validity of Linke and Ramscar's (2020) results. Notably, the distributions found by Linke and Ramscar (2020) showed much better fit to the expected distribution when the actually observed forms were analysed than when citation forms were used, suggesting that sublexical sound variation is systematic and functional. A sufficiently large and carefully annotated corpus of natural spoken communication would therefore be required to replicate the study.

Two other experiments were suggested in chapter 4. In section 4.2.2, I proposed an experiment that would map participants' sensitivity to unexpected and potentially misleading acoustic cues during a conversation. In

section 4.3.3 I proposed a corpus-based experiment concerning one of the most widely studied aspects of human communication – turn taking. The results of this experiment would indicate whether atemporal structure of utterances is relevant for processing (as representation-based views would predict), or whether processing is strictly incremental and it is therefore the number of discriminative elements that matters.

However, these experiments concern mainly the discriminative approach to language, rather than the integrated perspective as a whole. The value of the integrated perspective will therefore probably remain largely theoretical. From my point of view, it is nevertheless worthwhile to speculate about the conceptual background of a scientific field (even if practical consequences are not straightaway apparent). If a novel theoretical approach is appealing enough and spreads among the scientific community, it might eventually provoke hypotheses concerning topics that the author herself did not even think about. I naturally do not expect my master thesis to start a revolution within linguistics; however, I consider the two frameworks that I integrated here so convincing and promising that I dare assume that more researchers will notice them and pursue endeavours compatible with mine. I sincerely hope that under these circumstances, the theoretical work I have done in this thesis might be of use.

References

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada*, *11*, 295–328.

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438–481.

Boersma, P. (2009). Cue constraints and their interactions in phonological perception and production. In P. Boersma & S. Hamann (Eds.), *Phonology in perception* (p. 55–110). Berlin: Mouton de Gruyter.

Boersma, P. (2011). A programme for bidirectional phonology and phonetics and their acquisition and evolution. In A. Benz & J. Mattausch (Eds.), *Bidirectional optimality theory* (p. 33–72). Amsterdam: John Benjamins.

Boersma, P., Benders, T., & Seinhorst, K. (2018). *Neural network models for phonology and phonetics.* Unpublished manuscript, University of Amsterdam. (Downloaded 11. 5. 2020 from http://www.fon.hum.uva.nl/paul/papers/BoeBenSei45.pdf)

Bruineberg, J. (2018). *Anticipating affordances: Intentionality in self-organizing brain-body-environment systems.* Amsterdam: ILLC (FGw).

Bybee, J. (2010). *Language, usage and cognition.* New York: Cambridge University Press.

Casserly, E. D., & Pisoni, D. B. (2010). Speech perception and production.

WIREs Cognitive Science, *1*(5), 629–647.

Chemero, A. (2009). *Radical embodied cognitive science.* Cambridge: MIT Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–253.

Colombetti, G. (2014). *The feeling body: Affective science meets the enactive mind.* Cambridge: MIT Press.

de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, *82*(3), 515–535.

Di Paolo, E., & Thompson, E. (2017). *The enactive approach* [preprint]. Retrieved 2020-01-09, from `https://www.researchgate.net/publication/326120549_The_Enactive_Approach`

Diehl, R. L., Souther, A. F., & Convis, C. L. (1980). Conditions on rate normalization in speech perception. *Perception & Psycholinguistics*, *27*(5), 435–443.

Escudero, P. (2005). *Linguistic perception and second language acquisition.* Utrecht: LOT.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. J. (2013). Active inference and free energy. *Behavioral and Brain Sciences*, *36*(3), 212–213.

Friston, K. J., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, *159*(3), 417–458.

Galantucci, B., Roberts, G., & Langstein, B. (2018). Content deafness: When coherent talk just doesn't matter. *Language & Communication*, *61*, 29–34.

Ganong, W. F. (1980). Phonetic categorization in auditory perception.

Journal of Experimental Psychology, *6*(1), 110–125.

Gregoromichelaki, E., Kempson, R., Howes, C., & Eshghi, A. (2013). On making syntax dynamic: The challenge of compound utterances and the architecture of the grammar. In I. Wachsmuth, J. de Ruiter, P. Jaecks, & S. Kopp (Eds.), *Alignment in communication* (p. 57–86). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America*, *100*(2), 1111–1121.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception & Psychophysics*, *72*(5), 1218–1227.

Howhy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259-285.

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge: MIT Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. Cambridge: MIT Press.

Kemps, R. J., Ernestus, M., Schreuder, R., & Baayen, R. H. (2005). Prosodic cues for morphological complexity: The case of Dutch plural nouns. *Memory & Cognition*, *33*(3), 430–446.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107.

Levinson, S. C. (2016). Turn-taking in human communication – origins and implications for language processing. *Trends of Cognitive Sciences*, *20*, 6-14.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*(731).

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368.

Linke, M., & Ramscar, M. (2020). How the probabilistic structure of grammatical context shapes speech. *Entropy*, *22*(1).

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*(3), 253–257.

Machač, P., & Skarnitzl, R. (2010). *Principles of phonetic segmentation*. Praha: Epocha.

Newen, A., De Bruin, L., & Gallagher, S. (2018). *The Oxford handbook of 4E cognition*. Oxford: Oxford University Press.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 1–64.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psycholinguistics*, *13*(2), 253–260.

Powers, G. L., & Wilcox, J. C. (1977). Intelligibility of temporally interrupted speech with and without intervening noise. *The Journal of the Acoustical Society of America*, *61*(1), 195–199.

Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. Oxford: Blackwell.

Ramscar, M. (2019). *Source codes in human communication* [preprint]. Retrieved 2020-07-01, from `https://arxiv.org/pdf/1904.03991.pdf`

Ramscar, M., & Port, R. F. (2015). Categorization (without categories). In E. Dąbroska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (p. 75–99). Berlin: De Gruyter Mouton.

Ramscar, M., & Port, R. F. (2016). How spoken languages work in the absence of an inventory of discrete units. *Language Sciences*, *53*, 58–74.

Rescorla, W., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (p. 64–99). New York: Appleton-Century Crofts.

Rietveld, E., & Brouwers, A. A. (2017). Optimal grip on affordances in

architectural design practices: an ethnography. *Phenomenology and the Cognitive Sciences*, *16*, 545–564.

Rietveld, E., & Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology*, *26*(4), 325–352.

Roberts, G., Langstein, B., & Galantucci, B. (2016). (In)sensitivity to incoherence in human communication. *Language & Communication*, *47*, 15–22.

Roberts, S. G., Torreira, F., & Levinson, S. C. (2015). The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in Psychology*, *6*(509).

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology*, *110*(4), 474–494.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3, 4), 379–423, 623–656.

Sjerps, M. J., Fox, N. P., Johnson, K., & Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications*, *10*(1). (Article number: 2465)

Skarnitzl, R., Šturm, P., & Volín, J. (2016). *Zvuková báze řečové komunikace*. Praha: Karolinum.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(26), 10587–10592.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge: Harvard University Press.

Trubetzkoy, N. S. (1969). *Principles of phonology*. Los Angeles: University of California Press.

van Dijk, L., & Rietveld, E. (2017). Foregrounding sociomaterial practice in

our understanding of affordances: The skilled intentionality framework. *Frontiers in Psychology*, *7*(1969).

van Dijk, L., & Rietveld, E. (2018). Situated anticipation. *Synthese*.

van Westen, M., Rietveld, E., & Denys, D. (2019). Effective deep brain stimulation for obsessive-compulsive disorder requires clinical expertise. *Frontiers in Psychology*, *10*, 2294.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind*. Cambridge: MIT Press.

Warren, R. M., & Obusek, C. J. (1971). Speech perception and phonemic restorations. *Perception & Psychophysics*, *9*(3B), 358–362.

Wikipedia: Cognition. (n.d.). Retrieved 2020-06-27, from `https://en.wikipedia.org/w/index.php?title=Cognition&oldid=960738618`

Wittgenstein, L. (2009). *Philosophische untersuchungen* [revised fourth edition]. Chichester: Wiley-Blackwell.