Linguistics and Philosophy (2021) 44:773–812
https://doi.org/10.1007/s10988-020-09304-9

ORIGINAL RESEARCH

What would it mean for natural language to be the language of thought?

Gabe Dupre[1]

Published online: 26 June 2020

Abstract
Traditional arguments against the identification of the language of thought with natural language assume a picture of natural language which is largely inconsistent with that suggested by contemporary linguistic theory. This has led certain philosophers and linguists to suggest that this identification is not as implausible as it once seemed. In this paper, I discuss the prospects for such an identification in light of these developments in linguistic theory. I raise a new challenge against the identification thesis: the existence of ungrammatical but acceptable expressions seems to require a gap between thought and language. I consider what must be the case in order for this objection to be dealt with.

Keywords Philosophy of linguistics · Language of thought · Generative syntax · Philosophy of psychology

1 Introduction

Fodor (1975) introduced into contemporary philosophy of psychology the idea that cognition occurs in a language-like medium. This proposal leads naturally to the question of the relation between this proposed language of thought and natural languages, like English or Quechua, with which we are familiar. The simplest proposal is one of *identification*: the language of thought is natural language. Having a thought is tokening an expression of natural language, and token thoughts are individuated by their linguistic properties. However, despite this simplicity, the view that we ought identify

Josh Armstrong, Sam Cumming, Guillermo Del Pinal, Gabby Johnson, and Torsten Odland all read versions of this paper, and provided very valuable feedback. It also benefited greatly from discussion of numerous participants at the New York Philosophy of Language Workshop. Finally, special thanks to two anonymous referees for this journal for their unusually detailed and well-researched comments.

✉ Gabe Dupre
G.g.dupre@reading.ac.uk

[1] University of Reading, Reading, UK

the language of thought with natural language has not been widely adopted in philosophy or psychology.[1] A variety of traditional arguments, detailed in Sects. 3 and 4, have proved convincing. In particular, it has been argued that a variety of properties of natural language, in particular that natural languages are learned, variable, public, and ambiguous, cannot be properties of the language of thought, which is innate, uniform, private, and disambiguated.

A central difficulty with these arguments, as compelling as they have seemed, is that they rely on an intuitive understanding of what a natural language is. As linguistic theory has developed, especially in the last couple decades, however, a quite different picture of natural language has emerged. It has been one of the driving assumptions of generative linguistics that these everyday understandings of the term 'language' do not provide substantial constraints on linguistic theorizing, and so it is possible, indeed it appears likely, that the assumptions underwriting these arguments against the identification of natural language and the language of thought are false. Natural language, conceived of as the target of empirical linguistic research, is an aspect of individual's psychologies, and is plausibly innate, invariant, private, and disambiguated. More specifically, these intuitive properties of natural language are better viewed as properties of the *externalization* of natural language proper.

This suggests a re-evaluation of the traditional question of the identification. In fact, certain theorists working within this generativist tradition have argued that contemporary understanding of natural language does indeed suggest that identification is a plausible empirical hypothesis. Chomsky himself has been somewhat equivocal on this issue, although in recent years he seems to be more clearly promoting this thesis. In Chomsky (2007a) he states that "[i]t is often argued that another independent language of thought [i.e. independent of natural language] must be postulated, but the arguments for that do not seem to be compelling" (p. 16). However, in Chomsky (2015) he claims that we ought view "language as essentially an instrument of thought, even if we do not go as far as Humboldt in identifying the two" (p. 16), but later refers to the "underlying 'language of thought' provided by the internal language, the I-language, that everyone has mastered..." (p. 59). Most explicitly, in Chomsky (2007b), he states: "If the relation to the interfaces is asymmetric, as seems to be the case, then unbounded Merge provides only a language of thought, and the basis for ancillary processes of externalization" (p. 22), and this view is developed further in Chomsky (2017). In a series of papers, Hinzen (2006, 2011, 2013, 2014, 2015,

[1] With some exceptions, such as Kaye (1995). It is worth mentioning here, if only to largely put the issue aside, of the prominent view that language has 'explanatory or ontological priority' over thought. The driving intuition here is that language is required to impose a sufficient degree of specificity and determinacy on pre-linguistic cognition. This position has been argued for especially by Dummett (1991) and Dennett (1991). See Clark (1998) for an attempt to state this position in line with more recent work in the cognitive sciences. Davidson (1975) has presented a similar argument for the conclusion that language and thought are explanatorily inter-dependent. While related, these positions are not equivalent to the identification hypothesis which will be investigated in this paper. Crucially, defenders of the priority of language over thought typically assume that it is *particular* natural languages, understood as communication systems, which enable us to have genuinely propositional attitudes, and are thus skeptical of the internalist and nativist assumptions I will be utilizing in this paper. Likewise, this distinguishes the view defended in this paper from the Sapir-Whorf hypothesis, subject to much debate in linguistics, psychology, and anthropology, which claims that *particular* natural languages influence or determine the cognitive proclivities of speakers of these languages.

2017), following Chomsky, has provided the most detailed account of what such an identification would look like, how contemporary linguistics, and cognitive science more generally, makes it plausible, and what explanatory benefits would arise from making it.

While I believe they are lurking in the background of these works, the ways that contemporary generative theory undermines classical arguments against the identification of natural language and the language of thought have not been made fully explicit. My first goal in this paper will be to do just this. I hope to spell out more explicitly and fully than has been done previously how the picture of natural language presented by contemporary linguistic theory shows all of the traditional arguments for distinguishing between natural language and the language of thought to be unsound. However, while the metaphysics of contemporary linguistic theory are much more hospitable to this identification than traditional pictures of natural language were, the methodological developments suggest the opposite lesson. Whereas traditional linguistic theory assumed a relatively close correspondence between spoken language and linguistic competence, the demands of contemporary linguistic theory have expanded the gap between the two substantially. That is, as contemporary linguistic theory has advanced deeper and more abstract underlying grammars, it has been forced to exclude a wider variety of linguistic behavior from its purview. Of particular relevance for my purposes is the category of utterances deemed *ungrammatical but acceptable*. This class has been expanding as a result of phenomena traditionally taken to be paradigmatically grammatical being re-analyzed as relating to externalization. Perhaps the clearest example of this is word order itself: it is near-definitional of syntax that it studies of the arrangement of words in sentences, and from this it has traditionally, and reasonably, been assumed that the linear *order* of words is a grammatical/syntactic phenomenon, but much recent work views surface word order as instead relating to externalization processes (see chapter 7 of Hornstein et al. 2005). This increasing distance between grammar and observable properties of utterances predictably leads to breakdowns in the mapping between grammaticality and acceptability.

Utterances with this combination of properties pose a significant worry for attempts to identify natural language and the language of thought. If there are sentences that human speakers are able to interpret, but which are not licensed by the internal rules of the grammar, this seems to entail that speakers can have thoughts that are not expressions of natural language. But if this is so, the sets of possible thoughts and of possible (complete) linguistic expressions are not even extensionally equivalent, let alone identical. In Sect. 5, I shall go through some examples of expressions that generative linguists plausibly view as ungrammatical, despite the fact that they express thoughts which normal speakers can grasp.

In Sect. 6, I will describe several strategies that the defender of the identification of thought and language can use to respond to these kinds of examples. I will provide a qualified defense of the Identification Hypothesis, arguing that many apparent counterexamples of this sort seem plausibly explained. However, I will point to some examples which seem somewhat more difficult. I hope that this paper can make clear exactly what would need to be done to defend this alternative position.

2 The claim

Since Fodor (1975), the question of whether thoughts are conveyed by language-like vehicles has been one of the dominant threads in the philosophy of psychology. For the purposes of this paper, I shall not weigh in on these debates. I shall assume that something like the LOT hypothesis is correct, and will interpret it as making the following claims:

i. Cognition involves the manipulation of representations.
ii. These representations have constituent structure.
iii. Cognitive processes are defined over such structural properties.
iv. The semantic properties of these representations are not, in general, iconic.

Claims i–iii establish the representational theory of mind (see *inter alia* Fodor 1987 Chapter 1, and Fodor 1990 Chapter 1). Claim iv is intended to rule out representationalist theories which centrally posit non-language-like representational media, such as maps or images. As I said, I shall not be arguing for or elaborating on these claims. I mention them in order to make clear what the proposal is that I am evaluating, namely that the representations over which thought processes are defined are themselves the products of the language faculty, generated in accordance with whatever psychological principles govern this system. The idea that the language of thought is a natural language is of course only viable on the assumption that there is a language of thought, and so the issue only arises for those who accept i-iv. For certain stripes of connectionists (e.g. Churchland 1996) or dynamical systems theorists (e.g. Van Gelder 1995), then, the proposed identification cannot even be stated.

It is worth making explicit that the notion of a language of thought is being used here in a much narrower sense than it was in Fodor (1975). Fodor is concerned to show that representational theories of any aspect of the mind presuppose a structured medium over which computational operations can be defined. This will thus include the workings of the perceptual, navigational, and motor control systems, and any other systems which operate by manipulating representational states. It is clear that the identification of a language of thought with natural language is wildly implausible for most such systems. Non-human animals, from arthropods to apes, for which there is no reason to posit the possession of natural language, exhibit minds with these kinds of representational capacities.[2] In this sense, Fodor's broad use of the term 'language of thought' is misleading, as what he is really proposing is a language of mentation. It is a language of thought in a narrower sense, as applied only to *thoughts*, that I am interested in. Of course, it is is an open empirical question what the psychological kinds are, and which, if any, corresponds closely enough with our pretheoretic term 'thought' to be worthy of the name. This may involve some degree of *explication* (see Sect. 6.3), but for our purposes, it is sufficient to identify some seemingly important properties of thought, and view whatever empirical psychology discovers which has these properties (more or less) as 'thoughts'. Thoughts, as I use the term, are paradigmatically personal-level propositional attitudes.[3] That is, they are relations to complete propositional contents,

[2] See Burge (2014) for discussion.

[3] For brevity's sake, I shall restrict my attention to such propositional attitudes, but I am happy to include also attitudes that are constructed out of such propositions, such as interrogative (modeled typically as sets

attributable to a cognitive agent as a unified whole. Beliefs are paradigmatic examples of thoughts: agents believe *that p*, where *p* is a proposition. Thoughts are also, to use Stich's (1978) term, 'inferentially promiscuous': they systematically and reliably enter into rational transitions with other thoughts with a wide range of contents. Paradigmatic examples of thought processes include both practical and theoretical reasoning: I believe *that p* and *if p then q*, and so infer *q*, or I believe that *a-ing would bring it about that p*, and desire *that p*, and so I *a*. The proposed identification would then be that propositional, personal-level, inferentially promiscuous thoughts are natural language expressions. I will refer to the proposal that such thoughts are natural language expressions as the 'Identification Hypothesis' or 'IH'. IH is thus an empirical scientific hypothesis, not a piece of conceptual analysis or *a priori* metaphysics. The model here would be an empirical identity statement such as the identification of lightning with atmospheric electrical discharge. We can identify two psychological capacities, the ability to think and the ability to use and acquire a language. IH is the hypothesis that these capacities centrally depend on one-and-the-same underlying system; that exercises of one are exercises of the other.

The other relata of the proposed identification, natural language, will be the focus of this paper. We can identify two predominant ways of understanding this notion. One approach views natural languages as essentially public, shared entities. These entities consist of mappings from symbols onto meanings, and their properties are largely determined by social conventions. From this perspective, it is natural to ask how many people speak a particular language. The alternative approach views natural language as a psychological mechanism which maps one kind of representation onto another. This mechanism, in concert with many others, makes language use possible. An individual's language in this sense is token-distinct from that of any other speaker. Following Chomsky (1986), I will refer to languages in the former sense 'E-languages', and in the latter sense 'I-languages'. It will be my central contention that the plausibility of IH depends on interpreting 'natural language', in the latter sense.[4]

For different reasons, neither of these notions is perfectly transparent at this point. E-languages are largely continuous with folk understanding of what natural languages

Footnote 3 continued
of propositions) and imperative mental states (modeled as preference orderings on propositions). See e.g. Starr (2011) for a semantic argument for positing imperative mental states, and Friedman (2013) for a discussion of interrogative/inquisitive attitudes. Such attitudes seem to pose no immediate threat to the Identification Hypothesis, given the imperative and interrogative structures generated by the language faculty, which seem suitable for accounting for thoughts of this sort. *Sub-propositional* thoughts, as argued for in Grzankowski (2015) may likewise need to be accounted for with sub-sentential linguistic expressions, such as NP/DPs.
As throughout the paper, the issues here turn on as-yet-undecided empirical disputes. In particular, a 'mentalist' semantics for interrogative and inquisitive expressions is needed to identify these expressions with their corresponding psychological linguistic structures. A dynamic approach, which views the meanings of such expressions as abstract or social phenomena like changes to a discourse context, will pose deep problems for the account developed herein. See also fn.15.

[4] While I will be reading 'language' in the sense of *I-language* for the purposes of this paper, I am not assuming that there are no such things as E-languages. Just that I-languages are a reasonable referent for the term 'language' in the IH. If both notions of language pick out entities (as argued for in Stainton 1996, 2011), then 'language' will turn out to be ambiguous (as did 'mass' due to relativistic physics, see Field 1974), and IH will be true on one reading and false on the other.

are, and thus they inherit much of the imprecision characteristic of folk notions. However, there are certain features we can use to identify them: they are few relative to the number of speakers, people can grasp them more or less perfectly, they are tools primarily for communication, they exhibit significant variation, and are acquired through a process of learning from others. I-languages, on the other hand, are posits of a developing science, and thus claims about them are tentative and provisional. However, within at least the generative tradition, there is consensus that they are to a significant degree[5] species-universal and largely develop without much effort on behalf of either the learner or other speakers.

In the next two sections, I shall outline some standard arguments against IH. I shall show how these objections rely on an E-language understanding of natural language, and that when we replace this with the notion of I-language drawn from scientific linguistics, the force of these arguments evaporates. Along the way, the competing understandings of natural language should themselves become clearer.

Before getting into the theoretical arguments against the feasibility of this identification, we can note some general features in its favor. Firstly, to some the feeling that our thoughts are sentences of a natural language is highly intuitive. Carruthers (1998, Sect. 2.2) develops an introspection based argument that at least *conscious* thoughts occur in natural language.[6] The scope and force of such arguments are limited, in that firstly they only apply to *consciously accessible* thoughts and so may not generalize to thought in general, and secondly that it is far from clear that conscious introspection is a reliable guide to the workings of the mind. However, they offer at least a *prima facie* argument in favor of the identification of the language of thought with natural language.

Another very general motivation is parsimony. Given that we are already independently committed to the existence of natural language, if we could explain higher cognition with reference only to language of this sort, we make fewer theoretical commitments. However, as with all arguments from parsimony, this doesn't get us very far. We ought make as few theoretical posits as we can, *all else being equal*. That is, if positing only natural language, and no independent language of thought, were sufficient to account for all the relevant phenomena, then parsimony would favor making fewer posits. But, the question of interest is always whether all else is indeed equal. Only investigating the empirical prospects of the competing theories will settle this issue.

Perhaps more significantly, Hinzen (2013, 2014, 2017) has mounted an argument that there is a tight connection between what we can think and what we can express linguistically.[7] In particular, Hinzen argues that in many cases the best explanation for why certain thoughts are (im-)possible involves reference to which linguistic structures

[5] I will largely drop the qualifier 'to a significant degree' in the remainder of the paper, but as we are dealing with a biological object it will be assumed throughout that some variation is to be expected.

[6] Carruthers (2002) updates this view, and argues that language is the medium for *cross-modular* communication. This would serve to capture the inferential promiscuity of thoughts and would fit well with the view defended in this paper.

[7] Hinzen (2017) provides a different, and highly suggestive, argument for IH on the grounds that characteristic breakdowns in thought such as Autism Spectrum Disorder and Schizophrenia correlate with, and thus could be explained as, breakdowns in language. In Hinzen (2014), he *rejects* the existence of LOT, but only because he stipulates that LOT is a language of thought *in addition to* natural language.

are made (im-)possible by the language faculty. For example, Hinzen (2011, 2013) points out that lexico-grammatical properties of verbs appear to determine which kinds of thoughts we can have involving the concepts they express. Collins (2011) provides an apposite example:

1. Anton broke the bed.
2. The bed broke.
3. Anton made the bed.
4. *The bed made.

As these examples show, the thoughts we can have seem to track the expressions available in our language. More generally, the distinctions we make in thought track those made in language. Positing a language of thought independent of the language we speak seems consistent with the possibility that a *sentence* like 4 is ungrammatical but that nonetheless the thought it corresponds to is perfectly fine, perhaps analogous to sentence 2, indicating that something or other caused the bed to become made. But this is not what we observe. Sentence 4 is not merely ungrammatical, but *unthinkable*. Note the contrast between sentence 4 and ungrammatical sentences like “The child seems sleeping”. For these latter kinds of sentence, we can easily figure out what was meant (see Sect. 6.4), whereas sentence 4 simply doesn’t seem to provide a complete thought. Does it mean that the bed was made by someone or other, analogous to sentence 2, or that it made itself (analogous to sentences like “Anton washed”), or what? The ungrammaticality of 4 seems to preclude an answer to such questions. This is so despite there being no clear language-independent reason for the difference between these verbs, or the concepts they express. It is surely commonly understood that both making and breaking events require some independent force, agential or otherwise.

IH, however, provides a neat explanation for why sentence 2, but not sentence 4, expresses a thought. Ergative verbs, like ‘break’, allow for passive alternation, wherein the direct object (THEME) of a transitive construction can be raised to subject position in an intransitive construction. Non-ergative verbs, like ‘make’, preclude such an operation.[8] The explanation for this phenomenon is controversial and complex, centering around the claim that the lexical entries for ergative verbs mandate that the THEME (direct object) of these verbs is identified (‘theta-marked’), but identifying the AGENT (subject) is optional. Non-ergative verbs, on the other hand, mandatorily identify both their AGENTs and THEMEs. That is, ‘break the bed’ is a complete verb phrase, with no mandatory argument positions unfilled, whereas ‘make the bed’ is short an AGENT. Sentences 2 and 4 are formed by taking these verb phrases and raising their THEME arguments to sentential subject position. In sentence 2 no problem arises, as there are no further argument positions which need to be identified. However, in the attempt to form sentence 4, ‘the bed’ must be interpreted as filling the required AGENT role as well as the THEME role it has already been assigned, in violation

[8] The term ‘non-ergative’ comes from Collins (2011). The term is needed because ergative verbs are identified by two key properties (they are transitive and allow alternation), and other classes of agentive verb can be identified by their lack of either. Purely intransitive verbs, like ‘clap’ or ‘jump’, lack the former property and so are called unergative. Transitive verbs like ‘make’ lack the latter, and so can be called non-ergative.

of the Theta-criterion, which states that each argument must be assigned exactly one argument role (Chomsky 1981).[9] The details are not crucial for our purposes, what matters is that it seems that the thoughts we can have track the expressions available in our language. We explain the thinkability of 2 with reference to its grammaticality, and the unthinkability of 4 with reference to its ungrammaticality. This correlation between available thoughts and grammatical expressions is left unexplained if we posit a disparity between language and thought, but is predicted if we accept the view that the limits of language provide the limits of thought. This account thus inverts the perhaps standard view that we explain why we make the linguistic distinctions we do with reference to the kinds of thoughts we can have. In the example above, defense of this view would thus require some language-independent reason for treating 'break' and 'make' differently, which seems absent.

Despite these motivations, IH has not been widely accepted within philosophy. I will turn now to the primary reasons why not.

3 Traditional arguments against the identification I: the easy cases

I will examine four major arguments aimed to show that we cannot identify the language of thought with natural language.[10] As arguments concerning whether two systems are numerically identical, all four arguments can be stated as applications of Leibniz's law of the indiscernability of identicals. A property of natural language is proposed, which it is argued that the language of thought lacks, and so it is concluded that these cannot be the same language. I will describe these arguments in order of how serious a threat to the identification I think they pose.[11] In this section, I will detail two

[9] For the full details, see Hale and Keyser (2002). Given the overall strategy of this paper, it is worth considering an alternative hypothesis involving these data: that sentence 4 *is* grammatical, but is uninterpretable (hence unacceptable) for semantic reasons. This proposal inverts the analysis I have provided. It is conceivable that the syntax generates structures corresponding to sentence 4, but that semantic constraints, involving our concept MAKE, prevent us from interpreting them. That is, the difference between ergative and non-ergative verbs is not located in what arguments they require syntactically, but in what semantic roles must be filled. However, I believe that cross-linguistic data (e.g. Burzio's discussion of the distribution of the Italian particle 'ci' (Burzio 1986, Chapter 2), which seems to be allowed only in ergative constructions) suggests that this distinction is best drawn along syntactic lines, as is typical in the literature.

[10] There is one other major argument that I will not discuss, namely the argument from non-linguistic creatures. This argument aims to undermine the identification on the grounds that e.g. non-human animals and so-called 'pre-linguistic' infants, prior to developing their capacity to *produce* utterances, are able to think, but lack natural language. I will not discuss this argument as I take it to be basically an empirical question about which the verdict is currently out. While it seems unequivocal that such creatures have mental states, whether they have the kind of *thoughts* which concern this debate, i.e. personal-level, inferentially promiscuous, propositional attitudes, is controversial. Given the proposal that natural language is innate in humans, and the empirical data that human infants display competence with linguistic rules at a remarkably early age (see Yang et al. 2017 for a review), perhaps the best case for thought without language is in non-human animals. See the papers in Hurley and Nudds (2006) for discussion of the issues involved in attributing these kinds of thoughts to non-humans.
One possibility is that what makes human thought unique is that it occurs in human language, and that non-human animal thoughts occur in different languages. See Porot (2019) for discussion.

[11] There are certain other properties, such as conventionality or sharedness, which one could use to run similar styles of argument. For brevity's sake, I will leave these out as it should be clear how the argument would go, and what I say in response to the arguments I do discuss should apply equally to these.

such arguments, from publicity and underspecificity. I believe adopting an I-language approach to natural language undermines these arguments, or at least transforms them into empirical disputes concerning the details of linguistic theory. Once this approach is presented, and its power in defending IH is exhibited, I will turn in the next section to more serious traditional worries, from variation and acquisition. Responding to these will require more detailed and controversial claims about I-languages.

3.1 Publicity

Expressions of natural languages are, according to our intuitive understanding, *public* entities. That is, their properties are interpersonally available. This follows from our conception of language as primarily a tool for *communication*. If I were unable to pick up on the visual or auditory properties of your utterances, they would be unable to serve this communicative function. This public view of language is evidenced by the common assumptions that people *share* particular languages, that different people can speak such a shared language more or less well, and that linguistic expressions are essentially spoken or written (or perhaps signed). On the other hand, *pace* the behaviorist, thoughts are *private*. My having a thought of a particular sort does not result in any particular perceptually detectable property. Again, this falls out of the standard account of why we have language in the first place: language is needed precisely because it enables us to make our thoughts available for others. So this pair of contrary properties fits in perfectly with our everyday understanding of the relation between thought and language.[12]

3.2 Underspecificity

Despite the standard assumptions that language is *for* communicating thoughts, there are a variety of ways in which it seems *prima facie* to be less than optimal for doing so. In particular utterances often fall short of providing all the information in the thought they are used to express.[13] This can happen in a variety of related ways:

5. He stole them from her.
6. The dictator's behavior was sanctioned by the government.
7. She put the keys in the basket on the floor.
8. She loved him, and he her.

Sentences 5–8 all point to ways that our thoughts differ from the way they are expressed. Sentence 5 is an example of context-sensitivity or indexicality. Someone hearing this sentence may be unsure of whom or what the various noun phrases refer to, as the referents of these expressions can vary from context to context. However, in *thinking* such a thought, there can be no question as to whom or what one is thinking about. For me to think *He stole them from her*, I must know exactly at whom I

[12] It is precisely for this reason that Fodor spends a large amount of time responding to Wittgenstein's (1959/2009) arguments against the idea of a private language. Fodor viewed the language of thought as distinct from a public language partially in virtue of its being private. I shall argue instead that contemporary linguistic theory suggests that *both* the language of thought *and* natural language are private.

[13] This argument is advanced by Fodor (2001, 2008) and Gleitman and Papafragou (2005).

am addressing this (mental) accusation. Similarly for sentence 6. This sentence-form on its own does not determine whether it means that the government has endorsed or penalized this behavior, but the corresponding thought cannot fail to make clear which meaning it has. Sentence 7 is another case of ambiguity, this time of a structural, rather than lexical, sort. The sentence alone, spoken or written, is ambiguous between a reading according to which the keys begin in the basket and end up on the floor (perhaps in the basket, perhaps not) and one in which the basket begins on the floor and the keys are placed in it. But again, to think such a thought requires that one select a reading. Finally, sentence 8 is an example of ellipsis. The second coordinated clause ("and he her") appears to lack a verb, but interpretation of this sentence fills in this gap by replicating the verb from the first clause.[14]

What all of these phenomena are supposed to show is that certain properties are explicable only in strictly linguistic terms. Context-sensitivity, lexical and structural ambiguity, and ellipsis seem to be essentially properties of (public) language, not thought. Indeed it is hard to see how these *could* be properties of thought. And so by identifying language and thought, we lose the ability to account for these phenomena.[15]

Fodor (2008) makes this same point in terms of *compositionality*. Thoughts must, in order to explain their productivity and systematicity, be compositional. That is, the meaning of a complex thought must be determined by the meanings of its constituents (i.e. concepts) and their arrangement. However, language is non-compositional, as indicated by the above examples of underspecified linguistic expressions. The meanings of sentences are typically determined by variable features of the utterance context in addition to the meanings of lexical constituents and their arrangements, and so linguistic meaning is non-compositional. Thus, again we find a property of thought which is not a property of language: compositionality of meaning.

3.3 Why These Traditional Arguments Fail

As mentioned above, I believe the crucial failure in these objections to the identification of thought and language is the assumption that what natural language is is relatively accessible to our intuitions. There is a folk notion of 'language' which does indeed

[14] Related, but distinct, worries may arise with expressions which do provide the requisite information needed to grasp the thought expressed, but for which it is difficult to extract this information. Garden path sentences, such as "The doctor told the wife that the husband loved about her treatment" provide examples of this sort. Because 'tell' can select a complementizer phrase or a prepositional phrase in addition to its direct object ("tell the wife that *p*" vs. "tell the wife about *o*"), and 'that' can introduce a complementizer phrase or a relative clause, the parser in this case can be misled into constructing a structure which is then inconsistent with later encountered words. This makes assigning an interpretation to the whole sentence difficult. As in the cases discussed above, there is no corresponding property (in this case, difficulty of interpretation) of the thought.

[15] A more radical kind of worry may be raised by dynamic semantics. According to this program, the meanings of natural language sentences are different in kind from thoughts. While thoughts may be propositional and truth-conditional, these approaches view linguistic meaning in terms of potential to change the conversational context. If thoughts and sentences have different *kinds* of meaning, this suggests an identification will be impossible. While getting into the merits of such a project would take me too far afield, it is worth noting the attempts by e.g. Murray and Starr (2018) to show how dynamic meaning is grounded in updates to the mental states of conversational partners, which might lessen the difficulty of mapping language onto thought even within a dynamic theory of meaning.

have all the properties just mentioned, and thus differs from (a folk notion of) thought. But linguistic theory, as an empirical science, is not constrained to theorize about objects with these intuitive properties. One of the central aims of linguistics, just like all sciences, is to empirically determine which entities in the natural world are suitable targets for theorizing, and thus to identify the natural kinds in the target domain. And in fact, over the history of generative linguistics, the working conception of the proper target of linguistic theorizing has shifted radically away from this folk notion, in ways that appear to undermine these traditional arguments.

Chomsky's (1986) distinction between E-languages and I-languages encapsulates this shift. Chomsky intended for E-languages to correspond to our folk notions of a language. An E-language is an external, possibly abstract, object. Speakers of a particular E-language are similar in that they bear some cognitive relation ('knowledge') to this external entity. Speakers may, however, differ in how well they know the language they share. Young children, for example, may not yet have mastery of the language, but they are in the process of learning it, and thus acquiring mastery. I-languages, on the other hand, are states of an individual's psychology. In particular, they are states of the psychological computational system which functions to generate complex linguistic structures out of simpler linguistic structures. My I-language is the state of my language faculty. Other people's I-languages may be similar to mine, incorporating the same rules, but they are token-distinct psychological objects. Crucially, E-languages are individuated by the set of expressions (form-meaning pairs) they license, and which E-language a community or speaker knows is determined by which conventions they adopt.[16] If it is a convention in a given community that "The battle of Hastings was fought in 1066" is taken to mean that the battle of Hastings was fought in 1066, then that community speaks/knows an E-language containing this form-meaning pair as a member. I-languages, however, are type-individuated by the psychological processes which (partially) enable speakers to identify such form-meaning pairs. In principle, a single E-language could be spoken by a collection of agents with very different I-languages.[17]

Once the distinction between an E-language, as a social object, and an I-language, as a psychological object, is made, it opens the door to a variety of questions about the nature of this psychological object, and its relation to both this social object and to observable linguistic behavior. Once it is recognized that an I-language is supposed to be a genuine component of human psychology, it cannot simply be assumed that there is any simple relationship between this entity and these social objects or behavioral states. If the question of the relationship between thought and language is, as I take it to be, an empirical question about the relationship between two natural (psychological) kinds, then it is likewise an empirical question what these kinds are. That linguistic science has developed so as to investigate this internal, psychological object suggests that

[16] Lewis (1975) is the classical statement of this position.

[17] It is sometimes said that I-languages are *more fine-grained* than E-languages, in the sense just identified in the text. This is, however, incorrect. I-languages and E-languages are *orthogonal* taxonomies. As I shall be arguing, certain proposals in recent linguistic theorizing indicate that human speakers all have type-identical I-languages, even though they clearly have different E-languages. This confusion stems, I believe, from the assumption of a neat mapping between syntactic structures (expressions of an I-language) and possible public utterances (expressions in an E-language).

our intuitive judgements about which properties natural language has are insufficient. Whether natural language, *qua* target of theoretical linguistics, is indeed public and ambiguous is thus an empirical question. There are reasons to suspect that it is neither of these things.

Qua target of linguistic science, at least in the generativist tradition, natural language is an aspect of speakers' psychologies. Thus, even if thought and language are not to be identified, the one is no more public than the other. On this picture, while the language faculty *enables* (in concert with many other psychological mechanisms) the externalization, and hence publicity, of natural language, the internal computational system is not to be identified with whatever is thereby made public.

The underspecificity objection is similarly undermined in light of this conception of natural language. Structural ambiguity, on this account, strictly involves a many-to-one mapping from internal products of the language faculty to externalized public symbols. While utterances of the series of words in sentence 7 could be used to express distinct thoughts in distinct situations, it is a guiding assumption of much work in linguistic syntax and semantics that the language faculty generates structurally disambiguated expressions, and thus that this ambiguity is introduced only by the linguistically peripheral process of externalization, mapping these type-distinct structures into identical sounds, markings, or gestures. Likewise with ellipsis. Almost all work on this topic in generative linguistics assumes that the underlying, psychologically real, structure includes multiple copies of the elided material, but that some process of externalizing this structure deletes some of these copies. "She loved him, and he her" (8), and "She loved him, and he loved her" are thus, on this view, simply different ways of pronouncing the same linguistic expression. Thus the apparent gap between the linguistic expression and the thought it expresses, i.e. that one but not the other is in some way inexplicit, disappears. The assumption that ellipsis is a feature of pronunciation, not of differing underlying structures, is essential in explanations of why the elided material must be identical to some non-elided expression.[18]

The degree to which the view that the underlying structures of ambiguous expressions are themselves disambiguated is controversial varies depending on the cases. Analyzing sentences 7 and 8 in this way is basically uncontroversial. The ability to account for ambiguities (and lack thereof) in natural language with reference to different underlying structures has been one of the pillars of justification for generative grammar since at least *Aspects*, in which Chomsky explains the ambiguity of "Flying planes can be dangerous" with reference to distinct underlying structures (Chomsky 1965, p. 21). The ambiguity of sentence 7 is thus accounted for by positing two distinct syntactic structures, one in which 'in the basket' is a prepositional modifier of the noun phrase (NP) 'the keys' and 'on the floor' is the locative argument of the verb, and one in which 'in the basket' is the locative argument and 'on the floor' modifies the NP 'the basket'.

The worries raised by lexical ambiguity involve the assumption that the very same linguistic expression (e.g. 'sanction') can feature in distinct thoughts. This means that thoughts are individuated more fine-grainedly than lexical expressions and so the two cannot be identified. Again, this argument rests on a folk notion of language, according

[18] Although see Culicover and Jackendoff (2006) for critical discussion of this assumption.

to which words are individuated by their phonological properties. ‘Sanction’, on this conception, is one word with two different meanings. However, linguistic theory has no reason to stick with these everyday taxonomies.[19] And in fact, most theories of the lexicon instead view ambiguous expressions as involving the accidental sharing of phonological properties between two distinct lexical entries. So, on this view, ‘sentence 6’ is really a misnomer, as this does not denote a particular sentence, but a class of sentences pronounced in the same way. So the multiplicity of thoughts expressed is perfectly tracked by the multiplicity of linguistic expressions, and the objection disappears.

This also resolves the worries raised by Fodor (2008, p. 73) concerning cases from Kripke (1979) in which a single name is wrongly thought to refer to two distinct people. Fodor claims that this cannot be explained if we think in natural language, as natural language has just one expression here, whereas LOT can distinguish $Paderewski_1$ from $Paderewski_2$. Of course, the correct response here is that there are two *linguistic expressions* here, they are just pronounced in the same way (and, coincidentally, refer to the same person).

The treatment of deixis (as in sentence 5) requires slightly more machinery. On the face of things, the solution for cases of ambiguity can’t apply here, as different utterances of sentence 5 involve the same words. Whereas a mental lexicon will list multiple entries for ‘sanction’, it being more-or-less a coincidence that these words are pronounced identically, it will of course not list distinct entries for each use of ‘he’ or ‘them’. Thus sentence 5 seems to allow for variation in thought without variation in either grammatical structure or lexical items, the sole determinants of a linguistic expression.

What is needed here is a distinction between lexical *types* and *tokens*. These correspond to two distinct roles for lexical items in the use of language. On the one hand, lexical items are *repeatable*. For language to be useful, I must be able to re-apply the same expression in different contexts. This means I must store enough information about the expression that I can tell when it can be (re-)applied. The lexicon provides a store of just such information. On the other hand, lexical items are constituents of token linguistic expressions, constructed in real time in the process of producing and processing language (and, if IH is on the right track, thought). For a given speaker, there will be only one expression *type* ‘he’, but as many tokens of this expression as there are complex expressions in which it features. What indexical expressions demonstrate is that the meaning of a token complex expression is a function of the meaning of its *token* constituents, not of those of its constituent types.[20]

So, to defend IH from worries surrounding context-sensitive expressions, we must view it as a hypothesis about *token* thoughts: token thoughts are identical with token sentences. Token linguistic expressions are individuated by both their grammatical structure and their lexical constituents. In the case of stable expressions, whose contribution to a sentence is always the same, the type/token distinction could be fudged, but once we are dealing with expressions with variable semantics, it is absolutely

[19] And indeed whether this folksy taxonomy is coherent is debatable. Kaplan (1990) argues that public words should themselves be individuated more fine-grainedly than by their perceptible properties.

[20] This is obviously closely related to the distinction between content and character from Kaplan (1989).

crucial. While sentence 5 identifies a type of sentence which can express multiple different thoughts, each such thought corresponds to a distinct sentence token. This kind of argument generalizes to cover cases of polysemy as well. While tokens of the same word-type may contribute differently to different thoughts (e.g. ‘chicken’ in “I don’t think you should feed the chicken lamb” vs. “I don’t think you should feed the lamb chicken”), provided that each token thought is identical to some token linguistic expression, then IH can be maintained.

This response is similar to, but distinct from, Hinzen’s (2015) response to Fodor’s (2001) argument that there are elements found in thoughts which are absent in the language used to express them. Hinzen covers various different cases of this sort, with different strategies for dealing with each. In some cases, it is argued that the thought in question does not have the properties attributed to it (e.g. thinking, while in London, that it is raining is different from thinking it is raining *in London*), and in others that the linguistic expression does contain the meaning attributed to the thought (e.g. that “It is raining” does mean that it is raining *here and now* as a function of its grammatical structure). I shall not repeat all of Hinzen’s discussion here, but it is instructive, and seems to adopt a similar strategy to that in the previous paragraph of emphasizing the complexity and particularity which must be attributed to given linguistic structures.

4 Traditional arguments against the identification II: the hard cases

4.1 Variation

One of the most apparent properties of languages is their variation. On the surface, languages seem to be as different from one another as can be. Natural languages differ in their phonological properties (consider the complex consonant clusters of Czech, the rising and falling tones of Mandarin, and the clicks of !Kung), their morphology (compare polysynthetic Yupik to purely isolating Yoruba), their syntax (compare the strict word-order constraints of English with the relatively free word-order of Latin), and in myriad other ways as well. However, given the rejection of strong versions of the Sapir-Whorf hypothesis[21], it is widely accepted that the thoughts of speakers of these divergent languages do not show this same variation.[22] The way these thoughts are conveyed may differ in seemingly limitless ways, while the thoughts conveyed remain the same.

[21] This is, roughly, the idea that the thoughts one can have is substantially constrained by the particular language one speaks. See Pinker (1994, Chapter 3) for general discussion of why strong interpretations of this claim are no longer widely accepted.

[22] The invariance of thought is widely assumed, but it is still an assumption. If one were to allow for widespread variation in thoughts themselves, this would further undermine the traditional argument from variation in natural language against IH. It would then matter whether the variation in thought tracked the apparent variation in language. Of course, Whorfians argue along precisely these lines: variation in thought is explained by variation in language. For one recent instance of this debate see Li and Gleitman (2002) and the response by Levinson et al. (2002). But note that even in this debate, the linguistic differences are restricted to the lexicon, and further that the debates concern what is *natural* or *habitual* in thought, not which thoughts are or are not possible. Due to these restrictions, the Whorfian view argued for by Levinson and colleagues thus would not pose a deep problem for IH.

This disparity is what makes translation difficult but possible. If there were no linguistic variation, we would be able to communicate with everyone. But if thought itself varied, it is unclear whether communication between speakers of different languages, say through reading a translated work, would even be possible. It is because the thoughts expressed by "I'm hungry" and "Tengo hambre" are assumed to be the same, despite their quite different linguistic properties, that I am able to learn some Spanish by recognizing synonymies of this sort. Again, these sorts of phenomena do not even seem to be statable if we identify the language in which we think with the language which we speak.

4.2 Acquisition

A substantial chunk of Fodor (1975) is dedicated to arguing that natural languages cannot be the language of thought. His central argument is that if we do make this identification, we are faced with a regress. Natural languages, he argues, are *learned*. That is, children acquire a language by rationally responding to linguistic evidence in their environment, typically the utterances of nearby adult speakers. This fits in nicely with our layperson's picture of language. Intuitively, we learn the language we speak through various kinds of experiences we have with other people who have already learned it. This account of acquisition also seems to *explain* the variation we perceive: English speakers don't say things like "Tengo hambre" because the speakers from whom they learned had themselves learned rules prohibiting this kind of expression.

What Fodor was at pains to show, however, was that the language of thought could not be like this. That is, we cannot learn our language of thought. The reason for this is pretty straightforward. Fodor viewed learning as something like hypothesis testing. To learn, for example, whether one's language allowed unpronounced subjects, one forms the hypothesis "My language allows unpronounced subjects", and tests this hypothesis on the basis of one's primary linguistic data.[23] However, hypothesis testing presupposes a medium in which to state the hypotheses which are being tested. And so if the language of thought was to be learned in this way, then there must be some further language in which the learner is able to state hypotheses *about* the language of thought. And so on. Fodor's solution was to deny that we do learn the language of thought. If the language of thought is innate, i.e. it simply emerges as part of biological development, then there is no question about how to learn it. And it provides the medium in which hypotheses about natural language can be stated. This maneuver thus simultaneously showed how learning a language is possible (i.e via hypothesis testing), and undermined the looming regress (i.e. by claiming that the language in which such hypotheses are stated is innate). But, this proposal immediately precludes the identification of the language of thought with a natural language. If these are identified, we cannot leverage one into an account of the acquisition of the other.

Further, the innateness argument and the variation argument are mutually supporting. If Fodor is right that the language of thought must, in order to preclude a vicious regress, be innate, then those defending IH must likewise view natural language as

[23] Fodor's arguments focused on learning lexical semantic properties (e.g. learning that 'llama' applies to all and only llamas), but the arguments generalize to any account of learning based in hypothesis confirmation.

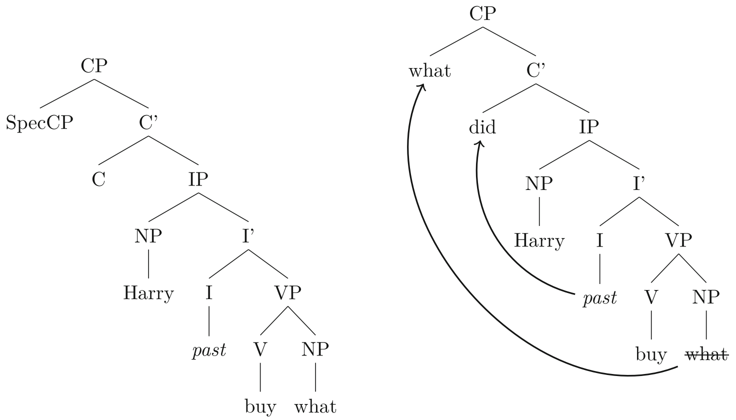
innate. However, the more languages vary, and the more this variation depends on subtle properties of the environment, the less plausible it is to view language as innately given.

4.3 Solving the hard cases

Much of what I said about the privacy and specificity of natural language is relatively uncontroversial. However, responding to the objections from variation and innateness requires going out further on a limb. For my purposes, the crucial proposal that has been developed in recent linguistic theorizing is that I-languages are species-universal. Whereas traditional approaches to generative theory assumed that much of the work in explaining linguistic variation was to be done within the I-language, certain recent work has suggested that we ought view this variation as instead a product of the different ways in which the same internal system is 'externalized'.[24] To see the difference, compare the following two possible explanations for the difference between a language in which (some) wh-expressions are pronounced at the beginning of a sentence (as in English) and those in which they are pronounced wherever in the sentence they receive their semantic interpretation (like Mandarin):

9. What did Harry buy?
10. Húfēi mǎi-le shénme.[25]
    Hufei buy-PERF what?

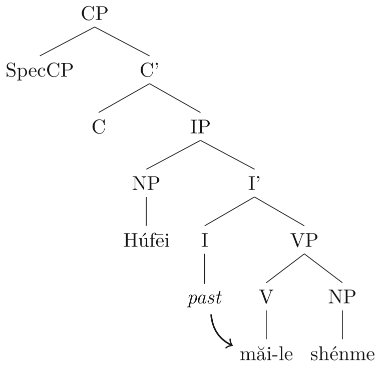
Traditional theories viewed this as a genuinely syntactic difference. The underlying structure of sentences 9 and 10 differed in that the wh-expression 'what' in 9 underwent movement:



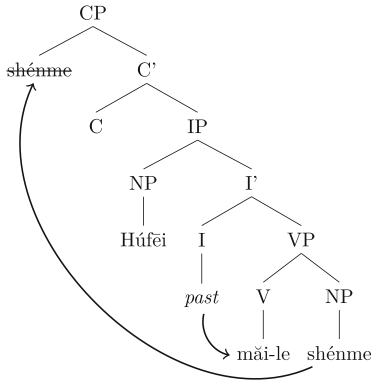
whereas the wh-expression 'shénme' in 10 did not:

[24] As well as the already discussed differences in the lexicon.

[25] The example is from Cheng (2003).



However, there is an alternative analysis, according to which the underlying syntax of both 9 and 10 is the same[26], and they differ only in how this structure is 'externalized', or pronounced (strike-out indicates that this expression is unpronounced)[27]:



These contrasting explanations suggest different understandings of the language faculty and of its states, I-languages. The former explanation assumes a framework in which language variation is explained *internal* to the language faculty, whereas the latter views variation as a product of the ways that systems outside of the language faculty handle the products of the faculty itself. On this latter account, the difference between wh-movement languages like English and wh-in-situ languages like Mandarin is constituted by differences in the ways that the sensory-motor systems interpret the generated linguistic structures, i.e. whether the lower or higher copy of

[26] I am ignoring the fact that in English tense is raised to C, whereas in Mandarin it is lowered to V, but a similar story could be told about this difference. In fact, given the assumed impossibility of rightwards grammatical movement (i.e. away from the 'trunk'), this is typically viewed as a merely phonological difference.

[27] The example is an illustration of the kind of analysis that I take to be the best case scenario for IH. I am not here defending this analysis. However, it does have some support from the fact that in-situ expressions seem to be subject to the same distributional constraints as moved wh-expressions. Where raising a wh-expression would violate constraints on movement, as when the wh-expression is an adjunct within a complex noun phrase, wh-in-situ languages preclude wh-expressions as well. This could be easily explained if the wh-expression is required to move in both languages. This proposal is however highly controversial. See Cheng (2009) for difficulties and alternative analyses.

a wh-expression is relevant for commands to the production systems.[28] Given this account of wh-movement, one could generalize and envision an approach to linguistic variation which treated *all* grammatical differences in this way, as contained within phonology, not syntax proper. This latter approach, viewing linguistic variation as a product of language-external 'externalization' systems has been suggested by Chomsky consistently over the last couple decades (see especially Berwick and Chomsky 2015) and has been forcefully advanced by Cedric Boeckx (see especially Boeckx 2010, 2014; Boeckx and Leivada 2013).[29]

On such a picture, the language faculty is invariant across the human population. I-languages, states of the language faculty partially responsible for the acquisition and use of language, consist of a computational system capable of constructing complex representations out of simpler representations. The syntactic principles governing such construction are the same for all human language users, as are the conceptual/semantic principles governing interpretation. The differences between languages, which are phenomenologically so overwhelming, are reducible to differences in the ways in which these identical systems interact with extra-linguistic systems of production and to differences within the lexicon.[30] This approach to the study of language suggests significant deviation from our folk notion, in ways which substantially undermine the traditional arguments against identifying natural language and the language of thought.

As in the above discussion of the division between the grammar and the lexicon, it is worth stressing that this is not merely a terminological question about what we *call* 'natural language'. The question is a substantive one: What are the components of the mind? What are the natural psychological kinds? The traditional explanation of linguistic variation proposes that there is a single psychological system responsible for both the similarities and differences between different natural languages. The more recent proposal denies this: similarities are accounted for by the species-universal language faculty, while differences are explained by divergent strategies for expressing linguistic structures. The former system is argued to be the distinctive feature of human psychology which makes human language possible, whereas the latter are more ancient systems, similar to those in many non-human animals, which have been co-opted for linguistic purposes. In this sense, it is an empirical *discovery*, not mere linguistic stipulation, that the species-universal language faculty *is* the natural phenomenon of natural language.

[28] Strike-out, in the tree-diagrams above, is, on this account, to be viewed as merely a theorists' notation device for keeping track of what use subsequent non-linguistic systems make of these structures, not reflected in the structures themselves as found internal to the language faculty.

[29] For some further evolutionary motivation for this proposal, see Huybregts (2017), who argues that assuming that the language faculty developed prior to the ability to externalize its products can shed light on some complex difficulties surrounding the evidence for the dating of the emergence of language.

[30] While this approach has appeared most plausible under the banner of the Minimalist Program, it could be stated easily enough in previous approaches: Deep Structure is uniform throughout the species and provides the Language of Thought, and all variation can be attributed to varying principles transforming Deep Structure into Surface Structure. While this view was rarely explicitly endorsed, one can view substantial amounts of early generative work as aimed at restricting language variation to Surface Structure. See Newmeyer (2005, Chapter 2) for discussion. Chomsky (1965, p. 117) attributes this position to the Port Royal Grammarians.

Of course, as with almost any scientific development, there will be some degree of linguistic decision in determining how to describe what has been discovered using everyday terminology. The purported 'discovery' that tomatoes are not fruit amounts to a decision to adopt a botanical taxonomy rather than a layperson's culinary taxonomy. The reason that the term 'fruit' can be retained through this lexical shift is that there is a close-enough correspondence between the earlier and later uses of the term. Likewise, I believe that linguistic theory provides a picture of what language is, in the sense of the ability to utilize a system of symbols unlike that of any other animal, which serves as a suitable replacement for our previous folk notion. Of course, much of the folk notion will not be retained (e.g. publicity), but neither will much of previous *scientific* understandings of language (e.g. rules governing surface word order). A plausible scientific theory posits the existence of a species-invariant computational system, part of human's innate endowment. A reasonable linguistic proposal is that such a system would deserve the name 'language'. IH is the combination of both of these.

If such proposals are correct, then apparent linguistic variation does not actually indicate different computational systems. If we identify this computational system with natural language, then there is no strictly linguistic variation. There is instead variation only in the way that language is externalized. This is expressed by Chomsky (1993, p. 50) when he claims that "[t]he 'computational system' of language that determines the forms and relations of linguistic expressions may indeed be invariant; in this sense, there is indeed only one human language." Our folk notion of language, which individuates languages partially in terms of such externalization properties, thus misled us into individuating languages much more finely than naturalistic inquiry suggests. Yet another intuitive difference between thought and language turns out to be at least a complex empirical issue, on which the debate is far from settled.

As mentioned in the setup of the objection, objections from variation and acquisition largely stand or fall together. Large amounts of environment-specific variation suggest learning, whereas a species-universal syntax and semantics suggests innate guidance. In fact, one of the arguments in favor of the view of linguistic variation as located exclusively in the forms of externalization is that it is precisely these properties that are available to the learner. That is, surface word order, phonological properties, and certain aspects of the lexicon, are largely made available in the primary linguistic data. Syntactic and semantic properties, on the other hand, are not strictly perceptible. Note that this view of linguistics insists on a strict demarcation between syntactic properties, like hierarchical constituency structure, and surface word order. The former are (alleged to be) species-universal, while the latter are extra-linguistic phenomena depending partially on the language faculty, but also on a variety of other psychological systems, especially 'interface' systems, and processes. It is only the latter that the learner can identify directly from the linguistic data, but due to movement and other phenomena which complicate the mapping from syntactic structures to utterances, these provide at best unreliable evidence for the former. Claiming that the underlying structures are known innately, and that all that must be learned is how these underlying structures are mapped onto externalized expressions massively reduces the difficulty of the acquisition problem.[31] Keeping with our assumed identification between natural language and this computational system, this concludes the case that traditional arguments against identifying natural language with the language of thought fail on account of an outdated understanding of natural language. Natural language, on this conception, is both innate and invariant. As we think that the language of thought must share these properties, the prospect of an identification is at least still open.

I wish to briefly touch on a possible objection which is not in the class of traditional worries about IH: the objection from psycho/neurolinguistics. While still very little is known about the neurobiological processes involved in the acquisition and use of language, future developments may be crucial in evaluating IH. In particular, IH predicts a very close relationship between syntactic and semantic processing. If semantic processing involves identifying the thought associated with a complex expression, then IH predicts that semantic processing requires syntactic processing. 'Syntax first' models of processing, such as Friederici (2002) are thus highly compatible with such a picture. However, proposals involving 'autonomous semantics', semantic processing occurring independently of syntactic structure, would pose a very serious worry.[32] Baggio (2018), for example, proposes that interpretation is at least partially independent of syntax. One primary source of data in favor of such a position is the existence of agrammatical aphasiacs who are able to interpret grammatically complex sentences when the semantic content is predictable on the basis of the lexical items, but not when it is not:

11. The apple that the boy is eating is red.
12. The cat that the dog is biting is black.[33]

In 11, but not 12, lexical meaning makes one assignment of arguments to their predicates highly plausible, and aphasiacs can utilize this information to interpret the sentence correctly. However, in 12, one needs to identify the grammatical/thematic relations between argument and predicate (i.e. that 'the cat' is the object of 'bite' and 'the dog' is the subject) in order to correctly identify the sentence's meaning. Whereas boys typically eat apples, and not vice versa, it may be assumed that cats and dogs bite one another frequently enough to not provide a strong cue for interpretation independent of grammatical constraints. This dissociation between grammar and semantic competence is apparently at odds with the predictions of IH.

The difficulty with interpreting such phenomena is that agrammatical *behavior* does not guarantee that there are deficiencies in the internal grammatical system. In particular, it is at least possible that these failures to interpret complex grammatical

[31] This approach is strengthened by what many now view as the failure of the Principles and Parameters approach. In particular, viewing language as the setting of parametric values has led to the positing of many highly-specific 'micro-parameters' in the face of apparent linguistic variation. The number and specificity of these has seemed to many to be non-explanatory. Given the deficiencies with the most plausible, and highly-touted, account of linguistic variation within the generative approach, we may be better off simply denying that there is any strictly grammatical variation, and thereby restricting variation to surface-level properties. See Newmeyer (2005) for a picture of language variation along these lines.

[32] Note that the concern here is strictly with compositional semantics, the identification of the semantic properties of complex expressions. IH makes no predictions about the processes of lexical semantic interpretation.

[33] From Caramazza and Zurif (1976).

structures are precisely due to failures to map external linguistic inputs onto internal grammatical structures. This interpretation would locate the failure as outside the strictly linguistic system, leaving the possibility that the thoughts of such subjects are expressed by linguistic structures a live option. Of course, it is an empirical question how such debates will be resolved. One motivation for viewing aphasia as a problem with performance (i.e. externalization) rather than competence is that production and comprehension can be dissociated, with one but not the other affected (see e.g. Friederici 1981). This suggests that it is the input/output systems which are damaged, as the core grammatical system is involved in both, and so damage to it should equally undermine production and comprehension.

The variety of neurolinguistic proposals in the literature suggests that there is little in the way of consensus here. While many parties accept a 'dual stream' model, according to which linguistic processing is divided into distinct processing routines in distinct neurological regions (the dorsal and ventral streams), there is much debate about which linguistic properties are processed in which streams. Mostly it is agreed that the dorsal stream is used for connecting sounds to action and motor control, but syntax and semantics have been argued to be processed together in the ventral stream (Hickok and Poeppel 2007), and to be interaction effects between both streams (Saur et al. 2008; Bornkessel et al. 2005). Both of these options are consistent with the claim that having a thought/interpreting a sentence requires being able to construct a grammatical structure. Given that such proposals do not treat syntax and semantics as processed independently (i.e. in parallel streams), they suggest that data like 11 and 12 ought be accounted for without positing strictly syntactic deficits. While I take these studies to be inconclusive on this issue, hopefully they point to a further area in which progress in answering the philosophical question about the relation between thought and language can be made by drawing on work in the sciences.

Another possible response to arguments of this sort could be given if we have reason to posit significant disparity between the strategies adopted by the parser and rules governing the grammar, as in Ferreira and Patson (2007) and Ferreira and Lowder (2016). If we posit two different kinds of structure-forming operations in language use, one process following the rules of the grammar and generating 'deep' structures, and another positing specialized and simplified heuristics of the parser and generating 'shallow' structures, we can account for the difference in 11 and 12 with reference to disruption only to the latter process. This would again involve the claim that aphasia is a performance, not a competence, phenomenon. Aphasiacs are unable to generate 'shallow' parses without the guidance of lexical semantic information and associations between concepts, and so cannot use these shallow trees as input to genuine sentence/thought generation, whereas they can do this with the shallow parses generated in response to sentences like 12 which provide the needed semantic clues.

Before finally moving onto the new problem with identifying thought and language, it is important to stress just how tentative all of this is. Positing genuinely syntactic differences between I-languages is still the most common approach to language variation in the linguistics literature, even within contemporary generativist work.[34] If

[34] See for example the wide range of syntactic differences between languages proposed in Cinque and Kayne (2005).

such an approach is correct, some (but not all) of the arguments given above will be successful. If English and Mandarin differ in that the grammar of one, but not the other, mandates wh-movement, but the thoughts that speakers of these language can have do not differ, then the language of thought cannot be natural language.[35]

5 The new problem: the acceptable but ungrammatical

While I take the conception of what natural language is, qua target of theoretical linguistics, as developed over the past few decades to be more amenable to identification with the language of thought, the methodology of these sciences has moved significantly in the other direction. In particular, while traditional generative theories, especially those in the transformationalist paradigm, assumed a close correspondence between the structures output by the language faculty and utterances, the increased abstraction of contemporary theories has led to a widening gap between these two phenomena. Significant proportions of linguistic behavior are thus viewed as non-reflective of the underlying system, as resulting from the influence of a variety of external, non-linguistic systems.

There are two kinds of gap between the outputs of the grammar and produced utterances: grammaticality without acceptability, and acceptability without grammaticality. Utterances are acceptable when native speakers judge them to be natural. Expressions are grammatical, roughly, when they are generable by the language faculty. That these two are not the same has been one of the central assumptions guiding the methodology of generative linguistics. However, while this gap has been central to generative theory since its inception, the magnitude of the gap has increased significantly. In the early days, it was largely assumed, explicitly or implicitly, that acceptability tracked grammaticality modulo certain kinds of 'deficiencies'. For example, Chomsky (1965) identifies "memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance" (p. 3) as sources of disparity between linguistic performance and competence.[36] Canonical examples of unacceptable but grammatical sentences include center-embeddings such as "the mouse the cat the dog chased caught squeaked", which despite being formed by perfectly normal grammatical rules place too substantial a burden on parser memory to interpret. However, this fairly simple relationship between acceptability and grammaticality has little to motivate

[35] In this case, one could defend a weaker version of the identification which viewed LOT as providing the *vehicles* of thought, so that tokening a thought necessarily involves tokening a linguistic expression, but that the same thought can be tokened by tokening different expressions. Language being invariant is the *cleanest* way of defending IH, but one might still consider it a vindication of IH if the weaker hypothesis, according to which minor differences in the vehicles of thought (e.g. wh-movement in one, but not the other) left the thought itself unchanged, were true. One could thus identify some 'core' aspects of grammar which determine which thoughts are available and the thought conveyed by each expression, and allow for variation in the 'periphery'.

[36] This attitude is endorsed in more recent work by Tonhauser and Matthewson (2015) when they claim (p. 19) that "native speakers judge a linguistic expression uttered in a context to be acceptable if and only if the linguistic expression is syntactically well-formed, felicitous and has truth conditions which are compatible with that context." making grammaticality a necessary, but not sufficient, condition on acceptability.

it. If extra-linguistic factors can serve to make grammatical sentences unacceptable, there is no reason why they should not also be able to make ungrammatical sentences acceptable.

It was the *methodology* of early grammatical theory, not the theoretical claims themselves, that suggested such an asymmetry. Given the tools of transformational grammar, it was possible to account for just about any linguistic data, and so sentences taken to be acceptable could easily be ‘predicted’ by grammatical theories by introducing new transformational rules. However, as linguistic theory has developed, the constraints on appropriate theory-formation have become significantly stronger, especially in the contemporary Minimalist Program. This means that it is often better to exclude the observations from the purview of the theory, and denounce even acceptable sentences as ungrammatical, than to complexify the theory so as to account for them.[37] Trotzke et al. (2013) provides a very clear recent statement of this approach: “[C]ertain attested utterances are explained outside the grammar proper. This permits a much simpler grammar than would otherwise be possible...” (pp. 26–27).

One complication in all of this is that claiming that an utterance is grammatical is, strictly, a kind of category mistake. This is why I flagged that the above definition of grammaticality as generability by the language faculty is only roughly accurate (see also Sect. 6.4 for further difficulties with this account of grammaticality). Along with the notion of a natural language, the notion of grammaticality has itself undergone significant revision in the development of generative grammar. As stated above, grammaticality is the property of being generable by the language faculty. But the language faculty generates structured psychological representations, not publicly observable utterances. The operational notion of grammaticality, as applied to sentences, is something along the lines of: producible via a relatively transparent mapping from the syntactic structure to a linearized utterance. Exactly what this mapping is is a matter of much debate in contemporary phonosyntax. Kayne (1994)’s Linear Correspondance Axiom, which states that linear order is determined by asymmetrical C-command, is one famous proposal for such a mapping. Some such proposal is needed in order to make sense of the notion of the grammaticality of a sentence. In this way, the grammaticality of a sentence, as opposed to a syntactic structure, is a derived notion dependent on both syntactic and phonological rules.[38], [39]

[37] Note that the claim is that it is *often* best to exclude observations, not that linguists have carte blanche to do this. Figuring out when it is best to revise the theory, and when it is best to exclude an observation is itself an empirical endeavor, and so the charge that this approach makes (versions of) grammatical theory unfalsifiable or otherwise empirically unsound, is misplaced. See Dupre (2020) for discussion.

[38] This point is particularly pressing given current Minimalist theories which view all sentential clauses as involving multiple copies of the same argument expression. This view is motivated by the claim that sentential subjects must satisfy two properties: 1. they are assigned a theta-role internally to the verb phrase, and 2. they are assigned case properties in SpecTP position, by a Tense head, outside of the verbal domain. This motivates the movement of the subject argument from within VP to the VP-external TP. However, we only ever pronounce the subject argument once, despite its occurring multiple times in the syntactic structure. So the grammaticality of the sentence must be a product of some phonosyntactic rule governing which copy to pronounce. See Lasnik et al. (2005) for an explanation of, and argument for, these proposals.

[39] Because of this, it will often be a difficult empirical question whether to analyze some utterance as ungrammatical but acceptable, as a result of some extra-linguistic influence, or as reflective of a somewhat complex phonosyntactic rule. For example, one could analyze topicalization as involving a grammatical rule of raising the topicalized argument to the left periphery (“That book, Marta loves”) in combination

With this background out of the way, we can get to the problem with identifying natural language and the language of thought: ungrammatical but acceptable sentences. These are sentences which speakers are able to interpret but which are not licensed by the rules of grammar. The problem is obvious: if speakers can interpret these sentences, i.e. the sentences express an available thought, but they cannot be generated by the language faculty, this suggests that the set of possible thoughts and possible natural language sentences are not even extensionally equivalent, let alone identical. Some thoughts are not expressible in our natural language, and so there must be some medium other than natural language in which they are expressible.

While the objection itself is relatively straightforward, identifying genuinely problematic examples is a little trickier, and turns on various empirical claims about the grammatical properties of the language faculty and the ways in which syntactic structures relate to utterances. Keeping in mind the distinction between the syntactic structure and the utterance form is crucial here, as the existence of a gap between linguistically licensed structures and possible thoughts is only demonstrated if the utterances in question do not correspond to (i.e. are not externalizations of) some underlying legitimate syntactic structure. And these expressions may indeed so correspond even if the way that they are pronounced introduces deviations from that predicted by the grammar. For example, as the above quote from *Aspects* makes clear, disfluencies may be viewed as creating a gap between competence and performance:

13. I... umm, went to the, uh, to the shop.

While the inclusion of filler terms like ‘umm’ and ‘uh’ and repetitions are, of course, not reflected in the grammatical structure of this utterance, deviations from the grammatically predicted sentence of this sort pose no problems for the proposed identification of thought and language. The thought conveyed by such utterances seems to be perfectly captured by a grammatical product of the language faculty. The aspects of the utterance which seem unreflected by this psychological linguistic expression are likewise not found in the thought it expresses, and so there is no need to posit a gap between the natural language expression and the thought. This strategy, of isolating the source of ungrammaticality in extra-linguistic processes will be one of the central ways of defending IH in the face of apparent counter-examples.

This account of disfluencies is similar to the above discussion of the difference between wh-raising and wh-in-situ languages. In both cases, it is claimed that one-and-the-same underlying grammatical structure can be realized by multiple different externalizations. The difference is that in the account of different question-formation operations, the multiplicity of ways of externalizing is explained with reference to specific phono-syntactic rules of which expressions get pronounced. In the account of disfluencies, however, the disparity is a product instead of much less well understood features of general, i.e. extra-linguistic, cognition and performance systems. We might,

Footnote 39 continued
with a phonological rule mandating that the lower copy is unpronounced, or as being an ungrammatical but acceptable result of an extra-linguistic, pragmatic strategy for pronouncing the topicalized expression in a location not licensed by the grammar. The status of a wide range of cases, including echo-questions and right-node raising, depends on which of these strategies is adopted. See Sects. 6.1 and 6.2 for more discussion.

following the distinction between the language faculty in the narrow sense and in the broad sense (Hauser et al. 2002), distinguish between narrow and broad phonology, where the former refers to linguistic rules of pronunciation, while the latter picks out whatever psychological processes are involved in 'translating' a syntactic structure into a public symbol. In general, then, when apparent divergence between linguistic expressions can be attributed to phonological processes, broad or narrow, this does not pose a problem for IH.

Other examples of acceptable but ungrammatical sentences require a different treatment. In many cases, we can recognize that a sentence is ill-formed in some way, but nonetheless understand it. Agreement violations and certain kinds of subcategorization violations provide examples:

14. They is happy.
15. Him told me the time.
16. The child seems sleeping.

In all these cases, we can recognize that something has gone wrong (number agreement, case agreement, and subcategorization requirements, respectively), but we can nevertheless understand what is meant. Calling such expressions 'acceptable' is a stretch, given that they do sound quite wrong. However, the importance of ungrammatical but acceptable expressions, for our purposes, was that they could be interpreted. And even though they sound bad, these expressions clearly meet this criterion. Further, these examples do not seem suitable for a 'phonological' explanation, as given for the examples above. It doesn't seem that in these cases there is a grammatically correct underlying structure which has been modified to produce these strange utterances. Instead it seems that the underlying structure itself is ungrammatical.[40]

Do these examples, then, provide the requisite case demonstrating an extra-linguistic medium for thought? Probably not. Such examples are probably best accounted for by some psycholinguistic 'repair' strategy, which maps these ungrammatical sentences onto corresponding grammatical structures. The literature on such parsing 'repair strategies' is large, but it fairly consistently adopts the view that given the lexico-semantic information made available by word-recognition, the parser is largely able to reconstruct the meanings of expressions despite grammatical deviance.[41] Again, if such approaches are on the right track, these examples pose no problem for the identification of thought and language. Quite the reverse: the parser's being required to produce a grammatical analog of these ungrammatical sentences, in order for the hearer to be able to interpret them, is exactly what would be predicted if the thought conveyed by a sentence was necessarily expressed in natural language.

The real problem for identifying thought and language can be seen by examining cases which seem to involve learning that grammatical constraints can be violated

[40] Of course, the morphology/syntax/phonology interface is highly debated, and it is possible that at least some examples of this sort do involve unusual phonological mappings. For example, one could view Case as a strictly phonological phenomenon, wherein 'him' and 'he' are different pronunciations of the same lexical item as found in different contexts. To the extent that such proposals are correct, these examples can be assimilated to the previous ones.

[41] See the papers in Fodor and Ferreira (2013) for a discussion of such strategies. Arregui et al. (2006) provide a particularly compelling analysis by showing that the degree to which a sentence requires repair (i.e. deviates from licensed grammatical structure) correlates with graded notions of acceptability.

in certain circumstances. Remember that it is crucial in defending this position that languages are neither learned nor variable: that the language faculty, as opposed to the lexicon and principles of externalization, is innate and species universal. However, there do appear to be acceptable linguistic constructions which violate the rules of the grammar, and thus indicate that some thoughts are not expressible as natural language constructions. One paradigm case of such constructions are adicity violations:

17. Ivan sneezed his tooth across the table.[42]
18. Siobhan danced the night away.[43]

The adicity, or argument structure, of verbs is one of the crucial ingredients in determining the grammaticality of structures containing it. That the adicity of an expression is part of the stored information attached to the term is a necessary part of explanations for a wide range of linguistic data. In particular, speaker intuitions about the unacceptability of adicity *violations* is highly robust. All competent speakers of English agree that "Ivan sneezed Siobhan" is unacceptable. This fact can be explained with reference to the shared knowledge that 'sneeze' is an intransitive verb, and thus cannot take a direct object. However, as Goldberg and Jackendoff, and others in the construction grammar tradition, argue, speakers are able to learn, via certain kinds of analogical processes, that there are constructions in which these verbs function differently. If the adicity of these expressions goes into determining which structures are generable by the language faculty, then the ability to interpret these constructions seems to require the ability to have thoughts that are not constructible within the confines of grammatical principles.[44]

A similar kind of worry comes from apparently interpretable violations of syntactic constraints. Consider:

19. This is the department which employs a teacher who speaks every language.[45]

To my ear, this sentence is ambiguous. On one reading, the identified department employs at least one amazingly gifted linguist capable of speaking every human language. On the other, it makes the much more modest, but still impressive, claim that each human language is spoken by at least some teacher in the department, although it

[42] From Goldberg (2006).

[43] From Jackendoff (2018).

[44] As in the other cases, what adicity violations show about the relation between thought and language depends on a variety of empirical assumptions. In particular, the problem raised in this section depends on the assumption that expressions like 'sneeze' and 'dance' are represented in the lexicon with their usual argument structures (i.e. as intranstive). If one relaxed this assumption, and allowed the lexicon to contain multiple argument structures for these expressions, or allowed 'on-the-fly' additions to the lexicon which accounted for sentences like 17 and 18, then these sentences would no longer count as ungrammatical. The motivation for the constructionist approach, however, depends on this assumption. Constructionists view our linguistic abilities with sentences like these to depend on the acquisition of *complex* constructions, into which particular lexical items can be placed, rather than on learning novel argument structures for particular expressions.

[45] Pietroski and Hornstein (2002) discuss related examples, such as "Mary believes that everyone who saw the richest man is happy". That inverted pseudo-clefts, in which the subordinated relative clause is found after the focused object, seem to allow movement out of locations where it is normally prohibited (e.g. relative clauses as in 19) is noted by Chung and McCloskey (1983) and examples of this sort are noted by Phillips (2013a) as problems for the view that island-effects are grammatical phenomena *sensu stricto*.

may be that no teachers speak all of the languages. While this second reading may be a little unnatural, it is, I believe, available, and informal polling has supported this. This poses another problem for the claim that thoughts are expressed in natural language. This is because this reading seems to involve interpreting the quantified noun phrase ‘every language’ as taking wide scope over ‘a teacher’.

On standard grammatical assumptions, going back to May (1985), quantifier scope is determined by relations of C-command after raising the quantifier expressions to the left-periphery of the expression. In order to get this second reading, then, ‘every language’ must be attached to the root of the syntactic structure after ‘a teacher’ is raised. The problem is that there are principled reasons to deny that ‘every language’ can be raised at all. Going back to Ross (1967), relative clauses have been viewed as ‘islands’: expressions which prevent extraction. This ‘relative clause constraint’ explains why the following question is ungrammatical and thus unacceptable:

20. *Which language did the department employ a teacher who speaks?

The attempt to raise the wh-expression ‘which language’ from the relative clause ‘who speaks which language’ results in ungrammaticality and thus unacceptability due to this island constraint. But why then are we able to raise the quantifier expression out of this clause in order to get the reading of sentence 19 wherein ‘every language’ takes wide scope? It seems we must be able to interpret this sentence *despite* the fact that the sentence expressing it (under this reading) violates the grammatical constraints imposed by our grammar. Thus we can think a thought our natural language cannot represent.

An analogy could be made here to learning an artificial language, such as a formal logic. Such languages are not consistent with the principles of natural language grammar, and thus cannot be acquired in the normal way that we acquire a (first) language.[46] They seem instead to be acquired through the use of more general psychological tools of inference, memorization, and extrapolation. Likewise it seems that, alongside the development of our natural language, we can acquire, through these more general learning processes, a variety of additions and exceptions to the grammatical principles governing the construction of natural language expressions. If these additions genuinely increase the set of thoughts that one can have, then the language of thought is not reducible to natural language. I believe the cases just described present the best case against this proposed identification.

6 Solution strategies

While I take the problem of acceptable but ungrammatical expressions to be a serious barrier to identifying thought and language, there are strategies that can be used in the attempt to explain away these difficulties. The first two of these, phonological explanations and repair strategies, we saw in the previous section. These aim to show

[46] Note that this fact is not sufficient on its own to show that natural language is not the language of thought, as it is possible that we learn an artificial language by learning to translate its expressions into expressions in natural language. This proposal will seem quite plausible to anyone who has taught a class in formal logic.

that, while the acceptable utterances may appear to violate the grammatical principles governing the language faculty, they may nonetheless be suitably related to an underlying legitimate structure. As IH says only that thoughts are expressed by the outputs of the language faculty, as long as our interpretation of these utterances is given by these underlying grammatical structures, such expressions do not pose a problem for this proposal. After elaborating on these strategies, I will turn to two more responses to the argument from acceptable but ungrammatical expressions. The first will involve claiming that the 'thoughts' grasped in interpreting these ungrammatical expressions may be quite unlike those grasped when we grasp grammatical thoughts, perhaps on analogy with the representational capacities of non-human animals. The second will draw a distinction between ways in which expressions can be deemed ungrammatical by the language faculty, and argue that only one of these ways poses a problem for IH. If all of the remaining apparent cases of acceptable but ungrammatical expressions are ungrammatical in this way, the hypothesis may be saved.

6.1 Complicate the morphophonology

As I have argued, the central development in linguistic theory which has made the revitalization of IH plausible is the distinction between the core processes of the computational linguistic system and the processes of externalization recruited to publicize the structured representations made available by this system. As several cases above made clear, this distinction is crucial in accounting for apparent linguistic diversity without committing to the claim that the underlying linguistic system, *the natural language* according to the proposal advocated in this paper, itself varies between speakers. The apparent difference between wh-in-situ and wh-movement languages can be accounted for with reference to different externalization strategies, and so the thesis that the syntactic/semantic properties of these languages are identical can be retained. If phonological processing of this sort is indeed peripheral, or subsidiary, to the core operations of language, this is an argument that English and Mandarin speakers really do speak the same 'language', in this technical sense. This thus undermines the argument against IH which claims that languages vary in ways that thoughts do not.

This raises the possibility that the examples of ungrammatical but acceptable expressions claimed to pose a problem for this identification can be handled in a similar way. On such a proposal, learned constructions ought be viewed as acquired conventions about phonology, not syntax. That is, sentences like 17 and 18 could be analyzed as learned 'pronunciations' of grammatically acceptable expressions, such as:

21. Ivan sneezed and thereby caused his tooth to move across the room.
22. Siobhan danced until the night was over.

If 17 and 18 have the same underlying structure as 21 and 22 respectively and differ only in the way this structure is mapped onto an externalized production, then their acceptability can be explained with reference to this structure, which is presumably generable by the language faculty. Thus the problem for IH disappears: no thought is available beyond those made so by the language faculty.

How plausible such a strategy is is a vexed empirical question, depending on complex issues concerning the relation between morphology, syntax, and phonology. Viewing surface forms as derived from apparently very different underlying structures is, however, a very familiar idea in these literatures, within the programs of lexical decomposition (e.g. Wierzbicka 1996; Jackendoff 1996), lexical semantics (e.g. Hale and Keyser 2002; Pustejovsky 1991) and distributed morphology (Halle and Marantz 1994).[47] Sentences like 21 and 22, as candidates for more transparent representations of the structures underlying 17 and 18, are in line with the general thrust of such approaches, which assume that it is only semantically "general" expressions such as 'cause' or 'until' which can be featured in the underlying representations without being pronounced.

Cross-linguistic work such as Dixon (2000) has shown that languages appear to vary in whether they allow causative constructions like 17 and 18, or whether the causality must be encoded with a causative morpheme. Japanese, for example, has a productive morphological rule for transforming a verb (e.g. 'agaru' *to go up* becomes 'ageru' *to raise* i.e. *to cause to go up*). That causality must be marked in surface structure in many languages provides some motivation for viewing causatives like 17 as simply the English strategy for externalizing what is, in its grammatical structure, akin to 21, and a similar cross-linguistic story could be told for resultative constructions like 18. Further work by Papafragou et al. (2002), building on discussions of Talmy (1988), has shown that while the surface properties of languages may vary in the ways they encode things like motion, this does not seem to influence other, non-linguistic, cognitive processes such as recall. This again is compatible with the theory here developed. If the differences between languages are restricted to mappings from thoughts to sensory-motor systems, we would predict that there would be no influence of 'linguistic variation' on other cognitive systems.

6.2 Repair

A closely related strategy is that of *repair*. Instead of viewing such anomalous (from the perspective of grammatical theory) expressions as opaque phonological mappings from underlying grammatical structures to externalized expressions, we can treat such utterances as genuinely ungrammatical (i.e. not products of normal phono-syntactic and phono-morphological rules applied to the products of the language faculty) but posits mechanisms by which they are 'translated into' grammatical expressions which can then be interpreted to give the contents of the thoughts.[48]

Which structures such processes produce is another empirical question, but the structures underlying sentences like 21 and 22 again seem like plausible candidates. This strategy and the previous one may well shade into one another, depending on how parsing (and production) mechanisms relate to the posits of phonological and morphological theory. If these 'performance systems' utilize the rules of the latter

[47] Although see Fodor and Lepore (1999) for a somewhat iconoclastic rejection of all such approaches.

[48] Some suggestive results in Friederici et al. (2006) suggest neurophysiological specializations for responding to ungrammatical expressions. This may be an indication of just the kind of 'repair strategy' advocated in this section.

theories, as argued for in Phillips (2004, 2013b) and Phillips et al. (2011), then there may be no difference: 'repair' would just amount to the application of such rules to public symbols so as to reproduce the grammatical structures from which they derive. If, on the other hand, parsing mechanisms utilize quite different strategies, perhaps the heuristics of Ferreira and Patson (2007), then there will be a clean divide between them.

While I believe these two strategies are plausibly the best bet for defenders of IH, there are serious obstacles to application of these strategies. The most serious is the *overgeneration* problem. Positing a repair strategy or phonological process by which apparently ungrammatical utterances can be mapped onto grammatical underlying structures is liable to overgenerate, and predict that expressions which are in fact unacceptable would be legitimized by these very processes. That is, one must ensure that any proposed strategies for exacting the mapping from acceptable but ungrammatical expressions to grammatical structures does not also suffice to map unacceptable expressions to grammatical structures.

For example, one possible strategy for 'repairing' adicity violations such as sentences 17 and 18 would be that rather than identifying the (usual) argument structure of the identified verb, and thus precluding the generation of a structure with the intended number of arguments, the parser first identifies the intended argument structure and creates a 'verbal skeleton', which has the right argument structure but with a dummy variable where the verb should be:

23. Ivan *V* his tooth across the table.
24. Siobhan *V NP* away.

These are perfectly normal grammatical structures (cf. "Ivan pushed his tooth across the table" and "Siobhan gave her money away"). The offending expressions, which can't typically be found in structures of this sort, can then be late-inserted and coerced into taking on the transitive meanings intended.

The difficulty with this is that it is unclear how to prevent overgeneration, predicting that sentences which are in fact unacceptable could be salvaged in these ways.[49] For example, Jackendoff (1997) describes a wide range of constructions closely analogous to sentence 18 that seem semantically plausible but which are nonetheless unacceptable. Consider, for example:

25. *Siobhan danced the Tango the night away.
26. *Siobhan danced happily the night away.

Likewise:

27. *Ivan sneezed violently his tooth across the table.

In all these cases, we can understand what these sentences *would* mean, but they are clearly bad. The difficulty then is explaining why such a repair strategy cannot likewise be used to salvage these expressions. If 'dance' and 'sneeze' can be treated as transitive verbs, why is this impossible for 'dance the Tango', 'dance happily' and 'sneeze violently'? Of course, there are things one can say about such constraints. It

[49] It is also inconsistent with standard Minimalist assumptions that there is no grammatical structure beyond the Merging of lexical items.

appears that these constructions allow this sort of coercion to apply only to verbal heads, not to VPs. This fact must itself be explained however: if pragmatic strategies allow for the mapping of ungrammatical expressions onto grammatical structures in the case of 17 and 18, why can similar processes not apply to 25–27?

It is worth noting that accounting for these examples in this way would not merely defend IH from this objection, but would provide positive support for it. In line with Hinzen's arguments discussed above, if interpretation of an utterance requires that we 'translate' it, mapping it onto a grammatical expression, this reinforces the idea that there is a one-to-one mapping between possible thoughts and possible grammatical expressions, as predicted by IH. It is not always noted that the assumption that ungrammatical sentences *must* be repaired in order to be understood itself requires explanation and justification. If we can grasp thoughts which are not expressed by grammatical sentences, then we might expect some ungrammatical sentences to be understood 'directly', i.e. by mapping them onto the LOT (where this is assumed to be distinct from natural languages) without repairing them. If we discover that such repairs are indeed always required, this provides strong evidence that understanding a sentence, i.e. grasping the thought it expresses, simply is constructing a natural language expression.

Examples like 19 may present even more serious overgeneration worries. Any strategy which loosens the constraint on extraction from relative clauses so as to allow for the ambiguity of 19 must not thereby predict that 20 is acceptable. One could propose that some repair strategy enables us to loosen the locality constraints on quantifier raising (but crucially not on wh-movement). However, this proposal similarly overgenerates:

28. It was a man who told me that every philosopher loves Frege.

Despite featuring a cleft construction which appears to avoid the constraints on movement in the cases above, this sentence is not ambiguous. It cannot be read as claiming that every philosopher is such that a man (read specifically or non-specifically) told me that *they* loved Frege. That is, one cannot read the embedded 'every philosopher' as scoping out of the that-clause and over the focused matrix subject 'a man'. Proponents of the repair strategy must not posit repair mechanisms for 19 which also predict scope ambiguities in 28.

6.3 Different kinds of thought

As Hinzen (2013, Sect. 7) points out, claiming that human thought occurs in natural language does not preclude the possibility that non-human (and thus non-linguistic) animals engage in some forms of 'thought'. The crucial idea behind IH is that human thought, as expressed by the structures of natural language, forms a natural psychological kind. This is consistent with there being many other kinds of psychological representation. Indeed, it is clear that the representational formats of large parts of cognition, such as vision (see e.g. Palmer 1999) or map-like locational representation (e.g. Camp 2007) are quite unlike expressions of natural language. One option for defending IH, then, is allowing that we can interpret ungrammatical sentences, and that we do so without mapping them onto grammatical structures, but that this involves a quite different kind of cognition than that used when we understand grammatical sentences. On this proposal, identifying thought and language is a kind of 'explication' in something like the sense of Carnap (1962): natural language expressions are the vehicles for a substantial amount of what is pretheoretically called 'thought', the class of natural language expressions forms a natural kind, and this class includes many of the 'core cases' of our pretheoretic notion, and thus we are justified in revising our conception of thought in line with this hypothesis.

It is, I assume, an empirical possibility that the everyday notion of 'thought' does not pick out a uniform psychological phenomenon (beyond the heterogeneity usually assumed by this term in philosophical discussions, which apply it to distinct psychological kinds like belief, desire, intention, etc.). There are several options we could take in response to such a discovery. One would be eliminativist, eschewing talk of thoughts altogether. Of course this would preclude IH. The proposal just sketched, however, would instead select some subset of the things we antecedently viewed as thoughts, and treat that as the extension of our new, scientifically useful, concept. If, for example, sentence interpretation in general turned out to centrally involve the construction of a syntactic structure in line with the constraints of Universal Grammar, but that in certain rare cases, when such a strategy was unavailable, interpreters resorted to the construction of a different kind of psychological structure, it may be defensible to hold onto IH by viewing only the former as the extension of the new, explicated, notion of 'thought'. Of course, how plausible/appropriate this linguistic maneuver is will depend on how much, and in particular how many of the 'core cases', of our traditional notion of 'thought' is covered by these linguistic structures. If non-linguistic interpretations are common and paradigmatic instances of thought, this explication will amount to little more than a stipulation of the truth of IH. If, however, exceptions are quirky and unusual, IH could be viewed as a genuine kind identity, and explication would thus be useful. Relatedly, if work on animal cognition suggests a close correspondence between animal and human cognition, this would pose a problem for such an explication of 'thought', as non-human animal thought, we are assuming, is not structured linguistically.

As above, this strategy is more plausible for some phenomena than others. In particular, it is often thought that non-linguistic representational media are very bad at expressing certain sorts of 'logical' content. Quantifiers, negation, disjunction, etc. seem difficult to express without language. This suggests that such an approach is unlikely to be of much use in handling the recalcitrant sentence 19.

6.4 Filters versus ungenerable expressions

The final possible response involves distinguishing between two ways in which an expression can be deemed 'ungrammatical'. Some expressions are ungrammatical on account of not being generable at all by the language faculty, whereas others are ungrammatical in virtue of violating some constraint on what the outputs of the language faculty must be like. This distinction largely originated with Chomsky and Lasnik (1977), and was then incorporated as one of the main features of Government and Binding Theory (Chomsky 1981). Traditional GB theory included general constraints on what kinds of structures could be produced (centrally, X-bar Theory and the sole transformational rule 'move α'), as well as a collection of 'filters', which served to further limit the set of acceptable expressions by excluding those structures which, though generated in a perfectly legitimate way, had some illegitimate property. Perhaps the most famous example of the latter is the Case Filter, which states that all overt NPs must be assigned Case. This constraint on grammaticality is posited in order to account for, *inter alia*, cases like the following:

29. *(It) seems the child to be sleeping.
30. The child seems to be sleeping.

For various reasons[50], it is important that grammatical theory not preclude the language faculty from generating structures like 29. However, it is clearly unacceptable. The Case Filter provides an explanation for this. Infinitive verbs ('to be sleeping') do not assign Case properties to their arguments. As 'the child' is the overt subject of this expression in 29, the Case Filter rules it out. This further explains why we find sentences like 30 in English. Because the embedded verb can't assign Case to the NP, the NP must move to the higher, tensed, verb, which can. This movement occurs even though 'the child' is the semantic argument of 'sleeping', and *not* of 'seems'. In 30, then, the overt NP is assigned Case, and so the filter is not violated and the sentence is grammatical, and thus acceptable.

The crucial point about this for our purposes is that, despite our ability to recognize sentence 29 as unacceptable, we do know what it means. This fact can be accounted for within the confines of IH if it is allowed that thoughts can be expressions of the language faculty, *even if* these expressions are marked as ungrammatical. On this view, we can have thoughts which are themselves ill formed, just as we can produce expressions which are ill formed. What is ruled out is having thoughts which the language faculty cannot even produce. We can thus distinguish two kinds of ungrammaticality. One kind involves the production of a full-fledged syntactic structure which is somehow 'marked' as ungrammatical.[51] The other involves a complete failure to even produce a structure. It seems that ungrammatical but acceptable sentences of the latter sort pose a deeper worry for the proposed identification of thought and language. If we can understand an utterance which cannot even be generated by the language faculty, it seems there must be some extra-linguistic medium which can serve as a vehicle for thought. But if the examples are ungrammatical in the former way, the language faculty will produce a vehicle for such thoughts, although it will indicate that this vehicle is in some way ill formed.

I have so far stated this response in the terms of GB theory. Doing so in the terms of the contemporary Minimalist Program is slightly more complicated. 'Filters' in this program have largely been replaced by 'interface conditions', demands imposed on the outputs of the language faculty which ensure that they are 'legible' by the semantic and phonological systems which are used to interpret and externalize the products of the language faculty. Standard Minimalist accounts of movement treat it as arising out of the need to remove uninterpretable features, i.e. properties of lexical items which

[50] Centrally, that generation of sentence 30 requires generation of 29 as an earlier stage in its derivation.

[51] This kind of ungrammaticality fits together rather nicely with the account of ungrammaticality as an 'error-signal' proposed in Gross (forthcoming).

result in failures at the interfaces. Case, for example, is viewed as an uninterpretable feature of arguments (DP or NP, depending on the theory) which gets deleted when the argument expression is in a local relation to a Tense expression. This will motivate movement when arguments originate within verb phrases (see fn. 38) which cannot perform this function of eliminating uninterpretable features.[52] Similar accounts are given of other agreement phenomena such as gender and number.[53]

Whether this Minimalist re-interpretation of what were traditionally viewed as filters can re-instate the traditional distinction between two forms of ungrammaticality depends on how we interpret the claim that unchecked/undeleted features are *uninterpretable* by the interfaces. At face value, this would seem to undermine the proposal that such cases of ungrammaticality involve merely 'marking' these structures as deficient in some way. If they are literally uninterpretable, especially by the semantic interface, then it seems that we cannot rescue IH by saying that such sentences can be interpreted despite being ungrammatical. However, it is not clear that one should read this term so strongly.[54] Sentences like 29 *are* interpretable, in the everyday sense of the term, despite being recognized as ill formed. I thus suggest that we ought read 'uninterpretability' at the interfaces as exactly in line with the account of traditional filter violations given above: uninterpretable expressions can be assigned meanings, and so can serve as vehicles for thought, although they are marked as grammatically defective.[55]

On the other side of this distinction between kinds of grammaticality, some expressions do seem to be genuinely ruled out by Minimalist accounts of the language faculty; not merely in that they are 'uninterpretable', in the sense just identified, but that they cannot even be constructed. Economy constraints, principles governing the workings of the language faculty which ensure that its operations are maximally computationally efficient, seem to operate in this way. The Subjacency Constraint, which provides a limit on the distance (defined structurally) an expression can be moved by a single operation, is motivated on the grounds that allowing the system to perform long-distance movement would create too substantial a computational cost.[56] Such a proposal is involved in explaining the difference between the following expressions:

[52] The extent to which arguments must undergo such movement is controversial, with extreme positions like that of Laenzlinger and Soare (2004) insisting that all arguments must be moved from within the VP, and more moderate positions suggesting that only subjects must be moved. However, it is very widely agreed that there must be some movement of arguments, and so the problem of which copy is pronounced arises.

[53] Pesetsky and Torrego (2007) provide an analysis of Case and other uninterpretable features along roughly these lines.

[54] And likewise for claims that uninterpretable features cause 'crashes' at the interfaces.

[55] Yet another empirical wrinkle in all of this is the recent program of 'Crash-Proof Syntax', which aims to show that the language faculty *never* produces uninterpretable expressions (see Frampton and Gutmann 2002). If such a proposal is correct, then the distinction between kinds of grammaticality will not be available, as no ungrammatical expressions will be generated.

[56] The rough idea is that the syntactic structure is constructed bottom-up, and movement from 'lower down' in the tree is motivated by the need to check features higher up. If long-distance movement were allowed, the construction would have to retain access to all lower-down structure, in case some element anywhere in the structure was a suitable target for such feature-driven movement. By insisting that all movement proceed in short steps, the construction of the structure can 'forget' about all structure except that in the immediate vicinity of the most recently added element. This drastically reduces the computational cost of constructing a syntactic representation. See Chomsky (2008) for a discussion of this framework.

31. What did [$_{\mathrm{IP}}$ Rahim claim [$_{\mathrm{CP}}$ that [$_{\mathrm{IP}}$ he read]]]?
32. *What did [$_{\mathrm{IP}}$ Rahim believe [$_{\mathrm{NP}}$ the claim [$_{\mathrm{CP}}$ that [$_{\mathrm{IP}}$ he read]]] ?

If movement is restricted so that it can cross at most one ‘bounding node’ (in English, IP or NP)[57] at a time, we can explain the above pattern. In 31, there are two IPs that must be crossed, but the wh-expression can make this movement in two steps, each of which crosses only one. ‘What’ can first move to the specifier of the embedded CP, as marked by the intermediate struckout ‘what’, and then to the sentence-initial position in which it is pronounced, and so no ‘long-distance’ movement is needed. However, in 32, while the wh-expression can legitimately move initially to the embedded CP, from there it must move to the sentence-initial position. But to do so would involve crossing two bounding nodes (the embedded NP and the matrix IP), and there is no intermediate ‘landing site’ which could be used to break up this journey. This thus explains the acceptability of 31 and the unacceptability of the otherwise quite similar 32.

Subjacency, and other constraints resulting from economy considerations, are ungrammatical in a stronger sense than violations of filters/interface conditions. It is a crucial part of the explanatory strategy of the Minimalist Program that such constraints prevent representations which would violate them from being constructed in the first place.[58] This is reflected in the more extreme unacceptability responses they generate. Whereas one can assign a meaning to 29 despite its obvious unacceptability, sentences like 32 are typically genuinely uninterpretable (in the non-technical sense). This fact provides further motivation for IH, as identifying thought and language enables us to explain the distribution of acceptability in ungrammatical sentences. Ungrammatical but acceptable sentences may involve the generation of structures which are subsequently marked as ungrammatical. Despite this marking, the fact that the structures are produced provides a possible vehicle for thought. However, when a sentence is ungrammatical in virtue of not even being generable, there is no vehicle for thought present and thus the sentence cannot be interpreted at all.

However, when sentences which are predicted to not even be generable are nonetheless interpretable, this poses a particularly deep problem for the proposed identification. The reading of sentence 19 where ‘every language’ takes wide scope appears to be of this sort. The constraint on extraction from relative clauses is typically analyzed as a subjacency violation.[59] This suggests that the interpretability of expressions like this, under readings which violate the subjacency constraint, are seriously problematic for IH, as the thoughts they seem to convey are not even producible by the language faculty. On the other hand, sentence 19 is somewhat marginal, so we may

[57] For the purposes of this paper, I am assuming that arguments are NPs rather than DPs. This is for simplicity and for consistency with discussions of historical statements of rules like the Case Filter. Nothing should turn on this issue. For the origin of the idea that arguments are headed by determiners, not nouns, see Abney (1987).

[58] The gains in computational efficiency allowed by economy constraints rely on the claim that the computationally inefficient operations are never even attempted, not just that when they are they lead to uninterpretability at the interface.

[59] In sentence 21, [who speaks which language] is an IP and [a teacher who speaks which language] is an NP, and both must be crossed. The Specifier of the CP intervening between these expressions could serve as a landing site for an intermediate movement, except that the subject of the embedded IP ‘who’ is already occupying this space, as a result of raising for Case assignment.

not wish to place too heavy an argumentative burden on sentences of this sort, which would be good news for IH.

While I believe these strategies are reasonably exhaustive, I don’t wish to commit to the claim that there are no other options for handling acceptable but ungrammatical expressions from the perspective of IH. The aim of this section is just to stress the ways in which a defense of this proposal relies on the outcome of several ongoing debates in the linguistics literature. If IH is to be successfully defended, I believe it will involve a combination of at least all of the strategies covered in this section.

7 Conclusion

In this paper I have hoped to show that the Identification Hypothesis is not as implausible as is often thought. In particular, once it is recognized that the relata in the purported identification relation are the language of thought and human I-language, many of the traditional problems with the view disappear. I also raised a novel obstacle to defending this proposal, acceptable but ungrammatical expressions, and pointed to several strategies for responding to worries of this sort. It is, at least, an open empirical question whether natural language, as identified as the target of scientific linguistic theory, differs from thought in the ways such arguments assume.

References

Abney, S. P. (1987). The English noun phrase in its sentential aspect. Ph.D. thesis, MIT.

Arregui, A., Clifton, C, Jr., Frazier, L., & Moulton, K. (2006). Processing elided verb phrases with flawed antecedents: The recycling hypothesis. *Journal of Memory and Language*, *55*(2), 232–246.

Baggio, G. (2018). *Meaning in the brain*. Cambridge, MA: MIT Press.

Berwick, R. C., & Chomsky, N. (2015). *Why only us: Language and evolution*. Cambridge, MA: MIT Press.

Boeckx, C. (2010). What principles and parameters got wrong. In C. Picallo (Ed.), *Linguistic variation in the minimalist framework* (pp. 155–178). Oxford: Oxford University Press.

Boeckx, C. (2014). *Elementary syntactic structures: Prospects of a feature-free syntax*. Cambridge: Cambridge University Press.

Boeckx, C., & Leivada, E. (2013). Entangled parametric hierarchies: Problems for an overspecified universal grammar. *PLoS ONE*, *8*(9), e72357.

Bornkessel, I., Zysset, S., Friederici, A. D., Von Cramon, D. Y., & Schlesewsky, M. (2005). Who did what to whom? The neural basis of argument hierarchies during language comprehension. *Neuroimage*, *26*(1), 221–233.

Burge, T. (2014). Perception: Where mind begins. *Philosophy*, *89*(3), 385–403.

Burzio, L. (1986). *Italian syntax: A government-binding approach*. Berlin: Springer.

Camp, E. (2007). Thinking with maps. *Philosophical Perspectives*, *21*, 145–182.

Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language*, *3*(4), 572–582.

Carnap, R. (1962). *Logical foundations of probability*. Chicago: University of Chicago Press.

Carruthers, P. (1998). *Language, thought and consciousness: An essay in philosophical psychology*. Cambridge: Cambridge University Press.

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, *25*(6), 657–674.

Cheng, L. L.-S. (2003). Wh-in-situ. *GLOT International*, *7*, 129–137.

Cheng, L. L.-S. (2009). Wh-in-situ, from the 1980s to now. *Language and Linguistics Compass*, *3*(3), 767–791.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: The MIT Press.

Chomsky, N. (1981). *Lectures on government and binding: The Pisa lectures*. Dordrecht: Foris.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Westport: Greenwood Publishing Group.

Chomsky, N. (1993). *Language and thought*. Wakefield: Moyer Bell.

Chomsky, N. (2007a). Biolinguistic explorations: Design, development, evolution. *International Journal of Philosophical Studies*, *15*(1), 1–21.

Chomsky, N. (2007b). Of minds and language. *Biolinguistics*, *1*, 009–027.

Chomsky, N. (2008). On phases. In R. Freidin, C. P. Otero, & M. L. Zubizarreta (Eds.), *Foundational issues in linguistics theory* (pp. 133–166). Cambridge, MA: MIT Press.

Chomsky, N. (2015). *What kind of creatures are we?* New York: Columbia University Press.

Chomsky, N. (2017). The language capacity: Architecture and evolution. *Psychonomic Bulletin & Review*, *24*(1), 200–203.

Chomsky, N., & Lasnik, H. (1977). Filters and control. *Linguistic Inquiry*, *8*(3), 425–504.

Chung, S., & McCloskey, J. (1983). On the interpretation of certain island facts in GPSG. *Linguistic Inquiry*, *14*(4), 704–713.

Churchland, P. M. (1996). *The engine of reason, the seat of the soul: A philosophical journey into the brain*. Cambridge, MA: MIT Press.

Cinque, G., & Kayne, R. S. (2005). *The Oxford handbook of comparative syntax*. Oxford: Oxford University Press.

Clark, A. (1998). Magic words: How language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 162–183). Cambridge: Cambridge University Press.

Collins, J. (2011). Impossible words again: Or why beds break but not make. *Mind & Language*, *26*(2), 234–260.

Culicover, P. W., & Jackendoff, R. (2006). The simpler syntax hypothesis. *Trends in Cognitive Sciences*, *10*(9), 413–418.

Davidson, D. (1975). Thought and talk. In S. Guttenplan (Ed.), *Mind and language* (pp. 7–23). Oxford: Clarendon Press.

Dennett, D. C. (1991). Two contrasts: Folk craft versus folk science, and belief versus opinion. In J. D. Greenwood (Ed.), *The future of folk psychology: Intentionality and cognitive science* (pp. 135–148). New York: Cambridge University Press.

Dixon, R. M. W. (2000). A typology of causatives: Form, syntax and meaning. In R. M. W. Dixon & A. Y. Aikhenvald (Eds.), *Changing valency: Case studies in transitivity* (pp. 30–83). Cambridge: Cambridge University Press.

Dummett, M. (1991). The relative priority of thought and language. In M. Dummett (Ed.), *Frege and other philosophers* (pp. 315–24). Oxford: Oxford University Press.

Dupre, G. (2020). Linguistics and the explanatory economy. *Synthese*. https://doi.org/10.1007/s11229-019-02290-x.

Ferreira, F., & Lowder, M. W. (2016). Prediction, information structure, and good-enough language processing. *Psychology of Learning and Motivation*, *65*, 217–247.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83.

Field, H. (1974). Theory change and the indeterminacy of reference. *The Journal of Philosophy*, *70*(14), 462–481.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (1990). *A theory of content and other essays*. Cambridge, MA: The MIT Press.

Fodor, J. A. (2001). Language, thought and compositionality. *Mind & Language*, *16*(1), 1–15.

Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford: Oxford University Press.

Fodor, J., & Ferreira, F. (2013). *Reanalysis in sentence processing*. Berlin: Springer.

Fodor, J., & Lepore, E. (1999). Impossible words? *Linguistic Inquiry*, *30*(3), 445–453.

Frampton, J., & Gutmann, S. (2002). Crash-proof syntax. In S. D. Epstein & T. D. Seely (Eds.), *Derivation and explanation in the minimalist program* (pp. 90–105). Hoboken: Wiley.

Friederici, A. D. (1981). Production and comprehension of prepositions in aphasia. *Neuropsychologia*, *19*(2), 191–199.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78–84.

Friederici, A. D., Fiebach, C. J., Schlesewsky, M., Bornkessel, I. D., & Von Cramon, D. Y. (2006). Processing linguistic complexity and grammaticality in the left frontal cortex. *Cerebral Cortex*, *16*(12), 1709–1717.

Friedman, J. (2013). Question-directed attitudes. *Philosophical Perspectives*, *27*(1), 145–174.

Gleitman, L., & Papafragou, A. (2005). Language and thought. In K. J. Holyoak & R. G. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 633–661). Cambridge: Cambridge University Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Gross, S. (forthcoming). Linguistic intuitions: Error signals and the voice of competence. In S. Schindler (Ed.), *Linguistic intuitions, evidence, and expertise*. Oxford: Oxford University Press.

Grzankowski, A. (2015). Not all attitudes are propositional. *European Journal of Philosophy*, *23*(3), 374–391.

Hale, K., & Keyser, S. J. (2002). *Prolegomenon to a theory of argument structure*. Cambridge, MA: MIT Press.

Halle, M., & Marantz, A. (1994). Some key features of distributed morphology. *MIT Working Papers in Linguistics*, *21*, 275–288.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *science*, *298*(5598), 1569–1579.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393.

Hinzen, W. (2006). *Mind design and minimal syntax*. Oxford: Oxford University Press.

Hinzen, W. (2011). Language and thought. In C. Boeckx (Ed.), *The Oxford handbook of linguistic minimalism* (pp. 499–522). Oxford: Oxford University Press.

Hinzen, W. (2013). Narrow syntax and the language of thought. *Philosophical Psychology*, *26*(1), 1–23.

Hinzen, W. (2014). What is un-cartesian linguistics? *Biolinguistics*, *8*, 226–257.

Hinzen, W. (2015). Nothing is hidden: Contextualism and the grammar-meaning interface. *Mind & Language*, *30*(3), 259–291.

Hinzen, W. (2017). Reference across pathologies: A new linguistic lens on disorders of thought. *Theoretical Linguistics*, *43*(3–4), 169–232.

Hornstein, N., Nunes, J., & Grohmann, K. K. (2005). *Understanding minimalism*. Cambridge: Cambridge University Press.

Hurley, S., & Nudds, M. (2006). *Rational animals?* Oxford: Oxford University Press.

Huybregts, M. R. (2017). Phonemic clicks and the mapping asymmetry: How language emerged and speech developed. *Neuroscience & Biobehavioral Reviews*, *81*, 279–294.

Jackendoff, R. (1996). Conceptual semantics and cognitive linguistics. *Cognitive Linguistics*, *7*(1), 93–129.

Jackendoff, R. (1997). Twistin'the night away. *Language*, *73*, 534–559.

Jackendoff, R. (2018). Representations and rules in language. In B. Huebner (Ed.), *The philosophy of Daniel Denett*. Oxford: Oxford University Press.

Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry, & H. Wettstein (Eds.), *Themes from Kaplan* (pp. 481–563). Oxford: Oxford University Press.

Kaplan, D. (1990). Words. *Aristotelian Society Supplementary*, *64*(1), 93–119.

Kaye, L. J. (1995). The languages of thought. *Philosophy of Science*, *62*(1), 92–110.

Kayne, R. S. (1994). *The antisymmetry of syntax*. Cambridge, MA: MIT Press.

Kripke, S. A. (1979). A puzzle about belief. In A. Margalit (Ed.), *Meaning and use* (pp. 239–283). Berlin: Springer.

Laenzlinger, C., & Soare, G. (2004). On merging positions for arguments and adverbs in the romance mittelfeld. In L. Brugè, G. Giusti, N. Munaro, W. Schweikert, & G. Turano (Eds.), *Contributions to the thirtieth Incontro di Grammatica Generativa* (pp. 105–128). Venice: Libreria Editrice Cafoscarina.

Lasnik, H., Uriagereka, J., & Boeckx, C. (2005). *A course in minimalist syntax*. Hoboken: Wiley-Blackwell.

Levinson, S. C., Kita, S., Haun, D. B., & Rasch, B. H. (2002). Returning the tables: Language affects spatial reasoning. *Cognition*, *84*(2), 155–188.

Lewis, D. (1975). Languages and language. In K. Gunderson (Ed.), *Minnesota studies in the philosophy of science* (pp. 3–35). Minneapolis: University of Minnesota Press.

Li, P., & Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. *Cognition*, *83*(3), 265–294.

May, R. (1985). *Logical form: Its structure and derivation*. Cambridge, MA: MIT Press.

Murray, S. E., & Starr, W. B. (2018). Force and conversational states. In D. Fogal, D. Harris, & M. Moss (Eds.), *New work on speech acts* (pp. 202–236). Oxford: Oxford University Press.

Newmeyer, F. J. (2005). *Possible and probable languages: A generative perspective on linguistic typology*. Oxford: Oxford University Press.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Papafragou, A., Massey, C., & Gleitman, L. (2002). Shake, rattle,'n'roll: The representation of motion in language and cognition. *Cognition*, *84*(2), 189–219.

Pesetsky, D., & Torrego, E. (2007). The syntax of valuation and the interpretability of features. In S. Karimi, V. Samiian, & W. K. Wilkins (Eds.), *Phrasal and clausal architecture: Syntactic derivation and interpretation* (pp. 262–294). Amsterdam: John Benjamins Publishing.

Phillips, C. (2004). Linguistics and linking problems. In S. F. Warren & M. Rice (Eds.), *Developmental language disorders: From phenotypes to etiologies* (pp. 241–287). New York: Lawrence Erlbaum Associates.

Phillips, C. (2013a). On the nature of island constraints I: Language processing and reductionist accounts. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects* (pp. 64–108). Cambridge: Cambridge University Press.

Phillips, C. (2013b). Parser-grammar relations: We don't understand everything twice. In M. Sanz, I. Laka, & M. K. Tanenhaus (Eds.), *Language down the garden path: The cognitive and biological basis for linguistic structures* (pp. 294–315). Cambridge: Cambridge University Press.

Phillips, C., Wagers, M. W., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. In J. T. Runner (Ed.), *Experiments at the interfaces* (pp. 147–180). Leiden: Brill.

Pietroski, P., & Hornstein, N. (2002). Does every sentence like this exhibit a scope ambiguity? In W. Hinzen & H. Rott (Eds.), *Belief and Meaning: Essays at the Interface*. Deutsche Bibliotek der Wissenschaften (Frankfurt: Haensel-Hoehenhausen).

Pinker, S. (1994). *The language instinct: The new science of language and mind*. New York: William Morrow.

Porot, N. J. (2019). Some Non-human languages of thought. Ph.D. thesis, CUNY.

Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics*, *17*(4), 409–441.

Ross, J. R. (1967). *Constraints on variables in syntax*. Ph.D. thesis, MIT. [Published 1986 as *Infinite Syntax!*, Norwood: Ablex].

Saur, D., Kreher, B. W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.-S., et al. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences*, *105*(46), 18035–18040.

Stainton, R. J. (1996). *Philosophical perspectives on language*. Peterborough: Broadview Press.

Stainton, R. J. (2011). In defense of public languages. *Linguistics and Philosophy*, *34*(5), 479–488.

Starr, W. B. (2011). A preference semantics for imperatives. Ms., Cornell University.

Stich, S. P. (1978). Beliefs and subdoxastic states. *Philosophy of Science*, *45*(4), 499–518.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100.

Tonhauser, J., & Matthewson, L. (2015). Empirical evidence in research on meaning. Ms., The Ohio State University and University of British Columbia.

Trotzke, A., Bader, M., & Frazier, L. (2013). Third factors and the performance interface in language design. *Biolinguistics*, *7*, 1–34.

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *92*(7), 345–381.

Wierzbicka, A. (1996). *Semantics: Primes and universals: Primes and universals*. Oxford: Oxford University Press.

Wittgenstein, L. (1959/2009). *Philosophical investigations*. Rev. 4th edition by P.M.S Hacker & J. Schulte. Hoboken: Wiley-Blackwell.

Yang, C., Crain, S., Berwick, R. C., Chomsky, N., & Bolhuis, J. J. (2017). The growth of language: Universal Grammar, experience, and principles of computation. *Neuroscience & Biobehavioral Reviews*, *81*, 103–119.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.