# Ontology of language, with applications to demographic data

S. Clint Dowland[a,b,*], Barry Smith[c], Matthew A. Diller[b], Jobst Landgrebe[c], and William R. Hogan[b]
[a] *Department of Biomedical Informatics, University of Arkansas for Medical Sciences, AR, USA*
*E-mail: clintdowland@gmail.com*
[b] *Department of Health Outcomes and Biomedical Informatics, University of Florida, FL, USA*
*E-mails: diller17@ufl.edu, hoganwr@gmail.com*
[c] *Department of Philosophy, State University of New York at Buffalo, NY, USA*
*E-mails: ifomis@gmail.com, jobstlan@buffalo.edu*

**Abstract**. Here we present what we believe is a novel account of what languages are, along with an axiomatically rich representation of languages and language-related data that is based on this account. We propose an account of languages as aggregates of dispositions distributed across aggregates of persons, and in doing so we address linguistic competences and the processes that realize them. This paves the way for representing additional types of language-related entities. Like demographic data of other sorts, data about languages may be of use to researchers in a number of areas, including biomedical research. Data on the languages used in clinical encounters are typically included in medical records, and capture an important factor in patient-provider interactions. Like many types of patient and demographic data, data on a person's preferred and primary languages are organized in different ways by different systems. This can be a barrier to data integration. We believe that a robust framework for representing language in general and preferred and primary language in particular—which has been lacking in ontologies thus far—can promote more successful integration of language-related data from disparate data sources.

Keywords. Language, linguistic competence, preferred language, primary language, Basic Formal Ontology (BFO)

## 1. Introduction

Here we present an ontological account of language and of related entities. We include a corresponding formal representation of what a language is that has been developed using classes and relations defined in publicly accessible ontologies. The goal of this work is to simplify the representation of language-related entities to support research on how languages are used and how language use is related to other demographic factors. We focus particularly on the health care domain, though the results of this work can be extended for example to the study of data relating to how different communities cope with language barriers associated with filling in government and other official forms in all areas.

### *1.1. Language data and health care*

Like data on race, ethnicity, and gender, data on the languages that people use fall into the category of demographic data. Where one researcher might find useful results derived from comparing how some factor differs across distinct ethnic groups, another might find useful results from a similar comparison across distinct linguistic communities. Moreover, there may be some overlap between data pertaining to

*Corresponding author. E-mail: clintdowland@gmail.com.

language use and demographic data in other categories. For example, sharing a language and sharing a linguistic heritage exemplify the sorts of cultural ties that can form the basis of ethnic groupings. In this sense, linguistic data may be of broad importance for research.

One area in which linguistic data are of importance is biomedical research. For example, data on the preferred language of a patient are among the data often gathered during a health care encounter and recorded in electronic health records (EHRs). These data capture something important about the communication between a patient and a health care provider in a setting in which miscommunications—such as might arise from a language barrier—could lead to negative outcomes.

The importance of language barriers extends beyond the context of health care encounters: in the United States, Low English Proficiency is regarded as a category of social determinants of health (SDoH); and it has been argued that it may in some cases exacerbate disparities that initially arise due to factors that fall under other SDoH categories.[1] Thus researchers may be concerned not only with patients' preferred languages, but also with a person's primary language or the collection of languages the person or the person's family use at home.

### *1.2. Interoperability, ontologies, and language data*

The integration of biomedical data enables researchers to make use of data derived from sources that use differing data models. Such integration is enhanced if the data are semantically interoperable, so that a computer algorithm designed to draw inferences from one data set can be applied to some other data set without requiring modifications to either the algorithm or the data to which it is applied. Semantic interoperability can be achieved by providing the data with computable semantics.

The use of a common data model (CDM) is commonly viewed as a means to address the lack of computable semantics of the sort needed for accurate and efficient data integration. However, Brochhausen et al. (2018) show that CDMs do not suffice to achieve that goal, for reasons which can be addressed to a degree through the use of axiomatically rich ontologies.[2]

Such ontologies enable the transformation of data as encoded according to some data model into a form that is suitable for use in a referent-tracking system.[3] A key advantage of such a system is that it represents not only the data themselves but also the entities the data are about. In this way data model conflicts, overlaps, and redundancies can be accounted for. Ontologies provide consensus terms representing the classes or types under which the tracked entities fall, terms that are then used to annotate the identifiers for the corresponding specific instances (patients, documents, encounters, and so forth). An ontology can also provide expressions representing the relationships among the tracked entities. If the ontology is axiomatically rich, then many of the class terms are equipped with axioms that represent how entities of a given type relate to entities of some other types (for example how a language relates to a speaker of the language or to the associated linguistic community). The axioms enable automatic reasoners to make use of the referent tracking data enhanced with ontologies in performing tasks such as returning query results, drawing inferences, and discovering inconsistencies in the underlying source data.

If we are to use ontologies to characterize the entities that language data are about, we must of course consider which entities we need to represent in order to reflect accurately what is captured, for example, in data about a patient's preferred language. And if we have data to the effect that *John's preferred language is English*, then the English language will be among the entities that the relevant data are *about*.[4] For this reason, part of developing an approach to modeling data about preferred languages is to model

[1] See Cohen et al. (2005); Eneriz-Wiemer et al. (2014); and Espinoza and Derrington (2021) for discussion of such issues.

[2] See also Blaisure and Ceusters (2018) on how realism-based ontologies can be used both to reveal ways in which a CDM falls short of its goals and to resolve those shortcomings.

[3] See Ceusters and Smith (2006) for more on referent tracking in EHRs.

[4] See Ceusters and Smith (2015) on aboutness.

languages themselves, and this calls for an underlying ontological approach to the question of what a language such as English or Swahili is. Thus, to enable semantic interoperability of preferred language data from disparate sources that might use different data models, what is needed is a set of axiomatically related terms intended to represent (a) languages and (b) information about a person's preferred language.

### *1.3. Overview of what follows*

Below, we present the results of our efforts to develop a controlled vocabulary of this sort. This approach includes, among other things, an account of what the relation is between someone who has an ability to use a language and that language itself, and how that relationship can be axiomatically represented.

We are working within the context of the Open Biological and Biomedical Ontology (OBO) Foundry community and its principles.[5,6] As such, we aim

(i) to reuse terms from pre-existing OBO ontologies wherever possible in order to maintain orthogonality of ontologies and thereby avoid redundant efforts;
(ii) to situate any newly created terms within an appropriate home ontology, in line with the OBO principle concerning scope;
(iii) to develop our approach within the context of Basic Formal Ontology (BFO), the top-level ontology typically used by the OBO Foundry.[7,8]

Regarding (ii), we determined that an appropriate home ontology for the terms presented below is the Ontology of Medically Related Social Entities (OMRSE), which focuses on social entities relevant to health care and is BFO compliant.[9,10] Linguistic data are medically relevant, and preferred language data are gathered during health care encounters and recorded in EHRs.

We begin our account of language in the next section with our treatment of the ability a person has to use a language, or in other words, of their *linguistic competence*. In Section 3 we begin addressing how to ontologically handle languages themselves, and in Section 4 we present our approach to it, which is based on the idea that a language is an aggregate of capabilities. Then, in Section 5, we show how our approach captures important aspects of the nature of languages. This includes, for example, addressing conventions regarding what are counted as the same or different languages at a given time or across a period of time. In Section 6 we turn our focus to language data, including data on a person's primary language as well as data on preferred language in the context of a medical encounter. Then in Section 7 we briefly discuss how the proposed view relates to other approaches to the ontology of language.

## 2. Representing linguistic competences and their realizations

In this section, we address one of the more easily categorized types of entities that are key to our approach, namely linguistic competences. To put it simply, a linguistic competence is something like the ability to use a particular language. It is worth clarifying that, while "linguistic competence" has been used with a specific technical meaning by Chomsky (1965) and other linguists, our use of this term here is not tied to such usages, nor is it based upon any stance towards them.

[5] Smith et al. (2007).
[6] OBO Foundry (n.d.).
[7] Arp, Smith, and Spear (2015).
[8] BFO is now an International Organization for Standardization (ISO) standard. See: https://www.iso.org/standard/74572.html.
[9] Hicks et al. (2016).
[10] http://purl.obolibrary.org/obo/omrse.owl.

### 2.1. Linguistic competences

Someone who learns just one word (for instance "arigato," for use in expressing thanks in Japanese) is not thereby demonstrating linguistic competence. Rather, someone is realizing their linguistic competence in a given language in the sense here intended whenever they use or understand a phrase or sentence in a language in which they have at least the sort of facility manifested by children from about the age of five years. By this age, most children can speak fluently and easily, and are able to both initiate and sustain conversations.[11,12] Karmiloff-Smith (1986) describes it as "a frontier age psycholinguistically," and notes that it is by roughly this age that children "have built up a series of juxtaposed procedures for language use and understanding" and mastered an "utterance grammar."[13] By this age most children have acquired their language's basic principles of phonology, and they can form sentences that contain all the elements found in complex sentences formed by adults.[14,15]

### 2.2. Linguistic competences as realizable entities

It is rather uncontroversial to posit that there are abilities to do such things as read, speak, and write, and to use language for such purposes as communicating or recording information. After all, some entities can do those things and others cannot. Indeed, we humans start out in the latter category, but many of us are in the former at some point, due to the acquisition and development of linguistic competences.

It is also rather uncontroversial to suggest that a class of such entities—*linguistic competence*—is, like other sorts of capabilities, a subclass of BFO: *disposition*.[16] In BFO, *disposition* and *role* are subclasses of *realizable entity*. While a role exists due in part to factors external to its bearer, a disposition exists only in virtue of the way its bearer is in itself. A possible exception might be a self-assigned role whose existence depends only on the decisions of the bearer, such as a restaurant owner who begins to bear the role of head chef immediately upon choosing to do so. But for many roles, an entity can begin or cease to bear the role without changing internally, for example when someone's promotion or demotion goes into effect at midnight on a certain date.

Beginning or ceasing to bear a disposition, however, always requires some physical change in the bearer. In BFO 2020, to say *b* is an instance of *disposition* means the following:

> *b* is a realizable entity & *b* is such that if it ceases to exist, then its bearer is physically changed, & *b*'s realization occurs when and because this bearer is in some special physical circumstances, & this realization occurs in virtue of the bearer's physical make-up.[17]

To gain or lose a linguistic competence cannot be achieved by mere decision or due solely to external factors, but instead would require some intrinsic neurological change in the bearer.[18] Thus a linguistic competence is not a role. Instead, it is a disposition. Like other dispositions it is realizable, and it is realized for example in processes of reading or writing.

Perhaps more controversial is the matter of how many linguistic competences a person can bear. There is a sense in which each person has just one capacity for using language, where that one capacity can be

[11] Karmiloff and Karmiloff-Smith (2002: 3, 153-154).
[12] Centers for Disease Control and Prevention (2022, May 11).
[13] Karmiloff-Smith (1986: 54, 460, 474).
[14] Lyytinen, Aro, and Richardson (2007: 458).
[15] Hoff (2013: 301-302).
[16] Researchers in the BFO community are experimenting with a definition of 'capability' as a BFO: *disposition* whose realization some organism or group of organisms has an interest in. See Landgrebe and Smith (2019), Koch (2020), and Merrell *et al* (MS in preparation).
[17] International Organization for Standardization (2021).
[18] We do not have in mind some particular type of change. One cannot gain or lose any abilities without changing in some way. One cannot learn a language or anything at all without one's brain undergoing various changes.

realized by using any language the person learns. But in another sense you acquire a new ability (or set of them) when you learn a new language. You become able to do such things as read or write or converse in that language. For this sort of linguistic ability—which we call *linguistic competence*—each bilingual person has at least two, while each monolingual person has at least one. A person might also bear two competences for the same language, but for distinct dialects thereof.

### *2.3. Realizations of linguistic competences*

Further examples of processes in which one makes use of one's linguistic competence are speaking, listening to a speaker, signing in a sign language, and gesturing with a thumbs up sign when one agrees with a speaker. It may seem that many of these could be grouped together under the heading of "linguistic communication," and that a corresponding *linguistic communication* as a subclass of OMRSE: *communication* would provide a simple way to characterize the sort of process in which a linguistic competence can be realized. However, there are at least two reasons to take a different approach.

One reason is that, by definition, each instance of OMRSE: *communication* is a case of successful communication. The term is defined as follows:

> A process in which some participant shares some generically dependent continuant with some other participant. The former utilizes some specifically dependent continuant that concretizes the generically dependent continuant intended to be shared, while the latter interprets that specifically dependent continuant as concretizing some particular generically dependent continuant, aiming to accurately infer the other participant's intent.[19]

For example, if you write a text message and send it to someone who receives and reads it, then a communication has occurred, and it consists both of your act of writing the message and the recipient's act of reading it. Your act of writing the message realizes your linguistic competence, and the recipient's act of reading the message realizes their linguistic competence. But your linguistic competence is realized independently of whether or not the message is received and read. There are events that make it true that your act of writing is part of a process of communication, but the communication process taken as a whole does not stand in the relation of realization to any single linguistic competence.

Another reason to not represent *realization of linguistic competence* as a subclass of OMRSE: *communication* is that someone's linguistic competence can be realized by processes that are neither communications nor parts thereof. When the message you write and send is never received, your act of writing realizes your linguistic competence, but it is not a part of any communication. You may take notes exclusively for your own use; read those notes at a later time; keep a journal with no intent to allow others to read it; or practice your speech in the shower when no one is listening.

### *2.4. Concretization and generically dependent continuants*

What all of these cases have in common is that someone is either creating or making use of, for example, vibrations in the air, in order to make concrete certain abstract entities that are words and sentences. Those abstract entities are referred to in BFO as *generically dependent continuants* (GDC). For BFO, each GDC is an abstract entity that exists if and only if it is concretized as some quality pattern. For example, suppose you own a copy of the novel *Frankenstein*, where the item in question is some bound collection of pages. Suppose further that I own a distinct bound collection of pages that is also a copy of the novel *Frankenstein*. Some might be tempted to suggest *Frankenstein* is thus a class that has these and other copies of the novel as its instances. But *Frankenstein*—that one thing of which we each have some copy or version—is not a class of things. It is itself an individual instance of the type we call *novel*. Thus, the

[19] http://purl.obolibrary.org/obo/OMRSE_00002068. On the meaning of 'generically dependent continuant' see 2.4 below.

relation of our copies of the novel *Frankenstein* to the novel itself is not instantiation. For BFO, it is what is called *concretization*. More precisely, each copy is the bearer of patterns that concretize the novel.

While the novel *Frankenstein* is an individual entity, there are many things of the sort that we have been describing as "copies" thereof. For BFO, the novel *Frankenstein* is an instance of *generically dependent continuant*. It is an abstract thing that we can neither see nor touch, except insofar as we can see or touch a material entity that bears some copyable pattern (for instance of piles of ink on paper) through which the novel is concretized. As a generically dependent continuant, its existence requires that it be *concretized by* at least one pattern in this way. Consider my copy of *Frankenstein*. On the pages there is ink, and that ink is arranged into various shapes to form letters, which in turn are arranged to form words, and so on. Those shapes are instances of *quality*, and that collection of qualities concretizes the novel. Likewise, your bound collection of pages bears its collection of shapes that concretizes one and the same novel.

In taking notes on a meeting, I put ink onto paper in certain shapes, where those shapes concretize—give concrete form in time and space to—abstract sentences that are *about* the meeting. In reading them later, I am making use of those same shapes. Similar processes occur when I communicate with some other person in writing. Or, if we replace the shapes of the ink with various patterns of sounds that I make when speaking, then we can see that concretization occurs in a similar fashion when I communicate vocally. While speaking, I would not only bring into existence the patterns that concretize words and sentences, but would simultaneously establish the concretization relation they bear to those words or sentences. The listener would then, if she understands me, interpret those sound patterns as concretizing those same words or sentences.

How, now, are we to understand the relation of concretization? First, we call upon the BFO distinction between two kinds of dependent continuants, namely the specifically and the generically dependent, or in other words SDCs and GDCs. Examples of the former include qualities, roles, and dispositions. Examples of the latter include abstract patterns such as words and sentences.

Each SDC is dependent on some specific bearer: I can have a suntan that is exactly similar to your suntan, but I cannot have *your* suntan. This is because qualities—and the same holds for roles and dispositions—cannot migrate from one bearer to another. I can, though, when I see a square on one sheet of paper, create a copy of this exact same pattern by drawing a square on a second sheet of paper. Such copyable patterns are what BFO means by 'generically dependent continuant.' To say that they are copyable means that they can migrate from one concrete bearer to another. GDCs are involved in all copying processes involving continuants, whether copying data from our notes on paper into a new spreadsheet, or copying a carpenter's design for a chair by shaping and joining pieces of wood, or copying a DNA sequence into a new molecule of messenger RNA.

### *2.5. Concretization-related processes: utilization and interpretation*

We divide concretization-related processes into two sorts: *concretization-utilization* and *concretization-interpretation.* We define these concretization-related acts as subclasses of BFO: *process*. In BFO, *p* is a process if and only if *p* is an occurrent that has some proper temporal part and for some time *t*, *p* has some material entity as a participant at *t*.[20]

A ***concretization-utilization process*** occurs when someone makes use of an SDC because it concretizes a certain GDC. An example is burning a flag because it is a flag with a certain pattern, or making a fist because this creates a shape that someone will find threatening. In each of these cases we are using an SDC because it concretizes some *already existing* GDC. When Kekulé created the symmetrical ring pattern representing the benzene molecule, he was using an SDC to *create* a new GDC.

[20] International Organization for Standardization (2021).

When you share information with someone by speaking to them, you perform a concretization-utilization process: the concretization relation is established between the patterns of sounds you produce and the piece of information you aim to share.

You might instead establish the concretization relation between a GDC and some pattern that already concretizes a different GDC. Suppose there is a sign that reads "Quarantine Area: Do Not Enter" outside the room *R1* in which there is a patient *P1*. Patterns on this sign are being *utilized as a concretization* of a certain GDC to serve as a warning about room *R1* and patient *P1*. When you move this sign to a different room *R2*, these same patterns then concretize a new GDC that is about this new room and about a different patient *P2*.

To be clear, instances of *concretization-utilization process* do not include just any process in which someone uses a pattern, or bearer thereof, that happens to concretize something. For example, throwing a hefty book at an intruder is a process of utilizing that book for some purpose, but it is not a *concretization-utilization process*.

A ***concretization-interpretation process*** occurs when someone *interprets* some pattern to be a concretization of some GDC. For example, when you see a "STOP" sign while driving, you effortlessly infer that it concretizes an instruction to stop within a certain area and to check for traffic before proceeding. Examples are not restricted to those in which you are entirely accurate in your interpretation. They are not even restricted to those in which the patterns in question already concretize anything at all. For example, if the wind happens to blow some fallen branches into a pattern that looks like "hi," you might interpret the patterns they form as concretizing a word, perhaps assuming that they had been intentionally arranged in that way. While that interpretation is incorrect, this does not change the nature of the interpretation process going on within you.

Our definitions and axioms for these terms, formulated using Manchester syntax, are as follows:

***concretization-utilization process*:**

| | |
|---|---|
| Definition: | Process in which some participant utilizes some SDC as a concretization of some GDC |
| Axioms: | SubClassOf: BFO: *process*<br>SubClassOf: *has participant* some (BFO: SDC and (*concretizes* some BFO: GDC))<br>SubClassOf: *has participant* some BFO: *material entity* |
| Examples: | Taking notes on a meeting is an example in which the concretization relation is newly established, since the SDCs that concretize the GDCs come into existence as the notes are written. In contrast, using slides prepared by someone else in order to convey information during a presentation is a case of *using* a concretization in which the performer neither brings the concretizing SDCs into existence nor is responsible for their standing in the concretization relation to the relevant GDCs. A nonlinguistic example is drawing the logo of one's favorite brand, in which one creates a pattern that concretizes a GDC that is also concretized by patterns on products of the brand. |

***concretization-interpretation process*:**

| | |
|---|---|
| Definition: | A process in which some participant infers that some particular SDC stands in the concretization relation to some particular GDC |
| Axioms: | SubClassOf: BFO: *process*<br>SubClassOf: *has participant* some BFO: *SDC*<br>SubClassOf: *has participant* some BFO: *material entity* |
| Examples: | For example, reading a text message that says, "It is raining outside," and inferring that the words on the screen are meant to convey information about the weather. The pattern on the screen that corresponds to the words is the SDC, |

which concretizes a GDC about the rain. Or, hearing your spouse say, "Can you come to the kitchen?" and then knowing that your spouse wants you to come to the kitchen. You have interpreted the patterns of the sound as concretizing a GDC about what your spouse wants. Or, seeing a drawing of a basketball player and inferring—perhaps incorrectly—that it is of Michael Jordan.

We can see from our examples that not all concretizations are linguistic in nature. The examples should make clear that interpretation includes lots of processes that happen so automatically that we rarely notice that they are happening, for example every time we hear and understand someone speaking.

For our purposes in this paper, we provide language-specific versions of our two definitions, as follows:

***linguistic concretization-utilization process*:**

Definition: A concretization-utilization process in which the utilized SDC concretizes the GDC through use of some language.

Axioms: SubClassOf: OMRSE: *concretization-utilization process*
SubClassOf: *has participant* some (BFO: *material entity* and *bearer of* some *linguistic competence*)

***linguistic concretization-interpretation process*:**

Definition: A concretization-interpretation process in which the SDC is inferred to concretize the GDC in a way that makes use of some language.

Axioms: SubClassOf: OMRSE: *concretization-interpretation process*
SubClassOf: *has participant* some (BFO: *material entity* and *bearer of* some *linguistic competence*)

### *2.6. Formal representation of linguistic competence*

Finally, here is the term we developed for representing linguistic competence, which makes use of the preceding two terms in its associated axioms:

***linguistic competence*:**

Definition: A disposition that inheres in some material entity and is such that that, if realized, it is realized by either some linguistic concretization-utilization process or some linguistic concretization-interpretation process.

Axioms: SubClassOf: BFO: *disposition*
SubClassOf: *inheres in* some BFO: *material entity*
SubClassOf: *realized in* only (OMRSE: *linguistic concretization-utilization process* or OMRSE: *linguistic concretization-interpretation process*)

Note that while we define *linguistic competence* in such a way that each instance is realized in only the sort of linguistic concretization-related processes defined above, this definition does not imply that each such instance realizes some linguistic competence. This is because you can utilize or interpret a linguistic concretization in a language for which you have no competence. For example, the use of a translation dictionary or automated translation service can enable you to create or interpret concretizations in a language you do not speak, which in turn might enable you to have some degree of success in communicating with speakers of that language. Additionally, you could know a few words of a language without having a competence for that language. For example many monolingual English speakers know the Spanish word "gracias," and one of them could thank another by using this word in what would be a successful communication. So while we take linguistic competences to be realized only by utilizations or interpretations of linguistic concretizations, we do not take the realization of a linguistic competence to be necessary for processes of these types to occur.

## 3. Representing languages

Having addressed linguistic competences and their realizations, we focus next on languages themselves.

### *3.1. Languages as classes or individuals*

So long as we agree that more than one language exists, it is uncontroversial to propose we treat each of those things as an individual that is an instance of some class. Call that class *language*.

Perhaps more controversial is the question of which sorts of things are instances of that class. This is closely related to the question how we should characterize the relationship between the class *language* and specific languages, such as the German language.

We adopt a view according to which each language is an individual continuant that is an instance of the universal *language*. Languages are continuants because they exist and change through time, but as contrasted with occurrents such as acts of speaking, they do not have temporal parts. The English language that existed during the 1970s is the same English language that has existed since at least the time of Shakespeare, just as a child is the same person as he will later become when he grows up into being an adult.

Words and sentences as GDCs are instances of the universals *word* and *sentence*, and individual spoken or written words or sentences are *concretizations* of these instances.

If you and I both speak English, then that is one language that we each speak. It is not as though you have your instance of English and I have mine. Of course, there is your competence to speak English, and there is my distinct competence to do so, but a competence to speak English is distinct from English itself. If I cease to exist, then so too does my linguistic competence, while the English language continues to exist.

We thus treat specific languages—such as the Spanish language or American Sign Language—as individual entities existing in time. This allows us to do justice to the fact that each language goes through changes as time passes. This is a characteristic feature of continuant entities, as opposed to occurrents and to types or classes. We address nuances, such as dialects and linguistic boundaries, in Section 5.

### *3.2. Is language a generically dependent continuant?*

Some might suggest that in concretizing the word “dog” by using this word in a sentence I would also be concretizing the English language, or at least part thereof. Using “word” to refer to a sort of thing that is concretized (as opposed to the sort of thing that concretizes), some might insist that the words of a language are *parts* of the language, in some sense of the word “part.” If those words are entities that can be concretized, then they are GDCs. If a language is something like a collection of words, or something that has words as its parts, then it, too, might be properly described as a GDC.

We grant there may be some such collection of GDCs for any given language. It is difficult to see how any language could lack one. The Spanish language, for example, could not exist without a collection of Spanish words. Nor could it exist without the corresponding linguistic competences and the speakers of the Spanish language who have these competences.

We are dealing here, however, with a certain kind of circularity. Without the Spanish language existing at some time, no words could be Spanish words, and no linguistic competences could be competences to speak Spanish. Without the Spanish language, the Spanish linguistic community’s members might still exist, but not as a linguistic community. Thus, while the users of a language, their relevant competences, and the associated collections of GDCs (of words, sentences, texts) are all necessary for a language to exist, so too is the language necessary for those things to be the types of things they are. Defining them all without running into circularity would be impossible, and thus we must choose one as a starting point

and declare it primitive or basic. For this we would then provide merely an *elucidation* and use the chosen primitive to provide *definitions* of the remaining terms as needed. We here choose as our primitive term 'linguistic competence' as presented in Section 2.5 above, thereby identifying each instance of *language* as a collection of linguistic competences of the sort described in the foregoing.

## 4. A dispositional account of language

The proposed account is one for which *language* is defined in terms of BFO: *disposition*, for languages are aggregates of linguistic competences and a *linguistic competence* is a *disposition*. Like other dispositions, these competences are *realizable entities*, in BFO terms. Each linguistic competence is something that is *realized* in certain activities—reading, writing, speaking, and so forth. These are all examples of processes that realize a capability of a certain sort.[21] It is entities of this sort, we believe, that can yield the most coherent account of what a language is.

A dispositional approach to language can do justice to the fact that a language can exist even while none of the associated processes of realization are happening, for example when a given language has but a small community of users and during some interval of time it so happens that none of them is making use of the language. And as we discuss below, taking a language to be an *aggregate* of dispositions allows for capturing the ways in which changes to and subdivisions of the language are mirrored in changes to and subdivisions of the relevant linguistic community.

### *4.1. Languages and linguistic communities*

The processes that realize linguistic capabilities are actions performed by the bearers of those capabilities. We can thus identify the language itself with the aggregate of those linguistic competences that are enjoyed by the members of the relevant group and enable communication among combinations of those members. We cannot say it is the aggregate of *all* linguistic competences of those people, since we must exclude those additional linguistic competences which exist where speakers of a given language are multilingual. A language is the aggregate of the *relevant* linguistic competences on the part of all the members of the aggregate of users of the language. Thus, there is, by definition, exactly one such aggregate of language users for each language, though who the members of this aggregate are will change, incrementally, over time. Let us use the term *linguistic community* to refer to each such aggregate of users. Our approach can now be summarized in the following elucidation:

*language* = an aggregate of capabilities that inhere in the members of the relevant linguistic community.

Having already presented *linguistic competence* above, we next present our formal treatment of *language*, as well as of *linguistic community*:

***language*:**

| | |
|---|---|
| Elucidation: | Aggregate of linguistic competences that are considered as forming a distinct group on the basis of perceived common characteristics, such as mutual intelligibility among their bearers, in addition to historical or cultural factors. |
| Axioms: | SubClassOf: (*has member* some OMRSE: *linguistic competence*) and (*has member* only OMRSE: *linguistic competence*) |

[21] Again, see Landgrebe and Smith (2019), Koch (2020), and Merrell *et al* (MS in preparation) for more on recent research within the BFO community on defining 'capability.' See also Jobst Landgrebe and Barry Smith's *Why Machines Will Never Rule the World* (2022).

***linguistic community*:**

Definition: A maximal collection of humans each member of which bears a linguistic competence for the same language.[22]

Axioms: SubClassOf: OMRSE: *collection of humans*
SubClassOf: *bearer of* some OMRSE: *language*
SubClassOf: *has member* only (*bearer of* some OMRSE: *linguistic competence*)

Note that while we acknowledge the role of mutual intelligibility in individuating languages, it is not the sole standard that can be used for this purpose. As we discuss in Section 5 below, languages are divided into subgroups—dialects for example—and it is sometimes a matter of convention or historical contingency whether two subgroups are regarded as being of the same language.[23] A result is the possibility of two subgroups of different languages sharing greater mutual intelligibility than two subgroups of the same language. To adopt some degree of mutual intelligibility as the sole criterion for individuation would thus be more prescriptive than descriptive. Our proposed elucidation of *language* has the advantage that it can accommodate real world deviations from any mutual intelligibility-based standard.

### *4.2. Languages and linguistic communities as aggregates*

As we have proposed an account for which both languages and linguistic communities are aggregates—of linguistic competences and of their bearers, respectively—we now need to say more about what we mean by this.

Note that OMRSE: *linguistic community* is a subclass of OMRSE: *collection of humans*, which in turn is a subclass of BFO: *object aggregate*. In BFO-2020, *object aggregate* is defined as "a material entity consisting exactly of a plurality ($\geq 1$) of objects as member parts which together form a unit." An object aggregate has no parts other than its members.

BFO allows that an object aggregate can at some times have just one member—for example when the population of an endangered species is reduced to just one organism—so long as it has multiple members at some other time. Furthermore, an aggregate in BFO is able to preserve its identity while changing its members. For example the English-speaking linguistic community gains a new member when someone learns English, and loses a member when someone who knows English dies. This is in contrast to the sort of view of collections outlined for example in Bittner, Donnelly, and Smith (2004), wherein collections have a fixed set of members, and are atemporal entities that are fully present when all members exist and partially present at times when some but not all of its members exist.[24]

The notion of an aggregate that changes members over time is uncontroversial from the perspective of common sense. We think and talk about such aggregates on a regular basis. For example a football team consists of a number of players, and the same team can continue to exist even if some players leave the team or new members join. The aggregate of persons who work for your employer is an aggregate of which you are one member, and you likely care to some extent about changes of membership in that

[22] This will recall the idea of the *speech community* in (Gumperz, 1968). Like a linguistic community, a speech community is a "human aggregate" that is "set off from similar aggregates by significant differences in language usage." However, where Gumperz specifies that a speech community is characterized by "regular and frequent interaction" among its members, a linguistic community as we conceive it can stretch across multiple subcommunities that do not interact, for instance for reasons of geospatial disconnectedness. Another contrast with linguistic communities is that speech communities for Gumperz can be set apart from one another on the basis of less drastic differences, for example through the use of distinctive jargon or slang.

[23] Hence our use of the phrase 'are considered as' – see also Section 7.1 below.

[24] We do not deny the importance of representing what Bittner *et al.* call "collections," but we must also represent other sorts of pluralities such as aggregates that gain or lose members over time. Otherwise there will be difficulties when it comes to dealing, for example, with families existing into the future in virtue of bearing children in successive generations. How, in such circumstances, would the needed whole atemporal set of members be determined?

aggregate such as might occur if someone leaves for another job or if a new employee is hired. A family is an aggregate of persons, and families often celebrate when they gain a new member—for example through birth, adoption, or marriage—and mourn when they lose a member through death. These are all examples of aggregates that begin to exist at some time and can undergo gains and losses of members over time. They illustrate also a certain phenomenon of vagueness as to who or what counts as member of (for example) a family at some given time (Bittner and Smith, 2003).

The human population—understood not as a number but rather as the plurality of all living humans—is another example of an aggregate that is familiar and that undergoes changes over time. It gains a member with each new human life and loses a member with each human death. None of its current members are the same as any of the members it had two centuries ago. Overall it has tended to gain a member more often than it loses one, and thus one of the changes it has undergone over time is that it has grown in size.

Of the mentioned examples—football teams, employees of your employer, and members of a family—all are *subaggregates* of the human population. By *subaggregate of* we mean a relation in which an aggregate $x$ stands to an aggregate $y$ at time $t$ if and only if the members of $x$ at $t$ are a subset of the members of $y$ at $t$. There are many ways the human population can be subdivided into subaggregates along different axes such as race, ethnicity, gender, religion, handedness, and so forth.

As Smith (1999: 278-279) notes, some axes along which humans or other entities may be divided into aggregates—or "agglomerations" as Smith calls them there—are such that they "track more or less bona fide boundary lines," while others "are exclusively or primarily the product of fiat." The former "exist independently of all human cognition," while the latter are distinguished "as a result of human decision or convention." For example a particular colony of ants continues to exist while undergoing changes of membership as new members come into existence through reproduction and other members die. Colony membership affects their behavior. Bona fide boundaries then exist because ants of the same colony cooperate and live in a shared space, while ants of different colonies live separately and sometimes attack one another. Examples of an aggregate boundary which is the product of human choice or convention is the "pronouncedly fiat partition" of the human population into Americans and non-Americans (Smith, 1999: 279). United States citizenship conditions are the product of human decisions, and they determine who is in the aggregate of US citizens and thereby also determine who is not. Similarly, a football team and the aggregate of employees of your employer are aggregates of persons whose membership is determined by human decisions.[25] As we discuss in Section 5.4 below, the boundaries between what are regarded as distinct languages are also in many cases fiat partitions, as they, too, are determined to a degree by human convention.

For representing how a member of an object aggregate relates to the aggregate, BFO has until recently used a formal definition of what it calls *member part of* as follows:

> $b$ member part of $c$ at $t$ =Def. $b$ is an object; & $c$ is an object aggregate; & there is at $t$ a mutually exhaustive and pairwise disjoint partition of $c$ into objects $x_1, \ldots, x_n$ (for some $n \neq 1$) with $b = x_i$ (for some $1 \leq i \leq n$).[26]

This implies that *member part of* is limited to holding between an object and an aggregate at a time at which the aggregate has multiple members, as indicated by "for some $n \neq 1$." In the most recent version (BFO-2020), however, an object aggregate with multiple members can survive being reduced to just one

[25] Since linguistic communities are distinguished on the basis of the languages they speak, and convention plays a role in determining the boundaries between languages, convention thereby also plays a role in determining the boundaries between linguistic communities. In this sense, linguistic communities are social categories, comparable to races and ethnicities. OMRSE researchers are currently developing a BFO- and OMRSE-based approach to representing social categories and the categorization schemes that demarcate them. See Dowland, Diller, and Hogan (MS in preparation).

[26] The Relation Ontology (RO) defines the similar relation *member of* as "a mereological relation between an item and a collection." See http://purl.obolibrary.org/obo/RO_0002350. For more on RO, see Smith, Ceusters, Klagges, *et al.* (2005).

member. This yields a modified definition of the relation *member part of* in which "for some $n \neq 1$" is replaced with "for some $n \geq 1$."[27]

Since aggregation is not limited to objects, even if there is no parallel BFO class of *object aggregate* labeled "disposition aggregate," we can generalize the sense of aggregation at play in BFO to apply also to entities of other types. In this case we are concerned with an aggregate of linguistic competences. Many of the key aspects of the *linguistic community* with respect to its being an aggregate are mirrored in the associated aggregate of *linguistic competences*. In many cases, one gains or loses a member at the same time as the other.[28] For example when a speaker of a language ceases to exist, so too does that person's linguistic competence; and when a person learns a new language, that person becomes a new member of the corresponding linguistic community. Like an object aggregate, the aggregate of competences for a given language can clearly persist through gains and losses of members. And as with other object aggregates, the aggregate that is a linguistic community survives the loss of member parts so long as at least one speaker survives.

## 5. Languages, dialects, and linguistic boundaries

If we conceptually divide linguistic competences into the largest possible groups such that any two competences in the same group enable mutual intelligibility between their respective bearers, then there would be overlap between some pairs of these maximal aggregates of mutually intelligible competences. The reason for this is that—as we see most clearly in the case of dialects—mutual intelligibility is not transitive, and so it is possible for some dialect *D1* to be mutually intelligible with each of two others *D2* and *D3*, even if *D2* and *D3* are not themselves mutually intelligible. Thus one group of the type just described would include competences for both *D1* and *D2*, while another would include competences for *D1* and *D3*.

Note that we leave open in this description of these three dialects whether any of them are dialects of the same language. Thus our description is consistent with a number of different ways these dialects could be grouped into languages, including (i) the three are dialects of the same language, (ii) just two of the three are dialects of the same language, or (iii) each one of them is a dialect of a different language. This is because—as we discuss in more detail below—the ways in which linguistic competences are in fact grouped into languages are in some cases matters of convention or historical contingency.[29]

### *5.1. Language continua and dialect continua*

While the boundaries between countries are normally crisp and for long periods unchanging, boundaries between languages as cultural phenomena are often vague and are subject to continuous change.[30] For example the Germanic languages each resulted from a different series of gradual changes to dialects of a common parent language referred to now as Proto-Germanic. As different dialects of Proto-Germanic underwent those changes, what are now considered distinct languages formed out of them in a process which extended over centuries. Proto-German branched into West Germanic, North Germanic, and East Germanic. West Germanic split into Anglo-Frisian (which further split into English and Frisian) and Netherlandic-German (which further split into Netherlandic and German). North

[27] See https://github.com/BFO-ontology.

[28] There are exceptions, for as we discuss above, a speaker of one dialect of a language could gain a competence for another dialect of the same language, thereby increasing the number of competences for the language without adding a new member to the linguistic community.

[29] As Max Weinreich (1945) puts it, "A language is a dialect with an army and navy."

[30] Palander, Riionheimo, and Koivisto (2018: 8).

Germanic split into West Scandinavian (which further split into Icelandic, Faroese, and Norwegian) and East Scandinavian (which further split into Danish and Swedish). East German developed into Gothic.[31] But the boundaries between languages that form from such splitting are vague and result from gradual changes over time. This means that two languages may early on share mutual intelligibility, but lose this feature in later versions of those languages due to changes in the underlying populations occurring over time.

In some cases languages that have developed from a common parent may retain some degree of mutual intelligibility even while undergoing substantial changes. For example Norwegian, Swedish, and Danish have a high degree of mutual intelligibility and yet they are regarded as distinct languages, rather than as dialects of the same language.[32] Additionally, some West Germanic dialects are considered dialects of German while others are classified as dialects of Dutch, while at the same time there is little or no mutual intelligibility between certain pairs of dialects of German.[33] Similarly, varieties of Karelian and Finnish form a *language continuum* wherein White Sea Karelian and Eastern Finnish are to a degree mutually intelligible, while Western Finns and Olonets Karelians have much greater difficulty understanding one another.[34] There are *dialect continua* which involve a lack of mutual intelligibility between dialects at opposite ends.[35,36] Some varieties of Karelian are mutually intelligible with some varieties of Finnish, but not all varieties of Karelian are mutually intelligible with one another.

The geographical boundaries between dialects, too, are paradigmatically vague. Dialects change and spread continuously, driven, for example, by the desire of potential speakers to gain acceptance in a new group. Dialect distinctions trace not merely geography, but also age cohorts, class distinctions, and behavior.[37]

### *5.2. Dialects as subaggregates of languages*

One noteworthy virtue of the view of a language as an aggregate of linguistic competences is that it automatically provides a simple framework for understanding dialects. Dialects are *subaggregates* of the relevant circumcluding language. The users of a dialect of some language then form a subaggregate of the members of the broader linguistic community of that language. Speakers of a Manchester dialect are *ipso facto* speakers of English, and they are realizing their capabilities for speaking English by speaking one or other Manchester dialect. Additionally, dialects can be subaggregates of other dialects in the same way that dialects can be subaggregates of languages. For example, the New England English dialect is a subaggregate of the North American English dialect, which in turn is a subaggregate of the English language.

### *5.3. Dialect-related entities*

Since dialects are ontologically similar to languages in our account, and since we have presented OMRSE terms for representing linguistic competences and linguistic communities, it is simple now to give an overview of parallel terms for dialects. Indeed it is largely a matter of replacing “language” with “dialect,” and replacing “linguistic” with “dialectal,” in the labels and definitions of those terms.

[31] Buccini & Moulton (2016).
[32] Chambers and Trudgill (1998: 3).
[33] See Palander, Riionheimo, and Koivisto (2018: 8), and Chambers and Trudgill (1998: 3).
[34] See Palander, Riionheimo, and Koivisto (2018: 47), and Koivisto (2018).
[35] See Kunnas (2018: 124).
[36] See also Gumperz (1968: 385) on dialect chains.
[37] See Palander, Riionheimo, and Koivisto (2018), and Chambers and Trudgill (1998: 9).

A *dialectal competence* is a *linguistic competence*. Indeed, unless someone speaks a language without doing so in one of its dialects, that person's linguistic competence is also a dialectal competence. As noted above, a Manchester dialect competence is an English competence.

The community of speakers of a given dialect forms a *dialectal community*. As a dialectal community is a subaggregate of a linguistic community, so the corresponding dialect is a subaggregate of the corresponding language. This means that, just as the collection of competences for a Manchester dialect are a subaggregate of the total collection of English competences, so too is the community of speakers of a Manchester dialect a subaggregate of the linguistic community of English speakers. Just as linguistic communities overlap in the case of bilingual persons, so too may dialectal communities within the same linguistic community overlap. For as mentioned in Section 2.2 above, it is possible for a person to bear two linguistic competences for the same language by bearing competences for different dialects of the same language.

### *5.4. Linguistic boundaries and groupings*

As we acknowledge in our elucidation of the meaning of 'language' and as we see from the examples above, degree of mutual intelligibility alone does not provide a single, consistent standard by which it can be decided what are the same or different languages. Above all, what are counted as different dialects of the same language may differ greatly, so that there can be a very low degree of mutual intelligibility between two persons who use what is acknowledged to be one and the same language.

In addition to (degrees of) mutual intelligibility, other—cultural, historical, and geographic—factors play a role. What is one language at a given time may at a later time have evolved into two, where the speakers at the geographic fringes where the two languages meet may well find their respective acts of speaking mutually intelligible, even though they are speaking different languages. Since we aim to account for *language* in a manner that allows us to treat actual human languages as instances of it, we do not ignore these ways in which the individuation of languages is in part a matter of convention. Thus, just as football teams and the aggregate of non-Americans are products of fiat divisions of persons into groups, so too are linguistic communities and languages.[38]

## 6. Language-related data elements

Having laid out the core of our approach, we next illustrate the utility of some of the terms presented above in some practical examples of demographic data elements. Each of these is a code or identifier for a language and is used to indicate a way in which a given person is connected with that language. We represent each of these data elements as subclasses of the Information Artifact Ontology's (IAO) *information content entity* or ICE, which is defined as, "A generically dependent continuant that is about some thing."[39]

We begin with a type of ICE that merely tells us that a given person has a linguistic competence for a given language:

***linguistic competence information content entity*:**

| | |
|---|---|
| Definition: | An ICE that conveys that a particular person has a linguistic competence for a particular language. |
| Axioms: | SubClassOf: IAO: *ICE* |

[38] See also Gumperz (1968: 385) on linguistic boundaries that "are defined partly by social and partly by linguistic criteria."
[39] http://purl.obolibrary.org/obo/iao.owl.

SubClassOf: *is about* some (*bearer of* some (OMRSE: *linguistic competence* and member of some OMRSE: *language*))

This term enables us to introduce subclasses whose instances convey further detail about the person and the language, such as the contexts in which the person uses the language and whether it is the person's primary language.

### *6.1. Primary language and languages spoken at home*

In some cases researchers are interested in gathering data on the languages that persons are most proficient at using. Corresponding queries are sometimes contained in instruments designed to gather data related to social determinants of health, as in (i) the PRAPARE instrument, which asks respondents what language they are most comfortable speaking, and (ii) the Healthy Planet module of the EHR system developed by Epic Systems Corp. (Verona, WI), in which respondents may select a language as the value for the "Primary Language" field.[40,41]

Additionally, data are sometimes gathered concerning the languages that persons use within their homes. Some people use one language more than others within the home while mostly using another language outside the home. Some use a number of languages within the home. For example in the United States, an immigrant might speak English more than other languages when outside the home, while speaking a mix of English and her native language at home. Two examples of instruments that gather data on the languages spoken in a person's home are the U.S. Census Bureau's American Community Survey (ACS) and the California Health Interview Survey.[42,43]

Clearly a person's primary language is one for which they have a linguistic competence. Likewise, we presume a person has linguistic competences for the languages they use at home. We thus represent ICEs about each of these as subtypes of the above *linguistic competence ICE*:

***primary language information content entity*:**

Definition: Linguistic competence ICE that conveys that the language is the one for which the person has their most proficient linguistic competence.

Axioms: SubClassOf: OMRSE: *linguistic competence ICE*

***language-at-home information content entity*:**

Definition: Linguistic competence ICE that conveys that the person uses the language at home.

Axioms: SubClassOf: OMRSE: *linguistic competence ICE*
SubClassOf: *is about* some (*linguistic competence* and (*realized in* some (*occurs in* some (*bearer of* some *residence function*))))

For now we rely on the textual definition of *primary language ICE* to specify what differentiates it from its parent class. A formal representation of degrees of proficiency for linguistic competences in particular calls for an account of degrees of proficiency of capabilities in general, which goes beyond the scope of this discussion.

### *6.2. Preferred language*

One demographic data element commonly recorded in EHRs is a code or identifier for a *patient's preferred language*. Such a data element is meant to tell us which language some patient prefers to use

[40] For PRAPARE, see National Association of Community Health Centers (2016).
[41] Dowland and Hogan (2022) present an ontological framework for some other data elements in these instruments.
[42] U.S. Census Bureau (n.d.).
[43] UCLA Center for Health Policy Research (2022).

when communicating within the context of some health care encounter. For example, if patient P prefers to communicate in Spanish during such an encounter, the record for that encounter might contain a preferred language field in which there has been entered the value that denotes the Spanish language. By entering that value for that field in some information system, someone records a piece of data to the effect that *P's preferred language during this encounter is Spanish*.

Preferred language data may be about the language that the patient prefers to use within the context not only of some particular health encounter, but also of all health encounters at a given facility. Such data are of importance because they may disclose important aspects of communication between a patient and a provider or other health care staff. The clinical encounter requires the effective establishment of the patient's medical history and an effective physical examination, and both require the achievement of intersubjective understanding between physician and patient. Such intersubjectivity is of crucial importance also in achieving an effective diagnosis and treatment. It not only requires that the physician speaks a language that the patient understands, but also that the patient feels that the physician understands the patient's needs. Intersubjectivity is based on the entire repertoire of human communication. It enables the patient not only to provide information that the physician needs in order to obtain a diagnosis, but also to understand any proposed treatment and to make informed decisions on this basis. While there are means to facilitate communication in cases where no language is shared between patient and provider—using a human translator, perhaps, or even a translation device—in such cases it would be difficult to achieve the same degree of intersubjectivity as that which obtains when a language is shared, and the patient's ability to share and understand information can be limited as a result. This has potential repercussions not only for health outcomes following from the encounter, but also for the patient's ability to provide informed consent.

### *6.3. Preferred language ICE*

Like the terms for language-related data elements presented above, we model the information about someone's preferred language as a subclass of IAO: ICE. However in this case we cannot presume the person bears a linguistic competence for the chosen language. For example, if you are a patient in a hospital in a foreign country and the staff all speak only one language for which you lack any competence, your best means of communicating with them might be via their language and your use of a translation device. This example also highlights that the preferred language need not be the language that the patient would most like to be using. Instead it is the one that has been chosen to be used, either by the patient or someone else.

Below are the two OMRSE terms that are most directly relevant to representing preferred language source data:

***preferred language information content entity*:**

Definition: An ICE that is about some person and some language, and that conveys the language chosen to be used in communications with that person during some process.

Axioms: SubClassOf: IAO: *ICE*
SubClassOf: *is about* some BFO: *object*
SubClassOf: *is about* some OMRSE: *language*

The axioms for *preferred language ICE* reflect that an instance thereof is *about* at least two things. One is the person whose preferred language it is, represented here as an instance of BFO: *object*. The other is the language that is recorded as that person's preferred language in the given context. Note that we do not relate the person to the preferred language by, say, requiring that the person bear a linguistic competence for that language. One reason for this is that, as noted above, there might be cases in which a person's chosen preferred language is not one for which they have any competence at all. The patient

might even lack *any* linguistic competence for *any* language, such as when the patient is an infant. In such cases, some organizations might enter in the 'preferred language' field a value that corresponds to the primary language or preferred language of the parents. The recorded ICE would then be about both the infant and a language, even though the infant does not have a linguistic competence for that or any other language.

Since the patient's preferred language can differ from their primary language and may even be one for which the patient has no competence, preferred language data might be most telling when available alongside information of the sort contained in linguistic competence ICEs. For example if primary and preferred language data are captured about the same persons, researchers could in principle use that data to investigate effects of language barriers in health care.

### *6.4. Preferred language and preferred spoken language*

A key reason for the presence of a preferred language field in EHRs is its inclusion in Federal Government guidelines for stage 1 Meaningful Use requirements.[44] Those guidelines define "preferred language" as "The language by which the patient prefers to communicate," and thus do not specify a mode of communicating (such as through speech or through writing).

However, definitions restricted to certain modes of communication can be found in the CDMs used by many organizations, even where that definition does not correspond to the available value options. Consider, for example, the CDM used by the National Patient-Centered Clinical Research Network (PCORnet) (2019), in which the field corresponding to the patient's preferred language has the label 'PAT_PREF_LANGUAGE_SPOKEN,' and is defined as, "Preferred spoken language of communication as expressed by the patient." That definition fails to address cases in which the patient is unable to communicate through speech. But there may be cases in which a patient has recently lost the ability to hear or speak, and she will likely then prefer written communication. In such cases, "preferred spoken language" is a misleading description. But presumably some value option is chosen, such as "other" or one that denotes the same language via which the patient is communicating in writing.

Additionally, some of the value options for the field are codes that specifically denote written languages: while *written* Chinese is common to speakers of both Mandarin and Cantonese, the latter are distinct *spoken* languages, yet PCORnet codes both as 'ZHO,' for Chinese. Following the direction to handle those languages in that way, the CDM states, "Within the ISO 639-2 value set, there is no distinction between the two." But the reason ISO 639-2 makes no distinction between those two spoken languages is precisely because ISO 639-2 is primarily intended as a standard code set for *written* languages.[45]

One result of this discrepancy is that accurately mapping such data elements to particular languages will be difficult, except perhaps where the raw data include reference to the spoken language. If all we know of someone's preferred spoken language is that it has been coded as written Chinese, then we do not know what language it is that they speak. One way to handle this is grouping together Chinese spoken languages as instances of a more specific subclass of *language*—such as *Chinese spoken language*—so that, even when we cannot assert *which* instance of that class is the person's preferred language, we can narrow it down to being at least some instance of that class.

Since the corresponding data item can be about language as written, or perhaps even about a lack of linguistic capacity, we do not restrict our definition of *preferred language ICE* to language as spoken. By using a class defined without that restrictive and sometimes false assumption, data resulting from such cases are annotated in a way that is less likely to misrepresent what really occurs.

[44] Centers for Medicare & Medicaid Services (CMS) (2014).
[45] See Library of Congress (2014).

### *6.5. Dialect data*

As discussed in Section 5 above, there are pairs of dialects of the same language that lack mutual intelligibility, as well as pairs of dialects of distinct languages that share a high degree of mutual intelligibility. There may then be cases in which language-level data do not suffice to convey for example whether a person has a communicative barrier in some situation, or whether some person is capable of translating for two others.

Dialect-related data items can be constructed that parallel those for more general language data. A simple example would be *dialectal competence ICE*: an ICE about a person bearing a competence for a dialect. Axiomatically the person is the bearer of the competence, the competence is a member of the collection of competences that is the dialect, and the dialect is a subaggregate of some language.

## 7. Discussion: ontology of language in philosophy and linguistics

The purpose of this section is to relate the proposed account to the larger, ongoing discussion of the ontological nature of language among philosophers, linguists, and other researchers of language. To give a detailed overview of individual approaches that have been developed and defended would go beyond the scope of this paper. Instead we focus on three categories of approaches to the ontology of language. Each of B. C. Smith (2008), Santana (2016), and Franken (2020) divide approaches to the ontology of language into roughly the same three categories of views, though with slightly different labels. None of the three categories is restricted to a single view or to the views of a single thinker, but instead includes a variety of views espoused by different researchers.

### *7.1. Three categories of views*

One category of views is what Smith (2008) and Franken (2020) label *Platonism*. Santana (2016) refers to this type of view as the *abstract ontology* of language. For these views language is something abstract, akin to the approach some have taken to the ontology of mathematical entities. We have already discussed an example of this sort of approach in Section 3.3 above, when discussing the stance that a language is a collection of GDCs. We do not deny the existence of such collections or that they are among the topics in which language researchers are interested. But they are not instances of OMRSE: *language*.

Next is the category of views labeled by Smith (2008) as the "Cognitive Conception," by Santana (2016) as "the psychological ontology of language," and by Franken (2020) as "psychologism." While the abstract approach to language treats language(s) as external to their users, psychological approaches tend to treat it as something internal to them. Approaches in this category take language to be primarily psychological in nature, and focus on cognitive structures, cognitive capacities, and mental activities. They might for example identify language with a single general language faculty, though of course one could also adopt an approach for which language is an internal and psychological entity while allowing that a bilingual person has or contains two entities of that sort. This would be akin to saying that each instance of OMRSE: *linguistic competence* is a language.

Finally there is the category of approaches that share the common feature of taking languages to be social entities or social objects. Smith (2008: 946) states that for social views, languages are "extrapolations from sets of common practices," and adds that, like the abstract views (Platonism), social views have it that "language is independent of any individual speaker," where for Platonism "language is independent of all speakers." Among social views, the linguistic community and its conventions are emphasized, with language seen as something shared by a community instead of restricted to an individual.

One key difference among these views is that for the abstract views a language is independent of and external to all its users, for the psychological views a language is internal to and dependent upon each individual user, and for social views a language is external to any individual user of its language but is not independent of the community as a whole.

The approach proposed here focuses in part on entities similar to those in the psychological category. We do not take linguistic competences and languages to be the same thing, but we take each instance of OMRSE: *language* to be *a collection of* instances of OMRSE: *linguistic competence*. Thus while *most* of a language is external to any one of its users, some member part of the language is internal to each of its users in the form of a linguistic competence inhering in them. If we add the arguably true assumption that linguistic competences are cognitive, then for our approach a language is a collection of cognitive competences inhering in language users. But since languages do not inhere in individuals, but instead are shared by communities, our account differs importantly from those in the psychological category.

Among the three categories of views considered in this section, our approach shares the most similarities with the social views of the ontology of language. For on our view a language exists independently of any particular one of its users, as the associated community can gain or lose members while the language continues to exist. But the language is dependent upon the linguistic community as a whole, since it is a collection of linguistic competences wherein each inheres in some member of the corresponding linguistic community. A language dies (ceases to be a living language) when the last speaker dies.

Furthermore, our account acknowledges that whether dialects or competences are considered to be of the same language is in part a matter of convention, which in itself might lead some to count our approach as a social view of language.

### *7.2. Mutual intelligibility and the Abstand-Ausbau distinction*

We can relate our approach to Kloss's (1967) distinction between what he calls *Abstand* languages and *Ausbau* languages, meaning "distance" and "expansion" respectively. The former consist of dialects grouped together solely on the basis of some degree of mutual intelligibility; the latter are languages defined to conform to the socio-political facts about what are considered to be the same or different languages. In the sense of *Abstand* languages, to say you and I know the same language suffices to imply we can communicate linguistically with one another in that language, given our current linguistic competences. But when the term "language" is used in the *Ausbau* sense, then to say whether you and I speak the same language may not suffice to convey whether we can communicate linguistically. This is for the very sorts of reasons outlined in Section 5 above, namely: two speakers of the same (*Ausbau*) language may be speakers of mutually unintelligible dialects—as with certain pairs of German or Italian dialects—while two monolingual speakers of different (*Ausbau*) languages such as Danish and Norwegian might share a higher degree of mutual intelligibility.[46]

It is languages in the *Ausbau* sense that we have elucidated in the foregoing. It is worth noting however that our account of *language* does not conflict with the additional representation of similar entities picked out by *Abstand*-oriented approaches to grouping dialects into languages. Tamburelli (2014) suggests that linguistics has become largely "*Ausbau*-centric," but cautions against restricting considerations to *Ausbau* languages, pointing out that "the existence of *Ausbau* languages does not exclude the possibility nor the importance of a linguistic taxonomy in terms of *Abstand* relations."

To be clear, the account of *language* we present here is not intended as prescriptive. Thus in particular it is not prescriptive as regards the proper subject matter of linguistics. Researchers may at times find it

[46] See the 'dialect' entry in Matthews (2003), as well as examples in Section 5.1 of this paper.

useful to employ linguistic categorization schemes that group dialects into *Abstand* languages on the basis of mutual intelligibility alone; and where such a criterion is employed, an approach similar to our proposed account could be applied, including making use of some of the resources presented here. After all, it would concern linguistic competences that are realized in concretization-related processes and that inhere in people who are members of mutual intelligibility-based communities. *Abstand* languages can in principle be ontologically represented alongside *Ausbau* languages, though in the present discussion, and with *language* as elucidated above, our primary concern is the representation of languages in the *Ausbau* sense.

## 8. Conclusion: improvement over the status quo

We believe that the controlled vocabulary presented above is an important improvement over existing resources for representing languages and data about them in ontologies. We have developed terms representing both the data items themselves and what they are about. For example in order to relate a person to that person's preferred language within some given context, we represent an instance of *preferred language ICE* as being about both *that person* and *an instance of language*. We have additionally developed representations of data items about a person's primary language and the languages a person speaks at home. As for how a user of a language relates to that language, we characterize the person as bearing a linguistic competence that is a member of the aggregate that is the relevant language.

We believe furthermore that the ontology of languages and dialects that we have developed can be quite generally useful. The axiomatically interconnected set of terms here presented can be used for annotating data relating to languages, linguistic communities, and linguistic competences and their realizations in a way that reflects key relationships among them and enables semantic interoperability among data sets from disparate systems.

## Acknowledgements

We would like to thank Mathias Brochhausen for his feedback on an early draft of this paper, as well as to acknowledge his role as co-developer of many of the terms presented here. We also thank this journal's reviewers of this paper for their feedback, which prompted beneficial additions and clarifications. A portion of this work was completed as part of the University of Florida's "Creating the Healthiest Generation" Moonshot initiative, which is supported by the UF Office of the Provost, UF Office of Research, UF Health, UF College of Medicine, and UF Clinical and Translational Science Institute. A portion of the research reported in this publication was supported by the University of Florida Clinical and Translational Science Institute, which is supported in part by the NIH National Center for Advancing Translational Sciences (NCATS) under award number UL1TR001427. Portions of this work were conducted using the Protégé resource, which is supported by grant GM10331601 from the National Institute of General Medical Sciences of the United States National Institutes of Health. Smith's research was supported by the University at Buffalo Clinical and Translational Science Institute under NIH award number UL1TR001412. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, or of any other person or organization acknowledged here.

## References


Arp, R., Smith, B., & Spear, A. D. (2015). *Building ontologies with Basic Formal Ontology*. MIT Press.

Bittner, T., Donnelly, M., & Smith, B. (2004). "Individuals, Universals, Collections: On the Foundational Relations of Ontology." In A. Varzi and L. Vieu (eds.), *Formal Ontology in Information Systems. Proceedings of the Third International Conference (FOIS 2004)*, Amsterdam: IOS Press, 2004, 37–48.

Bittner, T., & Smith, B. (2003). Vague reference and approximating judgments. *Spatial Cognition & Computation*, *3*(2-3), 137-156.

Blaisure, J. C., & Ceusters, W. M. (2018). Improving the 'Fitness for Purpose' of Common Data Models through Realism Based Ontology. In *AMIA 2017Annual Symposium proceedings*, 440–447. American Medical Informatics Association.

Brochhausen, M., Bona, J., & Blobel, B. (2018). The Role of Axiomatically Rich Ontologies in Transforming Medical Data to Knowledge. *Studies in Health Technology and Informatics*, *249*, 38-49.

Buccini, A. F., & Moulton, W. G. (2016). Germanic Languages. *Encyclopaedia Britannica*.

Centers for Disease Control and Prevention. (2022, May 11). Language and speech disorders in children. https://www.cdc.gov/ncbddd/developmentaldisabilities/language-disorders.html.

Centers for Medicare & Medicaid Services (CMS). (2014). "Eligible Hospital and Critical Access Hospital Meaningful Use Core Measures Measure 6 of 11, Stage 1, Record Demographics." https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/6_Record_Demographics.pdf.

Ceusters, W., & Smith, B. (2006). Strategies for referent tracking in electronic health records. *Journal of biomedical informatics*, *39*(3), 362-378.

Ceusters, W. & Smith, B. (2015). Aboutness: Towards Foundations for the Information Artifact Ontology. In *Proceedings of the Sixth International Conference on Biomedical Ontology (ICBO)*. CEUR vol. 1515. pp. 1-5.

Chambers, J. K., & Trudgill, P. (1998). *Dialectology*. Cambridge University Press.

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

Cohen, A. L., Rivara, F., Marcuse, E. K., McPhillips, H., & Davis, R. (2005). Are language barriers associated with serious medical events in hospitalized pediatric patients? *Pediatrics*, *116*(3), 575-579.

Dowland, S. C., Diller, M. A., & Hogan, W. R. (manuscript in preparation). On Drawing Lines within Society: Social Categories, Social Identities, and Demographic Data.

Dowland, S. C., & Hogan, W. R. (2022, September). *Using Ontologies to Enhance Data on Intimate Partner Violence* [Paper presentation]. 2022 International Conference of Biomedical Ontology (ICBO 2022): Ann Arbor, MI, United States. https://icbo-conference.github.io/icbo2022/papers/ICBO-2022_paper_6874.pdf.

Eneriz-Wiemer, M., Sanders, L. M., Barr, D. A., & Mendoza, F. S. (2014). Parental limited English proficiency and health outcomes for children with special health care needs: a systematic review. *Academic pediatrics*, *14*(2), 128-136.

Espinoza, J., & Derrington, S. (2021). How Should Clinicians Respond to Language Barriers That Exacerbate Health Inequity? *AMA Journal of Ethics*, *23*(2), 109-116.

Franken, D. (2020). Ontologie der Sprache. In J. Urbich & J. Zimmer (Eds.), *Handbuch Ontologie*.

Gumperz, J. J. (1968). The speech community. In D. L. Sills (Ed.), *International encyclopedia of the social sciences* (Vol. 9, pp. 381-386). New York: Macmillan.

Hicks, A., Hanna, J., Welch, D., Brochhausen, M., & Hogan, W. R. (2016). The Ontology of Medically Related Social Entities: recent developments. *Journal of Biomedical Semantics*, *7*(1), 1-4.

Hoff, E. (2013). *Language development* (5th ed.). Cengage Learning.

Information Artifact Ontology (IAO). http://purl.obolibrary.org/obo/iao.owl.

International Organization for Standardization. (2021). *Information technology—Top-level ontologies (TLO)—Part 2: Basic Formal Ontology (BFO)* (ISO/IEC 21838-2). Retrieved from https://www.iso.org/standard/74572.html.

Karmiloff, K., & Karmiloff-Smith, A. (2002). *Pathways to language: From fetus to adolescent*. Harvard University Press.

Karmiloff-Smith, A. (1986). Some fundamental aspects of language development after age 5. In P. Fletcher & M. Garman (Eds.), *Language acquisition: Studies in first language development*. Cambridge University Press.

Kloss, H. (1967). "Abstand Languages" and "Ausbau Languages." *Anthropological Linguistics*, *9*(7), 29–41.

Koch, P. (2020). A capabilities-based account of wellbeing. *The American Journal of Bioethics*, *20*(3), 85-87.

Koivisto, V. (2018). Border Karelian dialects: A diffuse variety of Karelian. *On the border of language and dialect*, 56-84.

Kunnas, N. (2018). Viena Karelians as observers of dialect differences in their heritage language. *On the border of language and dialect*, *9*, 123.

Landgrebe, J., & Smith, B. (2019). There is no Artificial General Intelligence. *arXiv:1906.05833*.

Landgrebe, J., & Smith, B. (2022). Why Machines Will Never Rule the World. Artificial Intelligence Without Fear. Taylor and Francis.

Library of Congress. (2014). Frequently Asked Questions (FAQ) - Codes for the representation of names of languages (Library of Congress). Retrieved from https://www.loc.gov/standards/iso639-2/faq.html.

Lyytinen, H., Erskine, J., Aro, M., & Richardson, U. (2007). Reading and reading disorders.

Matthews, P. H. (2003). *The Concise Oxford Dictionary of Linguistics*. Oxford University Press.
Merrell, E., Limbaugh, D., Koch, P., & Smith, B. (manuscript in preparation). "Capabilities."
National Association of Community Health Centers. (2016). PRAPARE implementation and action toolkit.
National Patient-Centered Clinical Research Network (PCORnet). (2019). PCORnet Common Data Model (CDM) Specification, Version 5.1. https://pcornet.org/wp-content/uploads/2019/09/PCORnet-Common-Data-Model-v51-2019_09_12.pdf. (Last accessed on October 1, 2020.)
OBO Foundry. (n.d.). "Principles: Overview." http://www.obofoundry.org/principles/fp-000-summary.html.
Ontology of Medically Related Social Entities (OMRSE). http://purl.obolibrary.org/obo/omrse.owl.
Palander, M., Riionheimo, H., & Koivisto, V. (2018). Introduction: Creating and crossing linguistic borders. *On the border of language and dialect*, 7-15.
Santana, C. (2016). What is language? Ergo, an Open Access Journal of Philosophy, 3.
Smith, B. (1999). Agglomerations, in C. Freksa, and David M. Mark, eds., *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science* (Springer Lecture Notes in Computer Science 1661), 267–282.
Smith, B., Ashburner, M., Rosse, C. et al. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, *25*(11), 1251-1255.
Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., ... & Rosse, C. (2005). Relations in biomedical ontologies. *Genome biology*, *6*(5), 1-15.
Smith, B. C. (2008). What I know when I know a language. In E. Lepore & B. C. Smith (Eds.), *The Oxford Handbook of Philosophy of Language*.
Tamburelli, M. (2014). Uncovering the 'hidden' multilingualism of Europe: An Italian case study. *Journal of Multilingual and Multicultural Development*, *35*(3), 252-270.
U.S. Census Bureau. (n.d.). About Language Use in the U.S. Population. https://www.census.gov/topics/population/language-use/about.html.
UCLA Center for Health Policy Research. (2022). California Health Interview Survey: CHIS 2022 Questionnaire Topics. Available: https://healthpolicy.ucla.edu/sites/default/files/2023-05/chis-2022-questionnaire-topics-source.pdf.
Weinreich, M. (1945). Der YIVO un di problemen fun undzer tsayt. *YIVO bleter*, *25*(1), 3-18.