1. Introduction

After a long period in which the study of the evolutionary origins of the human language capacity was avoided, some linguists now speculate and/or use computational modeling to try to infer the evolutionary process that led to modern human language. It is assumed that the evolution of language represents an increase in some sort of linguistic complexity: “pre-language” was less complex than modern human language in some ways, and—probably gradually—acquired the sort of complexity that modern human languages display.

Complexity in language is frequently measured in two ways: structural complexity of communicative signals, and the social-cognitive complexity of the interactional situations in which language or language-like communication is used. Some linguists are now attempting to measure structural complexity of contemporary human languages, after a period in which it was assumed that all modern languages were equal in overall structural complexity. There are undoubtedly significant differences among contemporary human languages in terms of the obligatory expression of certain grammatical semantic categories, and the formal morphological complexity of that expression. Nevertheless, all attested human languages display similar degrees of structural complexity in “design features” of language (Hockett 1960). These design features include the combinability and recombinability of meaningful units; some kind of hierarchical structure; recursion (but see Everett 2005, 2009), and duality of patterning.
Many of the design features of modern human languages, which presumably evolved from a lesser degree of complexity, are not specific to human language, but rather iconically reflect certain aspects of human cognition, which may or may not be specific to our species (see also Schoenemann 1999). In fact, even the categories of obligatorily expressed grammatical inflections and constructions, although not always obligatorily expressed in every language, are the result of the processes by which we verbalize our experience (Chafe 1977a, 1977b; Croft 2007a); these processes will not be discussed here for reasons of space.

Morphosyntactic complexity largely reflects conceptual complexity. Kirby (2002) argues that in an iterated learning model of language evolution, a grammar made up of recombining, hierarchical, and recursive units can emerge, given only a predicate-calculus-like semantic representation and a preference for an iconic mapping between form and meaning. But it is really the semantic representation and semantics-morphosyntax mapping that brings about the characteristic syntactic structures of modern human languages. The cognitive or conceptual capacities that lead to these morphosyntactic structures are an ability for conceptual analysis and compositionality; conceptual grouping (Langacker 1997); and a preference for diagrammatic iconicity in symbolic (form-meaning) mapping (Bybee 1985; Haiman 1980).

Recursive grammatical structures, another type of structural complexity found in most, if not all, modern human languages, commonly occurs in two semantic contexts. Noun-phrase and adpositional-phrase recursion is found in nested figure-ground relations specifying spatial locations (and metaphorical extensions thereof), e.g. the bowl in the cabinet above the stove. Sentence recursion is found in certain constructions when events
are functioning as modifiers (relative clauses), arguments (verb complements), and adjuncts (adverbial clauses). In particular, mental space builders (Fauconnier 1985) such as believe, want and hope introduce states of affairs in the “built” mental space—the space of beliefs, desires, hopes, etc.—and these are usually expressed structurally in terms of embedded clauses. So the emergence of recursion presupposes the emergence of these relatively complex cognitive structures. There are in fact other means of packaging these conceptual structures in sentences that do not necessarily involve embedding. But presumably the conceptualization of displaced experience and of other people’s minds is a basic human cognitive ability (see Corballis 2011, esp. ch. 7-9.) Also, it should be noted that recursion is not necessarily infinite, as assumed in formal syntactic theories. Spoken language rarely exceeds two levels of embedding (e.g., Croft 1995, 2007b). So the emergence of syntactic embedding may be gradual.

On the other hand, duality of patterning does not have a basis in the more-or-less iconic expression of cognitive or conceptual structures. Duality of patterning is defined here as the fact that utterances are made up of combinations of sounds and simultaneously made up of combinations of symbolic units (see Mufwene 2013 for discussion of the different ways that duality of patterning has been defined). Duality of patterning is presumably a response to selectional pressures brought about by the expansion of the inventory of signals. Unfortunately it is unclear at what point in language evolution the increase in the number of signals might lead to duality of patterning. I argue in §4.5 that language probably began as a restricted, domain-specific communication system but then evolved to a general-purpose communication system. A general-purpose communication system has to have the ability to express any situation
that arises and any type of entity that comes along. But duality of patterning is a response to the means to express a large number of situations. Duality of patterning could have emerged at a pre-language stage lacking word combination (a one-word stage). Phonological recombination would facilitate production (and remembering) of a larger vocabulary. But duality of patterning could also have occurred at a stage when multiword utterances occur. Multiword utterances reduces the need for a larger vocabulary, but the continued expansion of language to a general-purpose communication system still requires a large vocabulary, and perhaps at this point duality of patterning emerged.\(^1\)

I have argued here that at least some elements of the structural complexity of modern human languages are the consequence of the cognitive complexity of the conceptual structures being communicated. The leap from conceptual structures to linguistic structures is not an automatic step. The dependence of linguistic structures on conceptual structures only means that the relevant cognitive/conceptual structures must be in place for the formal linguistic structures to emerge. There still remains the mechanism for conceptual structures to be verbalized in language or a pre-language communication system. That is, there has to be a selectional pressure to lead humans to express concepts publicly, with something like language. That selectional pressure is provided by language’s supporting role in the achievement of joint action, which is the foundation of human culture and society. It is only in its social interactional context that the evolution of linguistic complexity can be understood. (Indeed, some might argue that selectional

\(^1\) It has been argued that individual speakers do not have that large a vocabulary; e.g. Cheng (2000) has suggested that individuals have knowledge of around 8000 words. A speech community’s lexical stock is quite a bit larger, but it may imply that there are other selectional pressures leading to duality of patterning than simply the communicative range of a general-purpose communication system.
pressures from social interaction led to the evolution of the conceptual structures described above, as well as leading to the linguistic structures to communicate them.) That is, the evolution of social-cognitive complexity (in terms of joint action) is a prerequisite for the evolution of structural complexity of linguistic signals.

The nature of joint action and language’s role in joint action is described in §2. Joint action presupposes a rich and complex social structure, even if it is “merely” the most basic structure for the emergence of human culture. In §3 I offer some cautionary examples to suggest that the evolution of linguistic complexity may be both gradual and slow. In §4, we look at the nearest sort of communicative system to language whose evolution is documented, namely the evolution of semasiographic systems. Commonalities among the evolutionary paths of different semasiographic systems provide further suggestions about the evolutionary path of modern human language. In §5, I offer speculations as to how modern human language may have gradually evolved, based on the observations in §§2–4.

2. Language as joint action

Language is fundamentally a joint action. Actually, language is just a part of the process of joint action, but it is language as joint action that determines most of the fundamental design features of language. To describe language as joint action is another way of saying that language can only be properly understood in its social interactional context. The nature of language follows from the part it plays in joint action.

The structure of joint actions has been analyzed by social psychologists such as Clark (1996) and philosophers of action such as Bratman (1992); I follow Bratman’s analysis here. Loosely, what makes a joint action joint is that it is more than just the sum of
individual actions performed by separate persons; in particular each individual involved
must take into consideration the other individual’s beliefs, intentions and actions in a way
that can be described as cooperative. It is this level of cooperation, and the features it
entails that are described below, that appears to be a distinctive characteristic of human
behavior and has allowed for the emergence of human culture and society (Tomasello
2008). Although great apes engage in some degree of sharing, reciprocity and
collaboration, it appears to be only the first steps beyond essentially competitive social
interactions (see Tomasello 2011; Tomasello and Vaish 2013, and references cited therein).

Human beings of course engage in all sorts of actions, including actions that involve
other human beings. However, an action involving more than one person must satisfy
certain conditions in order to count as a joint action—in Bratman’s terms, shared
cooperative activity. Each participant, for example each of two people performing a
musical piece, intends to perform the joint action, that is, their intention extends beyond
just their own individual actions as part of the joint action. The joint action is intended by
each person to be performed in accordance with and because of their meshing subplans:
that is, the overall intention of performing a musical piece requires each individual to
have subplans for contributing to the overall joint action that mesh with the other’s
subplans. Neither individual is coercing the other; the cooperation must be voluntary.
Each participant is also committed to mutual support: that is, if the other participant
falters in some way, the first participant will step in to help the other perform her part and
thereby allow the joint action to be successfully achieved.
Two other conditions are essential. All of these intentions and commitments on the part of the participants are shared knowledge among them, or as I will say following Clark (1996), it is part of the participants’ common ground. Finally, in the execution of the joint action, there is mutual responsiveness of the participants. That is, in executing the joint action we will coordinate our individual actions in order to ensure that they mesh with each other in execution and hence the joint action will be successfully carried out (to the best of our abilities).²

Coordination is essential in carrying out joint actions successfully. Yet coordination of our individual actions to achieve a joint action is a problem, because of another fundamental characteristic of human beings: we cannot read each other’s minds; we can only interpret the external behaviors of others. This is not always a bad thing, as a moment’s thought will demonstrate. But much of the time, we want or need to carry out joint actions, so we must find a way to overcome this problem. The means to solve coordination problems are coordination devices (Lewis 1969).

There are a number of coordination devices that people use to solve coordination problems. One of the most effective coordination devices is communication: we communicate to each other our intentions, thereby making them part of our common ground and allowing us to carry out our joint action. What precisely is communication? According to Grice, communication happens when one of us recognizes the other’s intentions. Communication is itself a joint action, however, as its Latinate etymology (commūnicāre ‘to make common’) implies. One person has to signal her intention and the other person has to recognize the intention. In other words, communication solves one

² The concept of coordination in joint action is unrelated to the concept of coordination in syntactic structure.
coordination problem, but poses another coordination problem in order for communication itself to succeed.

Fortunately, there is an effective coordination device for communication, although it must satisfy specific conditions in order to be usable. This is convention, which has been analyzed by both Lewis (1969) and Clark (1996). The Lewis and Clark definition of convention has five parts: (i) a regularity in behavior, (ii) which is partially arbitrary (iii) and is common ground in a community (iv), that serves as a coordination device (v) for a recurrent coordination problem. For example, shaking right hands is a regularity in behavior that is partially arbitrary (one could shake left hands, or kiss on the cheek, etc.), and it is common ground in the American community that shaking hands serves as a coordination device for the recurrent coordination problem of greeting someone.

Many human behaviors are conventional. In addition to conventions for greeting someone, there is the convention of driving on the right, or left, side of the road and the convention of paying for a meal before, or after, eating it. And language is a vast inventory of conventions. Every word and construction is a regularity in behavior—producing those words and constructions in utterances—that solves a recurrent coordination problem, namely communicating a particular intention or meaning. So language involves conventions for communication for the achievement of joint actions.

Powerful as convention is as a coordination device, it is not enough. Convention can work only under certain circumstances. A convention must become established in a community, that is, become common ground in that community. Hence it cannot be the first use of this coordination device (the particular linguistic form) for this coordination problem (the particular meaning intended). For a convention to be established, the
behavior must be used first to solve the coordination problem via nonconventional coordination devices (Lewis 1969).

One type of nonconventional coordination device is precedent: a behavior serves as a coordination device because it served before and was successful. For example, a new word was coined by a commentator in *The Economist* in its Lexington column in the 10 May 2003 issue: ‘Laugh at Bill Bennett, the erstwhile VIRTUECRAT, but don’t forget his message…’ Assuming that the reader understood what the commentator meant, he uses the new word again later in the column: ‘Who needs satire when you have the social conservatives?…Now Bill Bennett, the capo di tutti capi of the VIRTUECRATS, has been caught…with his hand glued to the slot machines of Las Vegas and Atlantic City’.

Precedent is not yet convention. It operates on the principle of ‘Hey, it worked once, let’s try it again!’.

Although it is logically possible that precedent does not even require shared knowledge of the precedent between the individuals, the definitions of precedent found in the literature on coordination devices (Lewis 1969:36; Clark 1996:81) are formulated as shared precedent: it worked for us once, so it might work for us again. I believe that the concept of shared precedent is the one that is relevant to human behavior and human evolution (see §5).

Coordination via precedent, of course, requires the precedent, and that precedent has to succeed via some other coordination device. And at this point we reach the grounding of all joint action: joint salience, made possible by the existence of joint attention among human beings (Tomasello 1999). For example, while I was driving in the rain a few years ago, my mother, who was a passenger, said ‘You need more wiper’—a novel use of the count noun *wiper* in a mass noun construction, not a convention of English. Her novel
coordination device (more wiper) succeeded because our joint attention was focused on the contextually salient fact that the windshield was becoming obscured because the wiper wasn’t going fast enough, and on our shared knowledge that one can adjust the speed of the wiper.

Joint salience is required for all utterances, for at least two fundamental properties of utterances (Clark 1996). First, almost every word in every utterance has multiple senses, leading to denotational ambiguity (if this is doubted, just examine a comprehensive dictionary like the Oxford English Dictionary), and that ambiguity has to be resolved ‘in context’, that is, via joint salience. Second, almost every utterance is communicating a specific event happening to a specific person or thing: I broke the teapot is about a specific speaker, a specific teapot, and a specific event at a specific time (expressed by the past tense verb form). Yet the words used—I, break, teapot—are linguistic conventions for general types, not specific tokens of the type. Using a word for a general type to refer to one specific individual and not another only succeeds ‘in context’, that is, via joint salience.

Linguistic convention is not yet enough in another way. Words and constructions must be realized in a jointly perceptually salient form: auditory, as in speech, or visual, as in signed language or writing. Only at this point is language fully grounded in joint salience. Thus, the full joint action using language has four levels (Clark 1996) (the paired actions are a reminder that all levels are joint):

• proposing and taking up a joint action, via

• signaling and recognizing the communicative intention, via

• formulating and identifying a conventional proposition, via
producing and attending to a perceivable utterance.

Central to joint action and to convention, including linguistic convention, is common ground: shared knowledge, beliefs and attitudes of members of the community. Common ground requires a shared basis (Clark 1996). Some common ground is shared between persons because of common experiences in their face-to-face interaction: things they have seen or done together. The shared basis for that common ground is joint perceptual salience. Other common ground is shared between persons because of what they have communicated to each other over the course of their interactions, even if it wasn’t experienced together. The shared basis for that common ground is the actional basis—their communicative acts, including linguistic acts. Like the shared basis of joint perceptual salience, the actional basis of shared communication requires direct interaction between the persons. Clark calls common ground founded on perceptual and actional basis personal common ground.

A third basis for common ground does not require face-to-face interaction. We share some knowledge simply by being members of the same community: chess players, birders, environmentalists, Mormons, linguists, University of New Mexico employees, etc. By virtue of membership in these communities, we have shared expertise that serves as a communal basis for common ground. For instance, as chess players, we can assume knowledge of the rules of chess and many strategies and moves in playing the game. But that is stating it a bit too glibly: where does that shared expertise come from?

Shared expertise is also not a state, but a process as well: the communities we belong to are communities of shared practices. Wenger (1998) defines communities of practice as possessing mutual engagement (i.e., joint action), a joint enterprise (i.e., the purpose
for the joint action), and a shared repertoire (i.e., commonly performed joint actions to carry out the joint enterprise). Shared expertise emerges from communities of practice. Conversely, communities persist and expand via the transmission of shared practices and shared expertise to new members. Even personal common ground is ultimately grounded in a process: the shared experiences and communications between the persons.

Language is joint action, but it is really part of a joint action that extends far beyond what linguists usually think of as language, as illustrated in Figure 1. Two persons wish to engage in a joint action. This requires a shared intention, meshing subplans, absence of coercion, and mutual support. Above all, it requires coordination of the individual actions of the persons. Coordination of a joint action can be achieved at least in part by joint salience. For example, two musicians look at each other to coordinate their playing. The joint action itself, like its coordination, requires common ground, which also emerges from joint salience but also a history of shared practice in the community to which the participants belong. Much joint action, certainly many sophisticated ones, use communication for coordination. Again, joint salience can facilitate communication. But convention, particularly linguistic convention, is highly effective and widely used—though only after joint salience and precedent have allowed the establishment of the convention in the speech community, and only when supplemented by precedent and joint salience. Finally, linguistic convention must ultimately be realized in a perceivable utterance, whose joint perceptual salience allows the linguistic convention to successfully serve its purpose. Language is only one part—albeit a very useful part—serving a much broader process.

3. The evolution of linguistic complexity
The description of joint action and social cognition in §2 describes joint action in modern human social interaction. It looks very much like a process that is tightly integrated, where each element depends on every other element. It seems that once the basic elements of joint action have evolved—common ground, helpfulness, and so on—all the rest will emerge automatically, and modern human language will therefore emerge with it. It is a little easier to conceive of a gradual evolution of the complexity of the signal described in §1 (and perhaps this is why so much attention has been focused on it). But one cannot separate the evolution of the complexity of the signal from the evolution of social cognitive complexity.

Nevertheless, we shouldn’t overestimate how rapidly or suddenly a complex suite of social cognitive skills can evolve. Examples from the evolution of semasiographic systems— which encode information in a lasting, visual medium, of which more in §4—and the ontogenetic development of social cognition show that hindsight is too powerful to form the basis of an evolutionary scenario.

For example, it is universally agreed that humans had fully modern language by fifty thousand years ago, if not much earlier. Around that same time (fifty thousand years ago if not earlier; see for example d’Errico et al. 2001), representational art appears in the archaeological record, including artistically quite sophisticated representations. One might think that putting the two together—auditory human language and visual representation—is not an enormous next step. Yet although all human societies have modern human language and representational art, it took tens of thousands of years before the first writing systems emerged; and the evolution of writing systems happened independently twice or at most four times, out of the thousands of human societies that
must have existed five thousand years ago. Clearly, putting the two together for the first
time was not easy at all. The enormous time lag was probably due to the absence of
selection pressure, which probably arrived in the form of large-scale stratified societies
where writing emerged. Even so, the actual emergence of writing appears to have been a
gradual process where it is documented (see §4.2).

Another far more mundane example illustrates the difficulty and gradualness of
evolution of semasiographic systems, which probably is characteristic of all human
invention. Musical notation, another semasiographic system that will be described in
§4.3, had evolved considerably by the seventeenth century. This included representation
of notes of different time values (longer and shorter), conventionally described as half-
notes, fourth-notes, and so on. By the seventeenth century musical notation represented
irregular note lengths, such as a note lasting for three sixteenths followed by a single
sixteenth note (3+1), by adding a dot after the longer note: in this case, an eighth note
followed by a dot denoted three sixteenths ("one and a half eighth notes"), and it was
followed by a sixteenth note. At the same time, a larger note value could be divided into
three shorter notes instead of two notes or four notes. But it wasn’t until after the middle
of the nineteenth century that a notation was devised for describing a 2+1 combination,
even though composers wrote music with 2+1 note values. Instead the 3+1 notation was
used (for example, in music as late as Chopin’s Polonaise-Fantaisie, published in 1846).
Even though there was selectional pressure to fix this notational problem—composers
produced both the 2+1 and 3+1 note value combinations frequently in their
compositions—it still took some two centuries for it to be solved.
The moral is that even when the cognitive parts are there—that is, they are part of human cognitive ability at that point in human history—putting the parts together to produce a new behavior does not happen automatically. In fact, it may take a long time before it even starts to happen at all. There has to be a reason for it to happen, that is, a selectional pressure in evolutionary terms—and even then, it takes a long time, even multiple human generations, for the new behavior to emerge and be propagated in the society. The selectional pressure is most likely to be social, not purely cognitive.

An example that suggests that social cognitive evolution is also gradual is the ontogenetic development of the human conceptualization of others. Tomasello argues that in ontogenetic development, a “revolution” occurs at around nine months of age, which leads to the ability of the infant to achieve joint attention with other human beings (Tomasello 1999:61-70). But other types of mental reasoning that are built on the notion of other minds having beliefs other than our own (called the “theory of mind”) take several further years to develop (e.g., Tomasello, Kruger and Ratner 1993:499-500).

The conclusion we can draw from these various examples is that it is eminently plausible to infer that the evolution of social cognition as described in §2 took some time, possibly a very long time—hundreds of thousands if not some millions of years. Tomasello suggests that although primates and some other animal species appear to engage in joint cooperative activity, the apparently “coordinated” behavior may be due to individual responses to causal inferences based on practical reasoning about conspecifics’ intentions and goals (Tomasello 2008:44-49, 173-85). Hence, it may have taken up to five million years (from the branching off of the hominin line to no later than fifty thousand years ago) for modern human social cognition to have evolved.
Unfortunately, there is no direct empirical evidence regarding intermediate stages of complexity in human cognition between primate group activities and signals and modern human joint actions and language. This is the reason that scholars avoided the question of modern human origins until recently. In the next section, I will look at the evolution of semasiographic systems, the nearest language-like systems for which we have direct empirical evidence, to see what they might tell us about stages in the evolution of language (cf. Coupé 2012, who compares the development of language to the development of photography). After taking a glance at language acquisition I will speculate on the evolution of language from pre-language in §5, using the logical priority of different elements of the model of joint action described in §2 in combination with the insights from the evolution of semasiographic systems.

4. The evolution of semasiographic systems

Semasiographic systems are used for ‘the communication of relatively specific ideas in a conventional manner by means of permanent, visible marks’ (Boone 1994:15; 2004:313). Semasiographic systems include images, numerals (counting), maps, calendars, bookkeeping systems, tables, mathematical notation, music and dance notation, graphs, and as a special case, writing. As noted above, semasiographic systems have been around for at least fifty thousand years, with the first production of images and statues. Semasiographic systems are similar to language in several respects. As noted above, they encode information, as linguistic expressions do. More specifically, they encode information in such a way that it can be shared, just as language involves sharing individuals’ intentions. Semasiographic systems employ, or evolve to employ, relatively complex information encoded by recombinable units, also like modern human language.
Most useful for the goals of this chapter, semasiographic tokens survive and leave a record, allowing us to observe stages of their evolution in at least some cases. In this section, I will briefly describe salient properties of the evolution of number systems, writing systems, musical notation, and dance notation, and offer some generalizations over the evolution of these semasiographic systems.

4.1. Numbers

Number appears to be the earliest semasiographic system after images. Artifacts from ten to twenty thousand years ago, such as the Ishango Bone (Marshack 1991:22; Rudman 2007:62) appear to involve tallies, the earliest type of number notation. In tallies, the number of repeated symbols (such as strokes or notches) denotes the equivalent number of entities. That is, tallies represent an additive number representation system (Pettersson 1996:796). Additive number representations are essentially iconic: the number of signs reflects the number of objects being counted.

Another example of an additive numeral system is represented by the undecorated tokens found in Middle Eastern sites dating back to 8000BC and apparently used to count goods (Englund 1998:46). A later development is the employment of distinct signs for larger numbers, such as 10 and 60, which effectively function as bases in the numerical semasiographic system. This development occurs in Mesopotamia at around the same time as proto-cuneiform around 3200-3100BC (see below; Nissen et al. 1993:25). Pettersson (1996:796) calls these sign-value systems, the earliest of which are also additive, repeating each sign to express multiples of the sign value. Familiar examples are the later Roman numerals, such as XXXIII for 33. However, the signs for each value are

---

3 Claims that the marks represent more complex number notation are highly speculative (Rudman 2007:62-65).
largely arbitrary: although signs for larger values are sometimes larger in size, or require more strokes, than signs for smaller values, the actual numerical value denoted by the sign is arbitrary.

One striking fact (to modern eyes) about the Mesopotamian sign-value notation is that the numerical signs varied depending on what is being counted or measured, and the same sign may have a different numerical value depending on what is being counted or measured (Nissen et al. 1993:25-27). For example, the sign resembling a dot represents 10 of the smaller units for discrete objects, 6 units for dry measures of cereals, and 18 units for surface measures (ibid., 131). This is an instance of the high degree of context-dependence that is commonly found in the earliest stages of semasiographic systems.

A still later development of number notation is a positional system (Pettersson 1996:796), in which the position of the sign determines its value by specifying a base amount that it multiplies. In a sign-value system, one must keep creating new symbols for larger quantities, but in a positional system, the only symbols necessary are the values up to the base (e.g. 10 in the modern decimal system); the position of the symbol determines the magnitude of the number. In Mesopotamia, a positional system evolved around 2000BC (Pettersson 1996:798)—several thousand years after the appearance of undecorated tokens, and over a thousand years after the emergence of a sign-value notation system. The counting board or abacus, which dates back at least to seventh century BC Greece (Menninger 1969:301), and the quipu system of knots, used in the Inca empire (15th century AD; Menninger 1969:252), also use the positional method to represent number. (Interestingly, the contemporary ancient Greek written numerals did not make use of the positional system found in its counting boards.)
The evolution of Mayan numeral notation also apparently proceeded through the same stages: additive, then sign-value (with different glyphs for powers of 20), then positional (Rudman 2007:118; he uses the term ‘additive’ for sign-value systems). The Ancient Greeks used a sign-value system adopting written letters for signs (Pettersson 1996:803), illustrating another characteristic of semasiographic evolution that we will encounter in other systems, namely, the adoption of signs from one semasiographic system for use in another semasiographic system. Hindu-Arabic numerals appear not to have been adopted from another semasiographic system. The earliest system (ca. 200BC) was partly additive (1-3 were one to three strokes), and partly sign-valued; a positional system emerged around 600AD (Pettersson 1996:804). Again, the modern Hindu-Arabic numeral system is the result of a gradual evolutionary process.

4.2 Writing

Writing essentially involves the adaptation of a semasiographic system in order to represent spoken language. Writing evolved independently at least twice. Undoubtedly independent are its evolution in Mesopotamia and Mesoamerica. Egyptian writing evolved at the same time as Mesopotamian writing. It was earlier thought to be inspired by Mesopotamian writing, but its structure is so different, and the timing so uncertain, that it is now thought more likely to have evolved mostly if not entirely independently (Baines 2004:175-76). The extant fragments of earliest Chinese writing—oracle bone inscriptions and clan names on bronze vessels from ca. 1200BC—are a full-fledged writing system in structure (Bagley 2004:198). There are no clear precursors in the archaeological record (ibid., 190); so the early evolution of Chinese writing is unknown.
It is possible, though now considered less likely, that Chinese writing developed via diffusion from the Near East.

By far the best documented evolution of writing is from Mesopotamia, because writing was mostly done on clay. The emergence of the writing system probably ultimately began with undecorated stone tokens referred to above. By 3300-3200 BC these undecorated tokens were enclosed in clay envelopes impressed with seals, presumably used for accounting purposes (Englund 1998:48). Seals were semasiographic representations of specific persons (or institutions such as a temple), that is, they denoted names of individuals. Thus the clay envelopes combined two semasiographic systems, the tokens counting goods and the seal impression designating the person or institution responsible for the goods. This is an example of a multimodal semasiographic system.

Some of the clay envelopes were also impressed with signs representing the number of tokens in the envelope and their form. There also occur numerical tablets, tablets with numeral signs (and seal impressions) but not enclosing tokens that they designate. These devices—sealed clay envelopes containing tokens, envelopes containing tokens but impressed with the number and type of tokens, and numerical tablets—are attested at around the same time and may reflect either a sequence of developments or alternative systems for accounting for goods. The envelopes bearing external signs illustrate another characteristic in the emergence of semasiographic systems: the often direct association of the signifier with the signified.

The next step occurred around 3200-3100 BC (the Late Uruk period and Uruk IV writing phase), in both Uruk and Susa (the seat of the ancient Elamite kingdom). In this period are found tags—small tablets—with ideographic signs that may represent personal
names (Nissen et al. 1993:20), though some ideographs may denote the goods (Englund 1998:57). In addition, there are numero-ideographic tablets, containing seal impressions, numeral signs and ideographic signs that designate different types of goods (Englund 1998:51-53). The numero-ideographic tablets mark the beginning of proto-cuneiform. This step involved the combination of a largely pre-existing set of numerical signs combined with a largely pre-existing set of ideographic conventions (Cooper 2004:77). The number of ideographic signs quickly expanded to some 900 signs (Englund 1998:68; Cooper 2004:68), while the context-sensitive system of sign-valued numerals described above constituted around 60 signs. The proto-cuneiform signs are almost all pictographic, that is, iconic, in nature (Englund 1998:71) or at least iconic-indexical (i.e., pictographic representation via metonymy and synecdoche; Cooper 2004:84), although some are abstract representations that had been in use for some time.

It is doubtful whether the proto-cuneiform tablets represent “writing” in the sense of reflecting in graphic form a spoken language. Both numerals and ideograms do not denote words in a specific language in the way that a writing system that records phonetic content, even only partly so, does. The proto-cuneiform tablets do not match spoken language syntax but ‘the “grammar” of the archaic accountants’ syntax’ (Englund 1998:63; 79). The information is structured instead by the shape of the tablet, the

---

4 Citing Buccellati (1981), Cooper (2004:77) considers the ideographs to derive from glyptic art. Schmandt-Besserat (1996) argues that the proto-cuneiform ideographs derive from decorated tokens, which contrast with the undecorated tokens referred to above. However, the decorated tokens are not found enclosed in the clay envelopes whose markings appear to have evolved into the precursors of proto-cuneiform, namely the numerical tablets (Englund 1998:47). Englund suggests that the iconic nature of the proto-cuneiform pictographs and the decorated tokens are instances of convergent evolution, rather than derivation of the former from the latter (Englund 1998:47 [fn. 92], 53).
ordering of cells in the tablet, and other nonlinguistic properties (Cooper 2004:80-81). In other words, writing originated as a semasiographic system for purposes other than to record spoken language:

…no writing system was invented, or used early on, to mimic spoken language or to perform spoken language’s functions. Livestock or ration accounts, land management records, lexical texts, labels identifying funerary offerings, offering lists, divination records, and commemorative stelae have no oral counterparts. Rather, they represent the extension of language use into areas where spoken language cannot do the job. (Cooper 2004:83)

One did not find written royal inscriptions in Mesopotamia until around 2700BC, in literature until around 2600BC, and in letters until around 2400BC (Cooper 2004:83). In other words, the semasiographic systems that evolved into writing originated in specific domains of use, such as the accounting functions of early Mesopotamian writing. The signs then came to be construed as denoting word-like concepts (the semasiologographic systems of Trigger 2004:47), and eventually words in a specific language. Thus, it is not surprising that proto-cuneiform did not have a language-driven syntax, and did not express grammatical morphemes. Indeed, it has been suggested that perhaps proto-cuneiform was not language-specific because ancient Uruk was a polity inhabited by speakers of different languages, although the Sumerians were the dominant group at first (Englund 1998:73-81). It was not until around 2400BC that the ‘spoken language determined the order of the script’ (Nissen et al. 1993:123), and it took until the end of
the third millennium BC for grammatical morphemes to be written (ibid., 116). ‘Full writing was the byproduct of a gradual discovery of new applications’ (Bagley 2004:233).

Over time, the signs evolved from a more iconic to a more arbitrary symbolic form (that is, no longer iconically representing the concept denoted by a word, not its phonetic form). This was in part due to the Mesopotamian writing technology, namely the use of the stylus on soft clay, which made it difficult to draw curved lines.

The origins of writing in Egypt have some similarities to and some differences from the origins of writing in Mesopotamia (Baines 2004; see also Stauder 2010). Its earliest known uses do not seem to be entirely economic. The earliest known ideographic objects are bone tags found in Tomb U-j in the Umm al-Qa’ab cemetery near Abydos (no later than the Naqada III era, ca. 3200BC; Baines 2004:153, 154). The tags bear numerals or ‘representationally based signs’, but normally not both at once, unlike the numero-ideographic tablets of contemporary Mesopotamia (ibid., 157). The tags were probably attached to goods, that is, the signifier is associated with the signified. The tags (apart from the numeral tags) are very likely to denote names, perhaps of prestigious entities (ibid., 164). A small number of pictographic (iconic) signs are attested, mostly different from later hieroglyphics. There are also jars, probably containing valuable oils or fats, with large painted signs similar to those on the tags (but no numerals; ibid., 157, 159). There is nothing language-specific about this semasiography, not unlike proto-cuneiform. The slightly later Hunters’ Palette (Naqada IIIa) and Scorpion Macehead (Dynasty 0) have similar signs (ibid., 168-69). In Dynasty 0 (before the beginning of the third millennium BC) the number of signs greatly expands to over a thousand, again not unlike
proto-cuneiform, before being consolidated to a few hundred by the middle of the third millennium (ibid., 172). In Dynasty 0, one finds larger tags that combine pictorial representation and linguistic (or at least semasiolographic) forms (ibid., 173)—another sort of multimodality. The function of the writing system expanded to other administrative functions in the 1st Dynasty (ca. 2950-2775BC), but ‘continuous discourse and full syntax were not notated for another couple of centuries’ (ibid., 174), again like Mesopotamian writing.

As noted above, little is known about the early evolution of the Chinese writing system. The earliest writings are already linguistic in the way that the earliest Mesopotamian, Egyptian and Mesoamerican systems are not. For example, the signs are fairly arbitrary in form, although more iconic than later versions of the signs; literate Chinese speakers retain an awareness of their original iconicity (Cooper 2004:89).

Mesoamerica represents an undoubtedly independent origin of writing systems, but the earliest history of writing in Mesoamerica is very complex and fragmentary. Part of the complexity is that the best-documented writing system, that of the Maya, is in fact a later adaptation of earlier writing system(s) used for other languages. All of the earliest signs have resisted decipherment, or their decipherment is highly controversial. The earliest occurrence of glyphs that look like later writing signs are found on Olmec carved heads dating from 1150-900BC (Houston 2004:288). The glyphs on the heads probably represent names of the individuals represented by the heads (ibid., 289). The direct association of signifier and signified was not decoupled until 900-600BC, when a stela is found with the glyph next to, not directly on, the headdress of the head (ibid., 290). By ca. 500BC there is a monument, La Venta Monument 13, which contains writing, defined
by Houston as representing language, detached from its signified, and occurring in linear sequences (ibid., 292). Nevertheless, the writing occurs on stelae denoting individuals who are probably named by the writing, as in the case of San José Mogote Monument 3 (ibid.), where a captive appears to be named by a calendrical date. This last monument illustrates a common Mesoamerican association of numerical, specifically calendrical, information with a logographic (or semasiologographic) system—not unlike the origins of Mesopotamian writing in a combination of numerical and semasiologographic signs (viz. the numero-ideographic tablets). The Isthmian (or “Epi-Olmec”) texts from 36BC-162AD are also associated with calendrical information (hence the precise dating; ibid., 296-97).

The Mayan scripts are adapted from the earlier Mesoamerican writing systems for the Mayan language(s). The scripts evolved over time, which makes decipherment of the fragmentary earliest texts difficult even though much is known about the more widely attested Classic Maya writing system. The earliest Mayan texts are lists of deities and name tags (ibid., 304), not unlike the earliest Mesopotamian and Egyptian systems. Again, grammatical morphemes are not encoded until the beginning of writing phase 1B (150AD; ibid., 305), around 250 years after the earliest Mayan texts.

We can make some general observations about the evolutionary emergence of writing from the origins of Mesopotamian, Egyptian, Chinese and Mesoamerican scripts. The emergence of writing is a gradual process, both in the transition from pre-writing semasiographic systems to writing as a representation of spoken language, and in the further development of writing to express grammatical structure as well as words (syntax
and grammatical morphemes). The signifier is often directly associated with the signified, or a representation thereof (e.g. the image of a ruler and his name).

Pre-writing is often multimodal, involving some combination of numerals, ideograms, seals and pictorial representations. Some of this multimodality may have been a trigger for the graphical representation of language, that is, the combination of discrete linguistic units. In pre- or proto-writing phases, the encoding of proper names appears to have been an important step. Proper names are indexical in function (picking out an individual) but do not operate via deixis. The earliest ideographs and logographs were pictographic, that is, iconic, or iconic-indexical (via metonymy and synecdoche). Even the development of phonetic representation originates in iconic signs metonymically via the rebus principle (a sign originally standing for a concept is extended to represent the associated phonetic form of the word denoting that concept). The form of signs later becomes more arbitrary. Another aspect of pictographic representation for language is that it adapts the signs of another semasiographic system, namely that of images. In this context it is not surprising that later writing systems emerged by adapting existing signs from writing systems that originally evolved for other languages, even using signs for different phonetic values than the originals, such as the Greeks adapting consonant signs to represent vowels.

The domains of use for writing were at first very limited, often for centuries, even though spoken language of course is a general-purpose communication system for all sorts of joint actions. This is because writing originates to solve specific information storage and communication problems, not to faithfully reflect spoken language. Even when the domains of use of writing expanded to include narrative, writing left much of the language out. Writing did not express grammatical elements until centuries after its
first emergence. In other words, substantial common ground between author and reader was required to interpret just the linguistic form encoded by early writing (this is still true of abjads, a writing system that indicates only the consonants; one must know the language to fill in the vowels).

4.3. Musical notation

Musical notation appears to have been invented independently or partly independently several times in the Near East and Europe. The earliest known fragment of musical instruction is an Old Babylonian text which gives the incipit of a hymn to Ishtar. The notation names the 6th string of the Babylonian lyre or harp, certain intervals that involve the 6th string, and the mode in which the hymn is to be played (Kilmer and Civil 1986:96). This fragmentary text appears to be a linguistic description of how to play the hymn rather than any specific musical notation.

Later Babylonian texts, from 1250-1200BC (West 1994) are less fragmentary but remain difficult to interpret. The notation was designed for a nine-stringed instrument, probably the bovine lyre (West 1994:166), and for the seven standard tunings of the instrument (ibid., 164). The texts bearing notation are in Hurrian but apparently in a Babylonian musical tradition. The notation consists of names of intervals followed by a numeral sign (ibid., 171). The interpretation of the numeral is disputed, but may indicate the repetition of notes. Nevertheless, we may make some general observations. The notation is adapted from other semasiographic systems—writing and numbers. As such, it is also multimodal. The music is also meant to be sung, and the lyrics are included. This provides another degree of multimodality, because sung music is also multimodal (words and melody). The notation is very domain-specific: it only can represent music for the
lyre, and only in one of the seven modes. Finally, the rhythm and other details of performance are not specified (ibid., 176); that is, common ground about the songs is required to perform it properly.

Musical notation was reinvented by the Greeks by the third century BC, and lasted no later than the fourth or fifth century AD (West 1992:254). It is generally agreed that the musical notation developed gradually. Pitches are notated by letters borrowed from the Greek alphabet for use in instrumental music. Some of the pitch letters are Ionian, but other letters appear to derive from the Argive script, a local Greek script of the sixth and fifth centuries BC (ibid., 261). The notation appears to have originated for the aulos (a flute). Vocal notation based on the Ionian alphabet was developed probably in the late fifth or fourth century BC (ibid., 263). Rhythm came to be notated in some detail in the second century AD, employing a separate set of symbols indicating length (ibid., 268). The history of Greek musical notation illustrates the gradual evolution of the notation over time. It is restricted at first to the aulos. Finally, it again presupposes familiarity with the music itself, and its performance style. The notation was probably used only by professional musicians; and was rarely used at all, as most lyric and dramatic texts lack musical notation (West 1994:270).

Musical notation was again reinvented in Europe in the middle ages. (Apparently Boethius [6th century] was aware of Greek notation using letters for pitches, but his writings did not lead to a similar notation in the middle ages; Taruskin 2005:17.) The earlier medieval musical notation is in the form of staffless neumes from the 9th and 10th centuries. Staffless neumes are basically dots and lines that indicate the overall pitch contour, but without the staves that allow one to identify musical intervals. Nor do
staffless neumes indicate rhythm. Instead, individual neumes appear to have specified particular melodic formulas (by their shapes, which were given Latin names) that were associated with particular chants and modes (Taruskin 2005:22). Hence staffless neumes presuppose much common ground: they do not indicate details of melody or rhythm (let alone other aspects of performance, including polyphonic accompaniment). The user of the musical notation must already know the music, which is still transmitted essentially orally; the notation at best serves as a reminder or an aid (Taruskin 2005:17). Musical notation was restricted at first to sacred music and later extended to elite secular music, namely the courtly songs of the troubadours, trouvères and minnesängers.

Neumes were positioned on a staff starting around the early 11th century (Taruskin 2005:16)—that is, two centuries after the first surviving documents of staffless neumes. This allowed relative pitch (i.e., intervals) to be represented. (Even so, manuscripts with staffless neumes were produced in the monastery of St. Gall until the 15th century.) But rhythm would not be represented until the beginning of the 13th century, another two centuries later (Taruskin 2005:176). Rhythm was first notated for the polyphonic music of Notre-Dame of Paris, by recruiting the neume shapes or combinations (known as ligatures) to express rhythmic patterns instead of melodic formulas. The rhythmic interpretation of ligatures was context-sensitive, depending on the overall pattern of ligatures (ibid., 177).

By this point, polyphonic music was being notated, but there did not exist explicitly notated means of aligning the voices such as the later bar-lines, apart from the iconic visual alignment of voices in the manuscripts. (Barlines first appeared in keyboard notation in the early 16th century; Apel 1953:9.) In the middle 13th century, so-called
Franconian notation was developed for notating the music of the motet style of the time (namely a style in which all voices are equally melismatic, but a text or texts were applied to the melismatic melodies). Again, the existing neume shapes and combinations were recruited, this time for a less context-sensitive way or represented note length and hence rhythm (Taruskin 2005:212-17). While each specific ligature (combination of note shapes and stems) has a unique rhythmic interpretation in Franconian notation, the individual note shapes did not, except in the context of the ligature. Ars Nova notation (14th century) is less context-dependent for its rhythmic interpretation, and the notation developed by the beginning of the 17th is essentially the same as modern notation with respect to rhythm (apart from odd anomalies such as the one described in §3).

Even with the development of more explicit notation of pitch intervals and values, and rhythm, other aspects of musical performance were left to common ground (leading to much debate over the “historically accurate” interpretation of medieval and renaissance music by modern early music performers). From the 17th to the 20th century, notation or articulation, dynamics and tempo were gradually added to musical scores. Much was still left to common ground and an oral tradition, such as how to ornament melodic lines or fill out a figured bass in Baroque music.

While a general notation system for European music evolved over the past twelve centuries, specialized notations continued to emerge. These specialized notations were mostly designed for instrumental music, and hence specific to an instrument. The most widespread example was lute tabulature, since lute was the most popular instrument in the Renaissance (Apel 1953:54-86). Lute tabulature used staff lines, but the staff lines indicate the strings of the lute, not specific pitches. Numbers or letters were adapted to
indicate which frets to use, and notes above the staves were adapted from general musical notation to indicate the rhythm. In other words, lute tabulature was multimodal, drawing from general musical notation, letters, and numbers to indicate strings, frets (which together specify pitch), and rhythm. A similar tabulature was used for German organ music through its flowering in the 17th and early 18th centuries (Apel 1953:21-47).

4.4. Dance notation

Dance notation is not standardized and numerous systems have been invented in the West. Like spoken language and music, dance requires strict temporal sequence (and hence linear order). But dance is far more complex a phenomenon to represent. The notations used in the West vary tremendously; many if not most choreographers invented their own notation. Nevertheless, the types of notation used present a historical sequence.

The earliest known dance notation was devised for the Renaissance basse danse ‘low dance’ (Guest 1989:1). Letters were adapted to indicate the steps to perform (letter notations continued for 200 years; ibid., 4). As with other semasiographic systems, the earliest dance notation was contextually highly restricted (only basse danse steps can be represented by the letters) and required much common ground—they can only be interpreted by those who know the steps to the dance.

The next types of dance notation in Europe (in terms of when it was first devised) were iconic in form. In the 17th century, a track notation was devised, which iconically represented the choreography as couples moved around the floor (Guest 1989:12-27). This notation developed in part because the dances of this period involved such movement. The notation also imagistically represented bodily movements, though it also used arbitrary symbols to indicate men vs. women (ibid.). Again, the notation was
restricted to representing Baroque dances. In the 19th century, another iconic notation was devised, namely stick-figure notation (Guest 1989:28-68). Stick-figure notation, unlike the earlier notation, was less restricted in the types of dance movements it could represent. All of these types of dance notation were multimodal in that the music to which one danced was also notated along with the dance steps and movements.

In the late 19th and early 20th centuries, a music-based dance notation was invented. It adapted musical notation for dance steps: note values indicate time—a dimension of dance not well notated by earlier systems. The notes were placed on different musical staves that indicated different parts of the body. Direction of movement was often indicated iconically, while modifications of the form of the musical note expressed bodily movements, sometimes iconically and sometimes not.

In the 1920s, purely abstract dance notations developed (Guest 1989:102-62). Guest also notes that an earlier abstract system was devised by a choreographer in 1831 but was an isolated invention. Unlike the Baroque and 19th-century notations, abstract dance notation was a completely symbolic, and in some instances anti-iconic, representation of dance movements, although some symbols were iconically motivated in part. On the other hand, the abstract systems are also highly general in representing human bodily movement, timing, and direction of dance—possibly a response to the freer dance styles of twentieth-century dance and the greater variety of movements that they encompassed.

The evolution of music and dance notation exhibits all of the properties of the chronologically prior evolution of number and writing. It was gradual (extending over centuries in the case of Western musical notation); it is multimodal, and adapts signs from other semasiological systems; its domain of use was very restricted at first, before
expanding; the interpretation of some signs was context-dependent; and the notation also required much common ground to interpret, becoming more explicit as it evolved. The chief difference is that the signifier was not directly associated with the signified, except when using musical scores in performance. This difference is presumably due to the fact that these semasiographic systems represent human performances rather than the conceptual structures represented by mathematics and language.

4.5. Some commonalities across the evolution of semasiographic systems

The evolution of different types of semasiographic systems have some common features, which may offer clues for the evolutionary origin of the language capacity. To begin with, the evolution of semasiographic systems, where we have documentary evidence, is gradual and incremental. Contemporary semasiographic systems for number, writing, music and dance are very different from their first notation, and the process leading to contemporary semasiographic systems included many intermediate stages and alternative notations. This observation reinforces the view expressed in §3, viz., that the evolution of social cognition and the language that facilitates joint actions was also gradual and incremental.

The earliest uses of semasiographic systems are very domain-specific and context-dependent. They are used for only a restricted functional domain of social (joint) actions. Numbers were used only for counting and for arithmetical functions, before being elaborated to more and more mathematical functions. Initial uses of writing were for specific communicative acts, and only later came to be used in more and more domains. In fact, only with the advent of electronic media is it increasing use in the most casual conversational interactions. Initial uses of musical notation were restricted to specific
types of music, for specific social or religious functions, or for specific instruments, as in the case of the creation of lute notation. Dance notations were only for specific genres of dance; and even now no general dance notation has come into widespread use, although the latest notations are highly general in their representation of bodily movements. Thus we might expect that language began in highly restricted functional domains, and its extension to become a general-purpose communication system was a long and gradual process in human prehistory.

Another respect in which early semasiographic systems are context-dependent is the minimal expression of what they signify. The earliest writing did not encode grammatical inflections, or all of the sounds of a language. Early European musical notation did not encode exact intervals or rhythms, let alone dynamics. Early dance notation was even more minimal in the information that it encoded. In other words, the interpretation of the semasiographic representation required much common ground; it served to minimally evoke what the relevant information or action actually was. Likewise, language initially functioned simply as a coordination device for joint action. Thus the “meaning” it conveyed was probably minimal—just enough for the (restricted) purpose of coordinating a particular type of joint action. Eventually, language came to be used to share knowledge, that is, to increase common ground in and of itself. At that point, language had to become more semantically specific (although as we saw in §2, language can never be totally “precise”). But this was probably a late stage in the evolution from pre-language to language.

In some semasiological systems, the signifier is often directly associated with the signified. This was the case for Mesopotamian tokens and clay envelopes, which were
associated with the objects they enumerated or denoted; and for Egyptian and Mesoamerican writing, which were often at first names associated with representations of the persons or deities that they signified. The semantic relationship is thus associative at first. However, the situation is somewhat different with performance notation. Musical notation at first was at most a mnemonic for the melodies that it encoded. In later periods, music was frequently performed from notation, but it is not clear if that was true at the beginning. Dance notation is not directly used in performance. Music and dance notations are more likely to be used for teaching rather than for final performance; they serve as a record of how to perform the composition in question. Still, the cases in which the signifier is directly associated with the signified suggests that the earliest uses of pre-language or language was to denote elements of immediate experience, and only later was language (or pre-language) extended to denote displaced or imagined experience, which must be inferred from the semantics of the sign. (This last suggestion is a common speculation about the evolution of language.)

Semasiographic systems are essentially coordination devices: graphic coordination devices. With respect to that function, the coordination devices for one semasiographic system is often parasitic on another semasiographic system. For example, writing symbols were appropriated for Babylonian and Greek music notation, for Greek numerals, Renaissance dance, and arithmetic symbols. Writing systems designed for one language were adapted for another language (Olmec for Maya, Phoenician for Greek, Sumerian for Akkadian, and then Hurrian, Hittite, Persian and other ancient languages; Chinese for Japanese and Korean; Greek for Slavic languages via Cyrillic, and the Roman alphabet for many languages). And images, the earliest semasiographic system,
were adopted for medieval music, for Baroque and nineteenth-century dance, and for the earliest writing and counting systems.

Thus we might expect that the earliest linguistic signals might be adopted from some other communicative system. It has been argued that primate vocalizations cannot be the source of modern human linguistic vocalizations, because primate vocalizations and their “meanings” (alarm calls or whatever) are innate. While modern human language definitely is not innate in the specific linguistic structures and vocalizations employed—it is extremely diverse across speech communities—it could still be the case that at some point in the evolution of pre-language, vocalizations or manual and facial gestures were adapted from some other function, including perhaps even innate vocal or other gestures. (This is the process known as exaptation, a phenomenon well attested in contemporary language change; see Lass 1990; Croft 2000.)

Semasiographic systems also appear to evolve to a (more) arbitrary symbol system. Again, this reflects a common speculation about the evolution from pre-language to language. Pre-language may originate with indexical or iconic gestures or vocalizations, before being supplemented with more arbitrary symbols. Of course, modern human language retains indexical and iconic expressions, along with more arbitrary symbolic forms. Indexicals form a central body of linguistic expressions in modern human languages: demonstratives are among the first words to be learned, are very ancient (not clearly derived from other words), are central to structuring discourse, and grammaticalize into ubiquitous grammatical functions, including definiteness, reference, and subordination (Diessel 2006). Diagrammatic iconicity is exhibited by the great majority of syntactic structures, and many signed language gestures are iconic.
In terms of the complexity of the signals themselves, the signs in the earliest semasiographic systems are frequently multimodal. Early writing combined symbols for numbers and for objects. Medieval and later music combined words and notes (pitches); later music combined notes for pitch and duration with other types of symbols for dynamics and articulation. All dance notation combines music with symbols for bodily movements. Baroque, music-based and abstract dance notation uses different types of symbols for different types of bodily movements. Hence it might be suggested that the earliest language was multimodal. There is an ongoing speculative debate about the priority of vocalization or manual gesture in the evolution from pre-language to language. It is perhaps most likely that both were used at first, without the separation that we currently conceive for the two (Arbib 2012, McNeill 2012). And of course modern spoken language is accompanied by manual gestures, and some have argued that intonation represents a vocal gesture that is at least partly separate from language in the sense of the latter being a segmentally-based symbolic system of communication (Bolinger 1985).

Finally, semasiographic systems exhibit virtually no duality of patterning. The closest thing to duality of patterning is the use of determinatives or classifiers and phonological complements in logographic writing systems (e.g. Cooper 2004:89; Nissen et al. 1993:115). The absence of duality of patterning may be due to the restricted domains of expression of most semasiographic systems, even late in their evolution, in contrast to the general-purpose communicative function of modern human language (see §3). But it might also suggest that the earliest pre-language did not exhibit duality of patterning at first either.
One important difference between semasiographic systems and modern human languages in their spoken or signed medium is that language in use must be processed in real time, while semasiographic systems, being lasting, are not. Thus the selectional pressures for semasiographic systems and language might be somewhat different, and the evolutionary characteristics may also differ. However, the properties drawn out from comparison of the evolution of different semasiographic systems—gradualness, domain-specificity and context dependence, minimal expression, association of signifier and signified, function as coordination devices, exaptation of functions, multimodality and lack of duality of patterning—are independent of the medium of communication. In fact, some of these properties (in particular context-dependence, minimal expression, association of signifier and signified) are often treated as more typical of the real-time medium of spoken or signed language than of writing.

Before proceeding to speculation on the stages of the evolution from pre-language to language, it is worth taking a glance at the process of language acquisition. Although ontogeny doesn’t always recapitulate phylogeny, there are some parallels in early language acquisition with the observations made on the evolution of semasiographic systems that are worth noting in supporting speculations on language evolution.

Language acquisition is a gradual process. This has been repeatedly demonstrated by recent research on both phonological and grammatical acquisition (Vihman 1996; Lieven, Pine and Baldwin 1997; Tomasello 2003). The earliest use of language is also restricted in functional domain, to a restricted set of activities that the caregiver and child engage in. It is also minimal, initially consisting of single words that evoke the relevant interaction; productive grammatical inflections do not emerge until later. The signifier
and signified are directly associated at first—the first utterances are about immediate experience. Finally, the earliest communicative gestures are multimodal—both vocalizations and bodily gestures. Thus, language acquisition also shares many features with semasiographic systems.

Of course, there are some important differences between child language acquisition and the evolution of pre-language from language. The caregivers that children interact with are already users of modern human language. And children, at least after the “nine month revolution” (see §3), have most of the social cognitive capacity that underpins modern human language. But the parallels to the evolution of semasiographic systems observed here (admittedly, also created by humans with modern social cognitive abilities) lend support to the speculations offered in the next section.

5. Speculations on the evolution of language

The observations on semasiographic systems, and the parallels with child language acquisition, suggest that some functions are likely to be earlier in evolution, and less complex in an evolutionary sense, while other functions are likely to be later, and more complex in the same sense:

<table>
<thead>
<tr>
<th>Less “complex”</th>
<th>More “complex”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context dependent:</td>
<td>Context independent:</td>
</tr>
<tr>
<td>specialized function</td>
<td>general purpose</td>
</tr>
<tr>
<td>minimal coding</td>
<td>explicit coding</td>
</tr>
<tr>
<td>immediate experience</td>
<td>displaced experience</td>
</tr>
<tr>
<td>(associative)</td>
<td>(inferred)</td>
</tr>
<tr>
<td>extensional (indexical)</td>
<td>intensional (iconic/symbolic)</td>
</tr>
<tr>
<td>multimodality</td>
<td>duality of patterning</td>
</tr>
</tbody>
</table>
I have used the term “complex” in scare quotes because complexity here is based on evolutionary sequence. Also, it is not based purely on the structure of the signifier but on how the signifier expresses the concepts being communicated. Nevertheless, it seems that the phenomena in the table do reflect lesser or greater degrees of cognitive complexity. A context dependent system is going to be less complex because it involves a small number of more specific concepts, while a context-independent system will be more complex since it will involve abstracting categories across a wide range of concepts, and explicit coding will require more signal complexity (grammatical inflections, more complex constructions, etc.). Displaced experience is more complex to communicate than immediate experience since it requires conceptualization of an experience apart from the speech act situation. Intensional symbols are more complex than extensional ones in that they require the formation of a category. Duality of patterning requires two alternative parsings of the same signal structure (phonological and morphosyntactic), whereas even a multimodal signal does not operate on two levels at once; both modalities are being used to convey the intended concept.

These asymmetries in communicative means can be used to suggest possible stages in the evolution from pre-language to language, as already suggested in §4.5. But at least as important, the logical structure of the suite of cognitive abilities described in §2 also suggests the evolutionary priority of some social cognitive abilities over others.

Joint action, coordination and communication all presuppose the presence of common ground. Hence we will take the evolutionary emergence of common ground—the ability
to recognize that others have mental states which can be shared (Tomasello 2008)—as our starting point.

In §2, we saw that common ground comes in different types that have different bases. Communal common ground emerges from shared practice among humans in a social group. If so, then personal common ground, based on common perception and action, seems prior: it is based on direct interpersonal interaction. But the actional basis presupposes the communal common ground of shared communicative conventions, in particular a shared language, in the community. Hence the actional basis must follow, or be part of, the shared practice that underpins communal common ground. The perceptual basis for personal common ground can arise from communal common ground resulting from shared practice among members of the social group. That is, the way that we categorize elements of our perceptual experience may be influenced by shared practice. But there is a more fundamental basis for common ground, namely a human being’s individual interaction with the world, including one’s own self (body, bodily functions and needs, etc.). We can call this natural common ground. Natural common ground is made up of an individual’s interaction with the world and one’s body. Its basis is a human’s perceptual recognition of species identity with another human being: this maximally inclusive “community” allows “natural” knowledge to be shared knowledge, that is, common ground.

Hence, an ordering of types of common ground based on logical priority would be something like this:

• natural common ground: individual ways of interacting with the world and one’s body, shared via recognition of other humans as conspecifics;
• perceptual personal common ground, shared via joint attention to and joint salience of the interacting individuals’ environment and actions;
• communal common ground, shared via shared practices between the members of the community for various social and cultural purposes;
• actional common ground after the emergence of language and other human communicative conventions.

With respect to coordination devices, a clear logical priority exists, as described in §2. Convention presupposes precedent: a convention cannot emerge until a coordination device is repeated due to precedent and then becomes common ground in the community. And precedent presupposes joint salience: joint salience allows the possibility of first-time (nonrandom) success in coordination, which can then be followed as a precedent. Joint salience and precedent require both current and prior interactions between the individuals, i.e. personal common ground. Convention is shared in a community, that is, conventions are part of a community’s communal common ground. And language is a conventional system for communication, so also presupposes communal common ground.

The logical relationships between types of common ground and types of coordination devices are given in Figure 1 (the dotted line indicates the relationship between joint attention and joint salience, and the double role it plays in the achievement of joint actions):
These logical relationships suggest the types of coordination devices that one might expect with different types of human interactions with the world, one’s self, and with others, that suggest stages in the evolutionary emergence of coordination devices including, eventually, language.

An evolutionary account of the emergence of these social cognitive abilities requires some sort of selectional pressure for hominins to develop each successive stage of these abilities. The obvious place to look is in the evolution of joint actions. In §2, we presented an analysis of joint actions that requires coordination and common ground, but also a degree of helpfulness, that is, commitment to the other’s success in carrying out their part of the joint action (Bratman 1992:336-38; Tomasello 2008, §5.2.1). Tomasello and colleagues have more recently developed an evolutionary account of the evolution of human morality as a basis for the emergence of joint actions (Tomasello 2011; Tomasello
et al. 2012; Tomasello and Vaish 2013). We will use this account to motivate the emergence of common ground, coordination devices, and thence language.

Tomasello and colleagues argue for two stages in the evolution of human morality beyond great ape “morality” (their interdependence hypothesis). They argue that while great ape sociality involves some degree of sharing, reciprocity and revenge, they are not committed to helping the other in collaborative activities in the way that humans are (Tomasello and Vaish 2013: 236). The first stage in the evolution of human morality is brought on by a selectional pressure for genuinely collaborative action (rather than just reciprocal altruism; Tomasello et al. 2012:673), which they take to be an environmental need to engage in collaborative foraging in order to survive (ibid., 674). This stage, which they call second-personal morality (Tomasello and Vaish 2013:240), involves helpfulness and commitment—treating the other as an equal, not as a ‘social tool’ (ibid. 236)—but specifically with individuals known personally to the agent. Reputation, critical to avoiding free riding, is based on direct experience of social interactions. The effect is that collaborative joint actions emerge, but only at a small scale.

The next stage is what Tomasello and colleagues call group-mindedness (Tomasello et al. 2012:681) or norm-based morality (Tomasello and Vaish 2013:245). They argue that this stage emerged as the result of increasing size of groups, and pressure from competing groups also increasing in size (Tomasello et al. 2012:681). The outcome is the ability to engage in collaborative joint activity with individuals who are not directly known to the agent, and hence do not have a reputation acquired by direct experience of interaction, but instead can be trusted by virtue of group membership. At this stage, there come to exist social norms and institutions that enable large-group collaboration.
Second-personal morality is probably the selectional pressure leading to the emergence of perceptual personal common ground and shared precedent as a coordination device (namely, the second level from the top in the figure). Perceptual personal common ground comes from direct interactions between the individuals, and shared precedent also requires direct experience. On the other hand, norm-based or group morality is probably the selectional pressure leading to communal common ground and convention as a coordination device. Collaboration is based on group membership; the common ground and coordination devices also have to be established groupwide, not by direct experience of interpersonal interaction.

The figure implies that, at least logically, there may have been two other discrete stages in the emergence of coordination devices, including language. The last stage is clearly an elaboration of group-mindedness, where the cultural transmission of knowledge through language (and, eventually, writing) proceeds at an abstract level and a large scale in a large society. The first stage may be an initial stage of second-personal morality. It presupposes at least some level of collaboration and it seems likely that such collaboration would first occur only in small groups of individuals known to each other—perhaps a kin group. Hence it may have been kin recognition rather than conspecific recognition that initially allowed for the emergence of (natural) common ground.

Finally, we may speculatively map more concrete proposals about pre-language and language based on the evolution of semasiographic systems and language development in Table 1 onto the logical sequence of common ground and coordination devices in Figure 1.
The first stage involving joint action, as a number of others have proposed, is most likely to employ only indexicals. Indexicals rely solely on natural common ground and joint salience or attention as the coordination device. Indexicals involve association and hence are restricted to immediate experience; they are context-dependent and typically multimodal—even linguistic demonstratives in their indexical uses are multimodal, accompanied by gestures (Diessel 2006:469-71). Iconic gestures might also serve as coordination devices at this stage, since they also require only natural common ground, particularly when used for immediate experience; they are also likely to be multimodal. The sort of joint actions that are coordinated by indexicals, icons and natural common ground would have to be very simple, immediate actions.

However, it is likely that individuals would engage in joint actions of this sort only with close associates, perhaps only kin. Second-personal morality might represent a more extensive group which engages in collaborative joint actions: known individuals, but not necessarily kin. Precedent as a coordination device would allow the emergence of more stable signals. One type of signal might be proper names for individuals, since individual identity is important for second-personal morality (proper names are also among the first things to appear in the evolution of writing). These proper names would be used at least as vocatives (i.e. in immediate experience) or possibly in displaced contexts. Proper names might begin with a gesture iconic of an associated property of the individual but may evolve to arbitrary symbols. Naming would facilitate the performance of polyadic joint actions. As such, naming might lead to vocalization as a more prominent or even primary modality: the broadcast nature of vocalization compared to visually perceived bodily gestures may be favored for selection in a polyadic joint action situation (cf.
Tomasello 2008:231). Stable symbols within the group might also lead to the emergence of semantic content in symbols (in fact, iconic gestures may also be used with semantic content). The joint actions are everyday activities that involve a number and/or variety of objects as well as agents (so are more complex in that sense), requiring explicit categorizing and labeling.

The advent of larger groups gives rise to communal common ground and conventional signals in that community. It seems likely that the evolutionary success of such groups would require engaging in planned joint actions. These joint actions are complex activities requiring subplans that need to be made explicit (that is, they are not simply part of communal common ground and cannot be executed by everyone without explicit negotiation or instruction). The content for communication is not just immediate experience but also a restricted set of imagined and displaced experience, namely the intended or desired subplans that must occur in the future in order to achieve the total planned joint action, and the subplans that have just been executed. Successful coordination of planned joint actions of this sort would require combinations of signals (different combinations may involve gesture + gesture, gesture + vocalization, or vocalization + vocalization). Combinations are required to construct imagined experiences, namely the intended subplans. Since these experiences are not in the here and now, and may not have existed before, they need to be evoked by combining units that do denote objects and actions in the here and now, but in combinations representing a novel situation. At this stage, combinations are likely to be structured by order, not grammatical inflection; the sort of fine-grained conceptual detail that calls for
grammatical inflection (see Croft 2007a) is not needed for coordination of everyday joint actions, even these more complex actions.

The final stage is the emergence of sharing of knowledge. The joint actions are complex, but now include the joint action of sharing propositional knowledge, beliefs, desires, evaluations, etc. In other words, they are not coordinating some physical everyday activities but enlarging the stock of knowledge (including of course cultural knowledge and attitudes) in the community. The content to be communicated is now not just immediate experience and imagined experience geared to executing complex actions, but also recalled experience and other types of imagined experience not leading to immediate action. At this point, actional common ground is produced by the joint action. Of course, the coordination devices include all three of joint attention, precedent and convention. The signaling system includes the signals from the earlier stages and combinations of such signals. At this stage, the combinations of signals serves to communicate displaced experience and the full range of imagined experience. The generalization of the function of the signaling system to including sharing of knowledge of all kinds will lead to the emergence of grammatical inflection and derivation, and the sort of grammatical constructions familiar from modern human languages. This generalization of function will also lead to duality of patterning in the structure of the signal, though it is possible that duality of patterning emerged earlier (see §1 and Mufwene 2013). Finally, this generalization of function leads to the primacy of vocalization in the signaling system (among hearing persons). In other words, the final stage leads to modern human language, though I would guess that this stage also
undergoes gradual and incremental evolution, and we might only be able to speak of modern human language at the end of the fifth stage described here.

This account of stages of language evolution is very speculative, of course (perhaps modeling would give some clues). Nevertheless, by grounding the account of the evolution of language in its function to facilitate joint actions among individuals in a community, we can develop plausible scenarios for the stages it must have gone through on the way to modern human language.

References


Coupé, Christophe. 2012. Semiotic investigations into early forms of symbolism and language. *The evolution of language: proceedings of the 9th International Conference*


d’Errico, Francesco, Christopher Henshilwood and Peter Nilssen. 2001. An engraved bone fragment from c. 70,000-year-old Middle Stone Age levels and Blombos Cave, South Africa: implications for the origin of symbolism and language. *Antiquity* 75:309-18.


