Language and Cognition **6** (2014), 327–369. doi:10.1017/langcog.2014.9

A perceptually grounded model of the singular–plural distinction

HAYDEN WALLES

ANTHONY ROBINS

AND

ALISTAIR KNOTT

Department of Computer Science, University of Otago, New Zealand

(*Received* 15 *July* 2013 – *Accepted* 22 *November* 2013 –
First published online 13 May 2014)

ABSTRACT

Embodied theories of language posit that the human brain's adaptations for language exploit pre-existing perceptual and motor mechanisms for interacting with the world. In this paper we propose an embodied account of the linguistic distinction between singular and plural, encoded in the system of grammatical number in many of the world's languages. We introduce a neural network model of visual object classification and spatial attention, informed by a collection of findings in psychology and neuroscience. The classification component of the model computes the type associated with a visual stimulus without identifying the number of objects present. The distinction between singular and plural is made by a separate mechanism in the attentional system, which directs the classifier towards the local or global features of the stimulus. The classifier can directly deliver the semantics of uninflected concrete noun stems, while the attentional mechanism can directly deliver the semantics of singular and plural number features.

KEYWORDS: grammatical number, visual attention, object classification, global precedence.

1. A proposal about the perceptual origin of the linguistic singular–plural distinction

Language allows us to talk about objects both as individuals and as groups. In many languages, a distinction between objects and groups is encoded in syntax, in the grammatical distinction between SINGULAR and PLURAL (e.g., in English, the distinction between *dog* and *dogs*). In such languages,

it is not just possible to identify a referent as an individual or a group, but syntactically obligatory to do so.

Where does the deeply ingrained singular–plural distinction in language come from? Researchers have discovered several prelinguistic systems for representing number in the brain, but none of these have quite the right properties to deliver the distinction between singular and plural found in language. In this paper we will make a novel proposal, namely that the linguistic singular–plural distinction has its origin in the visual attention system: specifically, in the system that allocates attention selectively to the local or global form of a visual stimulus. Attention to local and global form was first explored in the classic experiments of Navon (1977), using stimuli of the kind shown in Figure 1. Observers can identify either the global or local form of such stimuli. The global form of a stimulus is, roughly speaking, its shape: in Figure 1, 'S'. The local form of a stimulus is the shape of the homogeneous elements from which it is composed: in Figure 1, 'H'. When classifying a stimulus, an observer must choose whether to identify its global or local form, because these forms can differ, as in Figure 1. The attentional mechanisms that implement this choice have been studied quite extensively. What is less commonly noted is that choosing to attend to the global or local form of a stimulus commits the observer to a decision about the *number* of objects of the identified form. When we classify the global form of a stimulus, we necessarily identify a single instance of this form: that is what 'global form' means. When we classify its local form, we necessarily identify more than one instance of this form. A stimulus only *has* a 'local form' if it is composed of a homogeneous group of smaller forms.

In this paper, we propose that the linguistic distinction between singular and plural may have its origins in the neural circuitry that allocates attention to the local or global features of a visual stimulus. We express the proposal within a computational model of visual attention and object classification, which can identify both objects and groups, and can recognize both the local and global form of Navon stimuli. We begin in Sections 2 and 3 by situating our proposal within a wider account of the representation of type and number information in language and visual perception. In Section 4 we review existing proposals about how linguistic type and number representations may be grounded in perceptual representations, and introduce the main new features of the model we propose: a novel model of object classification, and a novel model of attention to spatial scale. In Sections 5 and 6 we give an overview of the model; Section 7 describes its performance in two perceptual tasks. Sections 8 and 9 compare the model to existing models of visual classification and attention, and assess its potential as the basis for an account of linguistic type and NUMBER information.

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

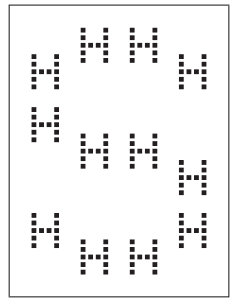

Fig. 1. A Navon stimulus, with global form 'S' and local form 'H'.

2. Type and number information in language

Human language conveys information about the type and number of objects quite separately. Information about object type is conveyed mainly through the system of noun stems. Through this system a speaker can directly identify a large open-class set of object categories, comprising classes of concrete objects (*dog*, *cat*, *xylophone*), abstract objects (*peace*, *idea*), and other types of semantic object denoting times, places, activities, and so on. However, noun stems do not systematically convey information about the number of objects whose type they denote. Number information is mostly provided by separate closed-class linguistic items, whose meanings must be combined with those of noun stems in various ways. In standard models of language semantics (see, e.g., Barwise & Cooper, 1981), a count noun stem contributes a *set* of objects (possibly empty) of a given type to the semantic representation of a sentence. We will call this set the 'noun set'. The cardinality of this set is determined by a combination of other elements in the sentence. It could be exactly one (*A dog barks*) or more than one (*Ten dogs bark*), or even zero (*No dogs bark*); the point is that elements other than the noun stem determine this cardinality.

There are two main systems in language for providing cardinality information to supplement the type information provided by noun stems. One is the system of quantifying determiners. In a standard model of semantics (see again Barwise and Cooper, 1981), these offer ways of precisely identifying the size of a set of individuals, either in absolute terms (*one*, *twenty-five*) or in relation to the size of some other set (*every*, *exactly half*). They also offer 'vague' ways of identifying group size, again either in absolute terms (*several*, *many*) or in relative terms (*most*).

A second linguistic number system is the grammatical system of NUMBER features. In most European languages, NUMBER features are expressed in bound morphemes or closed-class words, and signal the alternative values SINGULAR and PLURAL. (We will use small caps to denote syntactic feature

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

names and their values.) For instance, the English morpheme *-s* marks a noun as having the NUMBER value PLURAL; the Italian morphemes *-o* and *-a* mark a noun as SINGULAR. In many languages, the NUMBER feature has more than two possible values: for instance Polynesian and many Semitic languages include a DUAL class signalling 'exactly two', and in a few languages there is an additional TRIAL class signalling 'exactly three'. In these cases PLURAL means 'more than two' or 'more than three' respectively (Corbett, 2000).

NUMBER features have syntactic properties as well as semantic properties. In most European languages, for instance, within a noun phrase (NP) the NUMBER value of the head noun must agree with that of the determiner which introduces it, and the NUMBER of the subject NP must agree with a NUMBER feature on the main verb. In fact, the role of NUMBER features is often *primarily* syntactic; they do not always convey semantic information about cardinality at all (see, e.g., Hurford, 2003). For instance, *No dogs bark* has a syntactically PLURAL subject, but asserts that the set of dogs that barked is empty. Nonetheless, in NPs used referentially, to identify particular objects or groups, NUMBER features do frequently convey semantic information about number. Here are some examples:

(1) A dog walked in
(2) The dog barked
(3) John bought some cakes
(4) The cakes were tasty

The SINGULAR subject NP in example (1) introduces a single dog into the domain of discourse (Kamp & Reyle, 1993), and the SINGULAR subject NP in example (2) presupposes the existence of a single dog (van der Sandt, 1992). The PLURAL object NP in Example (3) introduces a group of cakes, and the PLURAL subject NP in Example (4) presupposes such a group. In these cases, the syntactic features SINGULAR and PLURAL directly deliver semantic information about the singularity or plurality of referents. In fact, the semantic notions of singularity and plurality do useful work even when they are not explicitly signalled. For instance, Chierchia (1998) argues that mass nouns (e.g., *wine*, *sand*), while syntactically SINGULAR, denote semantically plural referents: this explains why they share many syntactic properties with PLURAL count nouns (for instance, in English they can appear 'bare', without a determiner). So in language there are both syntactic and semantic distinctions between singularity and plurality.

In summary, linguistic information about the type and number of objects is conveyed by quite distinct components of syntax. Open-class noun stems

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

convey type but not number, while quantifying determiners and NUMBER features convey information about number but not type.

3. Type and number representations in the visual system

The primate visual system also computes information about object type and object number quite separately. In this section we review evidence for a distinction between type and number information in visual neural pathways, drawing on studies of both monkeys and humans.

3.1. REPRESENTATIONS OF OBJECT TYPE IN THE VISUAL SYSTEM

The neural assemblies that represent concrete object categories in humans are multimodal, and distributed over several areas of cortex, including superior and inferior frontal and parietal cortex (Just, Cherkassky, Aryal, & Mitchell, 2010; Sudre et al., 2012). However, the visual properties of objects which determine their category membership are primarily stored in the temporal and lateral occipital cortex (see Martin, 2007, for a review). For instance, Kreiman, Koch, and Fried (2000) found that the responses of single neurons in the medial temporal cortex are often specific to particular categories, such as faces, houses, and animals; in fMRI studies, Kriegeskorte et al. (2008) found distinctive patterns of activation in inferior temporal cortex for animate and inanimate objects, and within animate objects for faces and body parts; Shinkareva, Mason, Malave, Wang, Mitchell, and Just (2008) found patterns focused in ventral temporal cortex identifying the categories of tools and dwellings; Connolly et al. (2012) found activation patterns in temporal cortex which distinguished six animal species. Polyn, Natu, Cohen, and Norman (2005) found neural patterns associated with visually presented faces, locations, and objects, in which fusiform and parahippocampal cortex played an important role. Significantly, when subjects retrieved these object stimuli in a free recall task, the associated patterns in temporal cortex became active in advance of recall, providing evidence that they are functionally involved in representing categories, rather than epiphenomenal. Xue, Dong, Chen, Lu, Mumford, and Poldrack (2010) extended this finding in an experiment where subjects memorized pictures of faces during a study period, and later had to distinguish these faces from unseen faces in a recognition task. They found the similarity between the neural patterns encoding individual presentations of a given face during the study period predicted agents' later success in recognizing that face. This was confirmed separately for lateral occipital, ventral temporal, and fusiform gyri, again suggesting that patterns in these areas actively participate in

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

object representations. In macaque monkeys there is good evidence that cells in the anterior inferior temporal cortex (aITC) are sensitive to the visual properties of objects such as shape and colour, which play an important role in determining their type (see, e.g., Tanaka, 1996; Conway, Moeller, & Tsao, 2007; Zhang, Meyers, Bichot, Serre, Poggio, & Desimone, 2011).

3.2. REPRESENTATIONS OF NUMBER IN THE VISUAL SYSTEM

Three separate neural systems for representing number have been identified (see, e.g., Feigenson, Dehaene, & Spelke, 2004). One is a system for gauging the approximate number of items in a perceived group – a measure called 'numerosity'. The numerosity system operates on a logarithmic scale: it allows two groups to be distinguished if the ratio between their cardinalities is sufficiently large (around 7:8 in adults; Barth, Kanwisher, & Spelke, 2003). Numerosity is computed in the intraparietal cortex, both in monkeys (Nieder & Miller, 2004) and humans (Cantlon, Brannon, Carter, & Pelphrey, 2006; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Izard, Dehaene-Lambertz, & Dehaene, 2008). This is an area which is also heavily involved in computing representations of spatial location supporting directions of focal attention and visual search. A second number system is specialized for identifying the precise cardinality of small groups of up to four or five. This is called the 'parallel individuation' or 'subitizing' system. Its existence is motivated by two findings. First, humans can count a group of one to four objects in roughly constant time, suggesting that groups of this size can be counted in parallel (Trick & Pylyshyn, 1994). Second, there is evidence that both human infants and monkeys can successfully discriminate between pairs of numbers when they are both small (in the range one to three) but not when one is small and one is outside this range (Feigenson, Carey, & Hauser, 2002). This indicates that a system based on something other than ratios is being used to represent small numbers. The proposal is that small numbers in the range one to three are represented individually, in a scheme that does not encode the position of these numbers on a continuous scale. There is some evidence that subitizing has its own neural substrate: in humans, Ansari, Lyons, van Eimeren, and Xu (2007) found that counting subitizable groups preferentially activated the right temporoparietal junction (TPJ) compared to counting larger groups. A final number system, specific to humans, is the serial counting system. Counting involves systematically attending to each item in a group while engaging in a verbal counting routine; this allows identification of the exact cardinality of groups of arbitrary size. The counting system also recruits the attentional circuitry in intraparietal cortex, but relies in addition on superior lateral and medial precentral areas which are probably engaged in subvocal articulation and perhaps covert pointing gestures (Piazza, Mechelli,

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

Price, & Butterworth, 2006). There may also be learned associations between numbers in the counting system and points on a spatially arranged number line (see, e.g., Fias & Fischer, 2005).

3.3. DISSOCIATIONS BETWEEN NEURAL REPRESENTATIONS OF TYPE AND NUMBER

While inferior temporal cortex is sensitive to the type of visually presented objects, it appears quite insensitive to the number of objects of a given type which are presented, both in monkeys and humans. Nieder and Miller (2004) found that macaque aITC neurons are relatively insensitive to the cardinality of groups. They presented monkeys with objects in groups ranging from one to five in size. Animals were shown a sequence of two groups, and had to judge whether the second group contained the same or a different number of items as the first. Stimuli varied in number, but also in the type of their constituent items (which could be circles or a mixture of different shapes). Relatively few aITC neurons (8%) were sensitive to the number of objects in a stimulus display, and of these, only half (4%) were sensitive to number but not type. In another area, the intraparietal sulcus, 20% of neurons were sensitive to the number of objects in a display without regard for their type. (As discussed in Section 3, this area is involved in encoding the numerosity of objects in a group.) Clearly, information about the number of objects in a group stimulus is represented somewhat separately from information about the type of these objects.

A similar dissociation has been found in imaging studies in humans (Cantlon et al., 2006; Piazza et al., 2004; Izard et al., 2008). These studies use a habituation paradigm, identifying neural areas sensitive to selective changes in the number of objects in a display or in their type. Subjects were presented with a series of homogeneous group stimuli. Successive groups had different global configurations, but each stimulus contained the same number of items and the homogeneous groups all featured the same type of item. At a certain point, a deviant stimulus was presented, which had either a different number of items (but of the same type as previous stimuli) or items of a different type (but with the same cardinality as previous stimuli). This paradigm has been used with fMRI imaging, in both adults and four-year-old children (Cantlon et al., 2006; see also Piazza et al., 2004, for additional adult data), and with ERP recordings, in infants aged 0;3 (Izard et al., 2008). Cantlon et al. found that the intraparietal sulcus (IPS) responded selectively to deviations in number, while areas in the ventral and/or occipital temporal cortex responded selectively to deviations in shape, in both adults and four-year-olds. Izard et al. found a similar result with infants aged 0;3: deviations in number selectively activated intraparietal areas, while deviations in shape selectively activated

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

ventral temporal areas. These experiments provide support for the idea that number and type are computed in different processing streams in the human visual system.

There is also evidence from behavioural experiments that the human object classifier is relatively insensitive to the cardinality of homogeneous groups. One piece of evidence comes from an effect called 'redundancy gain': under some circumstances observers can be faster at classifying objects in a homogeneous group than individually (Theeuwes, 1994). This suggests minimally that objects in homogeneous groups can be classified in parallel. But there are also conditions under which observers can identify the type of visually presented stimuli, but fail to distinguish between single and multiple instances of this type. In a phenomenon called 'spatial repetition blindness', observers fail to distinguish between a single object and two objects of a single given type when these are presented for a brief period (Kanwisher, 1991). In patients with unilateral damage to parietal cortex, a related failure to compute cardinality is often manifested more persistently, in a variety of spatial neglect called 'extinction'. Extinction occurs when a subject fails to notice a stimulus in the contralesional visual field in the presence of a stimulus in the ipsilesional field. It is typically more severe if the two objects are similar (Baylis, Driver, & Rafal, 1993): paradigmatic cases of extinction involve an inability to distinguish between one and two instances of a given object type. Again this finding suggests that type and number are computed separately from visual inputs, and implicates parietal cortex in the processing required to individuate tokens of a given type.

4. A perceptually grounded account of the semantics of noun stems and NUMBER features

As just reviewed, the distinctions between type and number information in the brain's perceptual pathways bear some resemblances to those encoded in language. In this section we discuss the possibility that the linguistic distinctions *originate in* the structure of the perceptual system. The idea that structures in language supervene on structures in the perceptual and/or motor system has been articulated in several models, termed 'embodied' models of language (see, e.g., Barsalou, 2008; Feldman & Narayanan, 2004; Lakoff, 1987). These models posit that the human brain's adaptations for language recruit pre-existing non-linguistic mechanisms for interacting with the world, so that processing sentences directly engages these general mechanisms in some way. Embodied accounts of language pay particular attention to sentences that directly report the experience of concrete events in the world, such as examples (1–4). A common suggestion is that processing a sentence describing a concrete event involves evoking the same sensory or

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

motor representations as are generated when the described event is actually experienced (see, e.g., Bergen & Chang, 2005; Zwaan & Taylor, 2006). In this section we will review existing suggestions about how linguistic representations of type and number may have their origin in perceptual mechanisms, and introduce the novel features of our own proposal.

4.1. A PERCEPTUALLY GROUNDED ACCOUNT OF CONCRETE NOUN STEM SEMANTICS

Nouns denoting concrete objects evoke patterns of activity distributed over several cortical areas (see again Just et al., 2010; Sudre et al., 2012), but most theorists agree that the left temporal cortex plays a particularly important role in these patterns. In classic experiments, Damasio, Grabowski, Tranel, Hichwa, and Damasio (1996) showed that damage to the left temporal cortex selectively impairs the production of concrete nouns describing pictures of objects, and damage to different subregions of this area selectively impairs production of nouns describing objects of different types; Tranel, Adolphs, Damasio, and Damasio (2001) found this disruption even when subjects could identify the visually presented objects. Whether the temporal cortex is involved in representing the grammatical category of nouns *in general* is still controversial (see Vigliocco, Vinson, Druks, Barber, & Cappa, 2011). For the moment, we will focus on representations of concrete noun stems, in which temporal cortex uncontroversially plays an important role.

The observation we focus on in our model of concrete noun stems is the finding that neural representations of object type abstract away from number information, as just reviewed in Section 3. This fact is often overlooked in embodied accounts of noun semantics, but it strongly supports models in which the semantics of concrete noun stem semantics are read directly from object type representations. Like neural type representations, noun stems do not encode number information: for instance the noun stem *dog* is the same in the SINGULAR noun *dog* and the PLURAL noun *dogs*. This makes neural representations of object type ideally placed to deliver the semantics of concrete noun stems.

However, in a model of the perceptual processes which deliver the semantics of noun stems, it is important to provide an account of *how* the visual object classification system can abstract away from number information. If the human object classification system is able to directly deliver the denotations of noun stems, it must be able to classify homogeneous groups of objects as well as single objects – for instance, when presented with a group of dogs, it should be able to deliver the object type 'dog', just as it does when presented with a single dog. In fact, it must be relatively insensitive to cardinality, responding in essentially the same way to a group of dogs as to a single dog,

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

since a noun stem reports type but not number. Most computational models of human object classification are designed to operate on single objects rather than groups. A novel aspect of our computational model of object classification is that it can classify homogeneous groups as well as single objects, and is cardinality blind (see Section 6.1). We will argue that this classifier has the right properties to directly deliver noun stem denotations.

4.2. A PERCEPTUALLY GROUNDED ACCOUNT OF THE NUMBER FEATURE

Where in the human cognitive system does the linguistic NUMBER feature come from? While its role is no longer purely semantic, it may be that its role in language was originally largely semantic. There are good indications that monkeys can also distinguish perceptually between single objects and groups (Barner, Wood, Hauser, & Carey, 2008); human infants are also able to do this, even before they acquire the linguistic distinction between SINGULAR and PLURAL (Li et al., 2009). One possibility is that the NUMBER feature has its origin in a non-linguistic perceptual or cognitive system which distinguishes between single objects and groups, or between single objects, pairs, triples, and larger groups. The question is then which non-linguistic system can contribute this information.

4.2.1. *Subitization as a possible substrate for* NUMBER

The main proposal in the literature is that values of the NUMBER feature can be read from the subitization system, i.e., the system which represents the size of small groups in the range one to four. Hurford (2001) notes that language treats these numbers specially in several respects: for instance in many languages, the numerals one to four have idiosyncratic gender and case marking when used attributively to express a cardinality (as in *one dog*, *two cats*), and in many languages the ordinals for one to four have idiosyncratic forms (e.g., the English *first* and *second*). Critically, he notes that grammatical NUMBER systems allow alternative values between one and 'around three' in the languages he surveys. While Hurford does not want to propose simple correspondences between neural systems and number systems in language, he does suggest an association between linguistic NUMBER features and the subitization system.

Sarnecka, Kamenskaya, Yamana, Ogura, and Yudovina (2007) also argue for a connection between NUMBER and the subitization system. They find that infants whose native language contains an explicit NUMBER feature (English and Russian) are better at performing tasks requiring production or understanding of the number words one, two, and tree than infants whose

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

language has no explicitly marked NUMBER (Japanese). Their explanation is that the meanings of these number words are initially represented as values of the grammatical NUMBER feature, again within the subitization system.

As Sarnecka et al. (2007) acknowledge, an obstacle for a subitization-based account of NUMBER features is the feature value PLURAL. A simple suggestion that PLURAL signals a non-subitizable group is clearly wrong: for instance in English, sets of size two and three are PLURAL, but still subitizable. In fact the concept of PLURAL always encompasses a mixture of subitizable and non-subitizable numbers. This is even true in languages where PLURAL means 'more than three', since the subitization limit is four on the most conservative estimates (see, e.g., Trick & Pylyshyn, 1994). If the concept of PLURAL has its origin in non-linguistic semantic number representations, it is not yet clear what these representations are. Until this issue is resolved, it is also unclear whether the other possible values of the NUMBER feature (SINGULAR, DUAL, and TRIAL) denote representations within the subitization system: some connection would need to be found between the subitization system and the system representing PLURAL, to explain why SINGULAR, DUAL, TRIAL, and PLURAL are all possible values of the NUMBER feature.

4.2.2. *Local/global attention as a possible substrate for the distinction between* SINGULAR *and* PLURAL

As discussed in Section 1, our novel proposal is that the grammatical distinction between SINGULAR and PLURAL derives from the mechanism that allocates attention to the local or global form of a visual stimulus. As already noted, classifying the global form of a stimulus involves treating it as a single token, by definition. Classifying its local form requires treating it as a group, again by definition: a stimulus only *has* a 'local form' if it is composed of a homogeneous group of smaller forms.

There is some interesting recent evidence that the neural processing of SINGULAR and PLURAL features in language may coopt the neural system that allocates visual attention to local or global form. This comes in a recent study of the neural processing of NUMBER information by Domahs, Nagels, Domahs, Whitney, Wiese, and Kircher (2012). Their study directly compared the brain activity elicited by SINGULAR and PLURAL nouns. Subjects were presented with three types of noun: SINGULAR and PLURAL count nouns (e.g., *dog* and *dogs*) and mass nouns (e.g., *water*). Their fMRI responses to these nouns were analyzed. Domahs et al. found that PLURAL count nouns elicited more activity in the left temporoparietal junction (specifically the left angular gyrus) than SINGULAR count nouns. (This was the only area where differences were found between SINGULAR and PLURAL count nouns.) Interestingly, mass nouns also elicited more activity in the left TPJ than

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

SINGULAR count nouns; in fact the left TPJ response to mass nouns was indistinguishable from that of SINGULAR count nouns. Recall from Section 2 that mass nouns behave syntactically much like PLURAL count nouns, and many syntacticians regard them as semantically PLURAL, even if they do not carry PLURAL inflections (Chierchia, 1998). Given that the left TPJ selectively responds to PLURAL count nouns and to mass nouns, Domahs et al. postulate that this area has a general role in representing semantically PLURAL stimuli.

Domahs et al. (2012) suggest that the left TPJ's involvement in representing plurality derives from its involvement in spatial processing, but they make no specific proposals about how this process may occur. However, our suggestion that a visual stimulus is represented as PLURAL when an observer attends to its local form fits very well with Domahs et al.'s findings. There is a large body of evidence suggesting that the left TPJ is involved in allocating attention to local features of Navon stimuli. This has been shown in analyses of brain dysfunction (e.g., Robertson, Lamb, & Knight, 1988) as well as in PET and ERP studies (e.g., Fink, Halligan, Marshall, Frith, Frackowiak, & Dolan, 1996; Yamaguchi, Yamagata, & Kobayashi, 2000). Domahs et al.'s finding that the left TPJ is preferentially activated by PLURAL nouns is exactly what our proposal predicts. If subjects rehearse the attentional operations necessary to categorize the local form of a visual stimulus when they hear a PLURAL word, we expect to see activity in the left TPJ. So Domahs et al.'s findings are consistent with the model we propose.

It is also interesting to note that the neural region associated with attention to the *global* form of a stimulus is the *right* TPJ (Robertson et al., 1988; Fink et al., 1996; Yamaguchi et al., 2000). As discussed in Section 3, the right TPJ is the region activated by counting numbers in the subitization range (Ansari et al., 2007). Several theorists have suggested that subitization works by recognizing characteristic shapes associated with small groups of particular cardinalities; for instance, three objects characteristically form a triangle (see Palomares & Egeth, 2010). Identifying these shapes would clearly require attention to the global form of the group, so the involvement of the right TPJ in attention to global form certainly supports shape-based accounts of subitization. In summary, the system which allocates attention to the local or global form of a visual stimulus may provide the basis for an account not only of the distinction between SINGULAR and PLURAL, but also for an account of the other possible values of the NUMBER feature, DUAL and TRIAL. In this paper, we will restrict our attention to the SINGULAR/PLURAL distinction.

4.3. SUMMARY

In this section we introduced two new ideas about the perceptual mechanisms that underlie linguistic representations of type and number. We proposed

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

that the denotations of noun stems are delivered by an object classifier that is insensitive to cardinality information. And we propose that the semantic singular/plural distinction encoded by the NUMBER feature is read from an attentional system allocating attention to the local or global features of a visual stimulus. Note that these two proposals are linked. Classifying the local form of a stimulus always involves classifying a homogeneous *group* of shapes. So an account of the mechanism that selectively attends to the local and global form of a stimulus must include an account of a classifier capable of classifying both individuals and homogeneous groups.

In the remainder of this paper, we express these proposals in more detail in a computational model of how visual attention interacts with a cardinality blind object classifier, to allow the perception of both single objects and plural groups, and to recognize the difference between singular and plural stimuli. Our aim is to show that this model makes sense as a model of visual perception in its own right, as well as an account of the perceptual origin of linguistic noun stem and NUMBER representations.

5. Model overview and motivation

The basic structure of our model follows the well-established proposal by Ungerleider and Mishkin (1982) distinguishing two streams of visual processing: a 'what' stream in ventral cortex, subserving object classification, and a 'where' stream in dorsal cortex, subserving (among other things) spatial attention. In our model, the **attentional subsystem** (modelling aspects of the dorsal pathway) determines the salient regions on the retina, and activates these regions one at a time. The **classification subsystem** (modelling aspects of the ventral pathway) categorizes the retinal stimulus in the currently activated region; its output changes as different regions are selected. This basic architecture is similar to many others; see, e.g., Wolfe (1994, 2007), Rolls and Deco (2006), Mozer and Sitton (1998), Mozer and Baldwin (2008), Itti and Koch (2000), Walther and Koch (2006), and Navalpakkam and Itti (2005).

There are several lines of evidence which support the idea that the dorsal visual processing stream computes a map of salient locations; see, e.g., Gottlieb, Kusunoki, and Goldberg (1998) for evidence that intraparietal cortex computes salience, and Thompson and Bichot (2005) for evidence that the frontal eye fields do so. There is also good evidence that activity in this pathway modulates inputs to the object classification pathway. For instance, Moore and Armstrong (2003) found that micro-stimulation of a site in the frontal eye fields representing a particular retinal region selectively enhances visual responses in the corresponding region of extrastriate region V4, a key way station for visual information entering the object classification pathway in

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

inferotemporal cortex (IT). To confirm that attention affects representations in IT, Zhang et al. (2011) found that when monkeys covertly attend to an object in the presence of distractor objects, the pattern of neural activation in IT is shifted towards the pattern evoked when this object is presented in isolation.

In the attentional subsystem in our model, the saliency of a region is determined by two factors: one is local contrast (how different it is from the surrounding region), the other is homogeneity (how similar its texture elements are). Salient regions can contain isolated visual features that contrast with their surroundings, but also regions containing repeated visual features. Computations of saliency are performed at multiple scales, so salient regions containing isolated visual features can be of different sizes. Salient regions containing repeated visual features (i.e., homogeneous textures) can also be of different sizes.

There are several existing computational models of saliency that detect salient regions of different sizes (see in particular Kadir & Brady, 2001), and numerous models of texture identification which detect regions containing repeated visual features (particularly relevant is Kadir, Hobson, & Brady, 2005). There are also many existing computational models of classification that allow objects of different sizes to be classified, by taking as input primitive visual features at a range of different scales (see, e.g., Riesenhuber & Poggio, 1999). The main innovations in our model are in how the saliency mechanism interacts with the classifier. There are two innovations, which we will discuss in turn.

5.1. SELECTION OF A CLASSIFICATION SCALE

One novel feature of our system is that classification is influenced not only by the location of the currently selected salient region, but also by its size. Our classifier can work with primitive features of several different scales as input, but at any given point the scale it uses, called the **classification scale**, is selected by the attentional system. By default, the classification scale is a function of the size of the currently selected salient region, so that large regions are classified using correspondingly large features, and small regions with correspondingly small ones. Our model is novel in proposing that the scale of the salient region selected by the attentional system determines a default scale for the classifier to use.

The idea of establishing a default classification scale for an attended region based on its size is illustrated in Figure 2. Two salient regions are shown in the figure: a large one and a small one. In order to recognize a figure within each region, the primitive visual features which the classifier must use must be of an appropriate spatial scale – not too large and not too small (see Sowden & Schyns, 2006). If they are too large, they cannot be combined

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

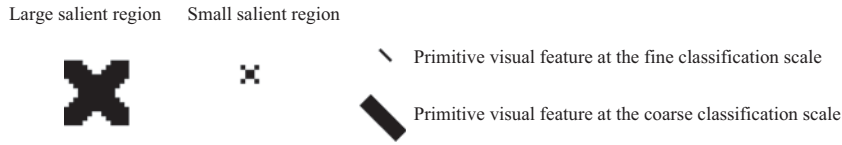

Fig. 2. Salient regions of different sizes, and associated classification scales.

to represent a complex shape within the region. (For instance, primitive features at the coarse-grained scale are of no use in representing a shape in the small salient region.) And if they are too small, then their combinations are not guaranteed to represent the global form of the figure occupying the region. (For instance, primitive features at the fine-grained scale are of no use in representing the 'global' shape of a figure in the large salient region.) In our model we implement a very simple treatment of the default classification scale: there are only two classification scales, which are appropriate for classifying 'large' and 'small' objects respectively. In reality, object classification involves a range of spatial scales rather than a single scale (Schyns & Oliva, 1999), so it would be more realistic to talk about a default *range* of classification scales. The important thing is that the relevant scales are specified in relation to the size of the object being classified (see Sowden & Schyns, 2006).

5.2. ALTERING THE CLASSIFICATION SCALE

A second novel feature in our system is that the selected classification scale can be changed *without changing the spatial region to be classified*, so that the classifier can reanalyze the currently selected region using finer-grained visual features. In our model, this attentional operation is crucial for the classification of homogeneous groups, and for an account of the difference between single objects and plural groups. The key idea is that in order to classify a group of objects occupying a given salient region, the observer must attentionally select a classification scale which is smaller than the scale established by default.

We suggest that the attentional mechanism which allows a choice between alternative classification scales for a given salient region is what underlies the visual ability to distinguish between singular and plural groups. If the classifier is analyzing a salient region at the default classification scale, any object type *T* it identifies will indicate the presence of a *single* object of this type in the region. If it is analyzing the region at a higher-than-default classification scale, any type *T* it returns will indicate the presence of *multiple* objects of this type in the region. The key idea is that the distinction between singular and plural is read from the current classification scale measured *in relation to* the default classification scale for the currently attended region.

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

5.3. SUPPORT FOR THE MODEL OF CLASSIFICATION SCALES

The model of classification scales just outlined is supported by several lines of evidence. First, it is well known that observers can selectively attend to the global or local features of visual stimuli (see, e.g., Fink et al., 1996). There is good evidence that this attention involves selective activation of particular 'spatial frequency channels' (see, e.g., Robertson, 1996; Flevaris, Bentin, & Robertson, 2010), which are analogous to spatial scales in our model. These findings motivate the mechanism in our model which selects a particular classification scale. But in addition, it has recently been found that the spatial frequency channels associated with local and global features of an object are defined in relative not absolute terms. Flevaris, Bentin, and Robertson (2011) conducted a priming study, where the primes were Navon stimuli, and the probes were stimuli containing both higher- and lower-frequency patterns. They found that attention to the local form of a Navon stimulus primed perception of the higher-frequency pattern, while attention to the prime's global form primed perception of the lower-frequency pattern, regardless of the absolute sizes (and retinal locations) of the patterns. Our model makes use of this notion of relative classification scale to support an account of group classification and of the distinction between singular and plural in the visual system.

6. Components of the model

The architecture of our model of visual attention and classification is shown in Figure 3. The classification subsystem is on the right; the attentional subsystem is on the left. In this section we describe these two subsytems in more detail. Technical details of both subsystems are given in the 'Appendix'.

6.1. THE CLASSIFICATION SUBSYSTEM

The visual classification subsystem is modelled by a convolutional neural network (CNN) based on that described in Walles, Knott, and Robins (2008) with a few refinements; see 'Appendix A.1' for details of the classifier's architecture and training scheme. It takes as input retinotopic maps of simple oriented visual features at one of two possible classification scales; the attentional system selects either large-scale or small-scale visual features. Activity in these input layers is propagated through subsequent layers which alternately combine features together in local regions of the input image and abstract over local regions of space. Responses of units in these layers model those of neurons in the ventral object processing pathway, which respond to progressively more complex shapes, over progressively wider areas of the visual field (see, e.g., Riesenhuber and Poggio, 1999).

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

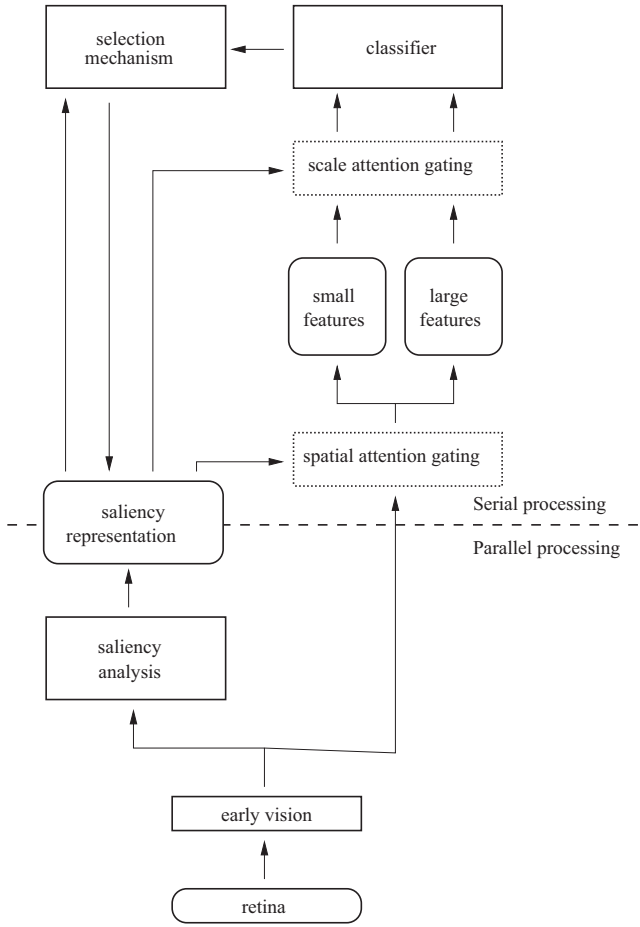

Fig. 3. Structure of the computational model. Boxes with rounded corners indicate representations and those with square corners indicate processes. Dashed boxes indicate gating to restrict the output to a subset of the input. Arrows indicate flow of information.

The classifier was trained with six shapes, each of which was presented at two different sizes, and at a range of locations (see Figure 4). The primitive visual elements of the small shapes are expressed at the fine-grained classification scale, and those of the large shapes are expressed at the coarse-grained classification scale. The large shapes were presented at varying pixel densities during training.

The classifier has seven output units: six of these provide localist encodings of the six shape categories and the seventh encodes the verdict 'unknown category'. The units have activations ranging from zero to one. We define the

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

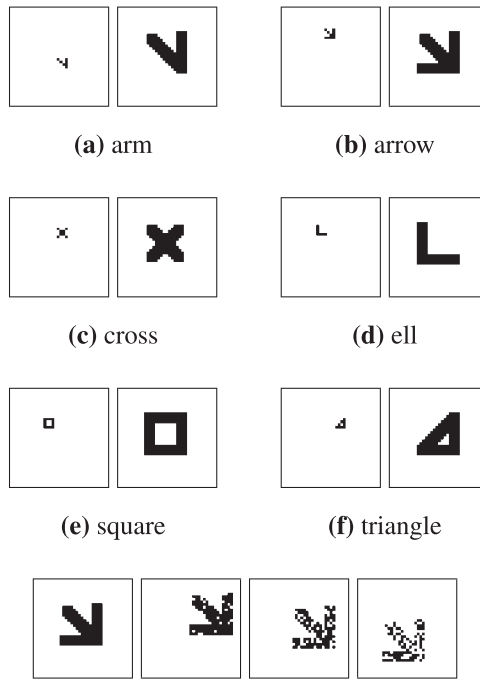

Fig. 4. The shapes used in the experiments. (a)–(f) show the shapes at the two sizes used. (g) shows how the large shapes were presented at four different densities.

classifier's decision to be the strongest output over 0.5. If no unit's activation exceeds 0.5, the classifier's decision is assumed to be 'unknown category'. In summary, the classifier provides two pieces of information: first, whether classification is possible, and, if so, what that classification is.

After training, the classifier exhibits two types of invariance which have been observed in the ventral visual system, in inferotemporal (IT) cortex (Logothetis & Sheinberg, 1996), and which are generally acknowledged to be crucial for a model of vision (Riesenhuber & Poggio, 2002; Ullman, 1996) – namely location (or translation) invariance and scale invariance. Location invariance is a result of the architecture of the CNN, which intersperses feature combination layers with layers that abstract over space (see Walles et al., 2008). Scale invariance depends on the input having been prefiltered for the desired scale. To classify the small shapes the fine-grained visual features must be used, and to classify the large shapes the coarse-grained features must be used. The attention system can select either the fine-grained or coarse-grained visual features. This allows the attention system to present an

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

input and ask the classifier if it can see anything familiar at a particular scale without interference from the other scales.

Most relevantly for the current paper, the classifier is also invariant to the cardinality of small shapes. (Cardinality invariance cannot be tested with large shapes because they are too big to present in groups.) Table 1 provides evidence of cardinality invariance for groups containing up to five items. For groups of two or more each input was formed by placing several small shapes on the retina with the constraint that no two shapes were less than two pixels apart. For each test example the units that passed the 0.5 threshold were tallied. For homogeneous groups of input type X, the number of times output X was over the threshold (correct group classifications) was recorded and is shown in the top part of Table 1. The number of times any output other than X was over the threshold (false positives) was also recorded; this is shown in the bottom part of Table 1. Clearly, the classifier is capable of identifying the type of items in homogeneous groups as well as the type of single items. In fact it is better at classifying groups than single items – a redundancy gain effect similar to that found in humans (see Section 3).

6.2. THE ATTENTIONAL SUBSYSTEM

6.2.1. *Overview*

As shown in Figure 3, the attentional subsystem can be divided into two interacting stages: a preattentive, or parallel, stage and an attentive, or serial, stage.

The preattentive stage includes an operation called 'saliency analysis'. The job of saliency analysis is to identify regions in the visual field that warrant individual processing: these regions are stored in a **saliency map**, which is computed in parallel across the visual field (Koch & Ullman, 1985; Itti & Koch, 2000; Walther & Koch, 2006). This map provides input to a serial selection mechanism which picks the most salient region in the map and allows the classifier to process information from this region. Once the classifier has processed the selected region, the saliency map is updated and a new salient region is selected.

In our model, the saliency map is called the **saliency representation**. Salient regions are identified as regions with high local contrast and high textural homogeneity. The former criterion is standard in saliency maps; the latter is novel in our system. The most salient location filters or 'gates' retinal input into the classifier, so that only visual features from this location are processed by the classifier.

The idea of restricting the classifier's inputs to a selected salient location is a well-known idea (see, e.g., Walther & Koch, 2006). In our model, the classifier's inputs are also independently gated by spatial scale. Given the

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

TABLE 1. *The classifier's performance on homogeneous groups of different sizes. The first five rows describe the response of the output unit X when presented with a homogeneous groups of one to five Xs. The second five rows describe the responses of the unit X to homogeneous groups of non-Xs.*

Pattern	Responses of unit X	Number of examples presented	Percentage
X	4493	4704	95.5%
XX	2958	3000	98.6%
XXX	2964	3000	98.8%
XXXX	2942	3000	98.1%
XXXXX	2912	3000	97.1%
Y	188	4704	4.0%
YY	79	3000	2.6%
YYY	105	3000	3.5%
YYYY	72	3000	2.4%
YYYYY	90	3000	3.0%

region and scale corresponding to the currently attended salient region, only primitive features corresponding to the selected region *and the selected scale* will reach the classifier from early vision. The novel aspect of the attention system is that a single selected spatial location can be classified first at a coarse-grained spatial scale and then later at a fine-grained one.

6.2.2. *Parallel attention component: saliency analysis*

The saliency analysis module parses the visual field and produces a saliency representation – basically a saliency map, which identifies salient regions at one of two spatial scales. Each region contains either a single shape, or a set of shapes which are treated as a single item (i.e., grouped) due to their proximity and/or similarity.

We determine saliency through a mixture of local contrast and texture homogeneity information. Two **local contrast maps** are computed, using Laplacian of Gaussian (LoG) filters tuned to two different spatial scales. A single **texture homogeneity map** at the fine-grained spatial scale is computed using the statistical histogram-based system of Liu and Wang (2000). These maps are combined to produce the saliency representation.

Prima facie, there may seem to be a conflict between local contrast and homogeneity as indicators of saliency. Saliency computed from contrast, and saliency computed from homogeneity (which implies *no* contrast) seem to push in opposite directions. However, while the principles may be in conflict at a single spatial scale, they are complementary at different spatial scales. Salient stimuli are those which contrast from their background *as wholes*, but whose *parts* show textural uniformity. Thus uniformity is required at a finer-grained spatial scale than contrast. Details of how the local contrast and

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

homogeneity maps are produced and combined into fine-grained and coarse-grained saliency maps are given in 'Appendix A.2', as are details of how the two saliency maps are combined into the master saliency representation.

Some illustrations of the regions found by the saliency analysis module are given in Figure 5. These demonstrate the module's ability to identify salient regions of different sizes: the input stimuli in Figure 5a are grouped into a single large region, those in Figure 5b are grouped into two medium-sized regions, and those in Figures 5c and 5d are identified as four small regions. They also show how the module reconciles conflicting local contrast and homogeneity cues to salience. If items are close enough (Figure 5a) then grouping can occur even among heterogeneous items. At an intermediate separation, grouping is determined by homogeneity: homogeneous stimuli are grouped (Figure 5b) and heterogeneous stimuli are not (Figure 5c). Finally, if items are separated widely enough, they are not grouped even if they are homogeneous (Figure 5d).

The contributions of the local contrast and homogeneity maps to overall salience are determined by the weights of two parameters, whose relative value determines the separation at which homogeneous stimuli are grouped. These parameter settings can be related to individual variations in grouping behaviour found in experiments on human subjects. Quinlan and Wilton (1998) explored the interaction of the Gestalt properties of similarity and proximity in humans. They found that proximity always dominates similarity if stimuli are sufficiently close, but that the distance at which this happens varies from subject to subject. The ratio between contrast and homogeneity weights in the computation of salience directly models this parameter of variation between subjects.

6.2.3. *Serial attention component: the selection mechanism*

The master saliency representation just described provides input to the serial attention component of our model, whose role is to selectively deliver bottom-up information from the retina to the classification subsystem (see the right of Figure 3). Selection occurs in two different attentional media. One is spatial location: the classifier can be restricted to receive input only from a particular region of the visual field. The other is classification scale: the classifier can be restricted to receive input from visual features at a particular spatial scale (fine-grained or coarse-grained). The way selection is implemented in each medium is described in 'Appendix A.3'.

Processing in the serial attention component takes the form of a sequence of **attentional operations**. There are two types of operation. One is the selection of a new salient region to attend to. When this happens, an appropriate classification scale is also automatically selected, namely the default classification

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

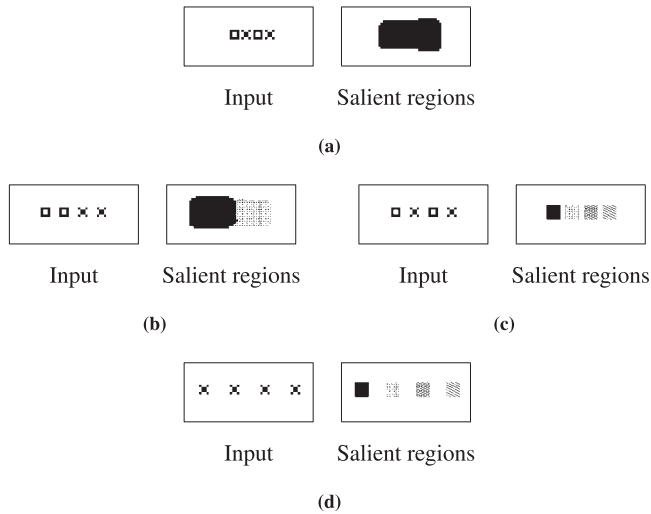

Fig. 5. Results of similarity–proximity conflict tests.

scale for that region. The other is the selection of a new classification scale, *without* a change in the currently attended location. Each attentional operation has lasting side effects on the master saliency representation. Selection of a new salient region involves inhibiting the currently selected salient region (if there is one), a process analogous to the spatial inhibition-of-return found in humans (Posner, 1980). Selection of a new classification scale also involves inhibition, namely inhibition of the currently selected classification scale. The way inhibition is implemented in each case is described in 'Appendix A.2.5'.

When a display is presented to the system, the first attentional operation is the selection of a random salient region in the master saliency representation, and the classification of this region at its default classification scale. The region's default classification scale is a function of its size: if it is large, the coarse-grained classification scale is the default; if it is small, the fine-grained scale is the default. If the region is large, it is then reanalyzed at the fine-grained classification scale, after which a new salient region is selected, by inhibiting the currently selected region and picking another one at random. If the region is small, it is not reanalyzed at a finer classification scale, since the model only features two spatial scales; instead, a new salient region is selected immediately. Thus each large region is analyzed successively at two spatial scales, while each small region is analyzed at just one. This cycle continues until all the salient regions in the original stimulus have been selected and inhibited.

There is one important point to note about the group classification of the local form of a large salient region. The classifier will only be able to provide

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

information about the form of these visual elements if they are *homogeneous*: as noted in Section 6.1, group classification is only possible if all objects are of the same type. But since textural homogeneity is one of the properties used to determine salient regions, as discussed in Section 6.2.2, this will quite frequently be the case.[1] By using homogeneity as a cue to the formation of salient regions, and by allowing salient regions to be reparsed at a finer classification scale, the attentional system naturally exploits the classifier's ability to operate on homogeneous groups.

7. Experiments and discussion

We now discuss the performance of our complete system, combining the attentional and classification subsystems, in two experiments, one to test its performance on Navon stimuli, the other to test its ability to distinguish between singular and plural stimuli. The system in both experiments uses a classifier which was trained on the simple large and small shapes described in Section 6.1.

7.1. PERFORMANCE ON NAVON STIMULI

The system was first tested on unseen hierarchically structured stimuli: large shapes made up of small shapes, as illustrated in Figure 6a. We created 90 unseen stimuli of this kind, formed from homogeneous groups of five small shapes each. In these groups, the small shapes were positioned closely enough together for the attentional system to recognize them as a single salient region, but otherwise as far apart as possible.

The system's typical performance is illustrated in Figures 6b and 6c. It first identifies the stimulus as a single (large) salient region, selects the default classification scale for a large region and classifies the stimulus at this scale, to identify its global form, as shown in Figure 6b. A successful classification at this scale indicates the presence of a single object in the attended region. The system then selects a finer classification scale and reclassifies the stimulus to identify its local form, as shown in Figure 6c. A successful classification at this scale indicates the presence of multiple objects of the given class, i.e., of a plural group. Over the 90 unseen Navon stimuli, the system successfully identified both the global and local shape of the stimulus in 78% of cases.

[1] The local visual elements in the selected region could also be heterogeneous, if they are grouped closely enough, as in the stimulus shown in Figure 5. But in fact our classifier provides useful information about this case too, by failing to offer a classification at all. In this case an additional attentional routine can be planned to analyze each local element of the region in turn. We have implemented this operation, but it goes beyond the scope of the current paper.

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

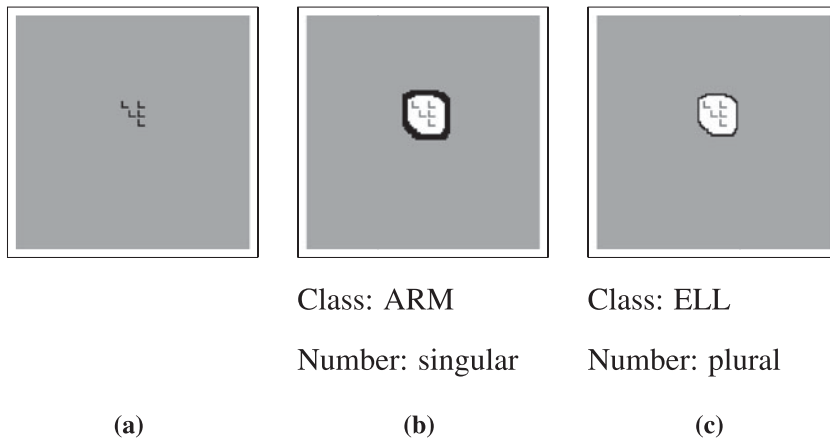

Fig. 6. Illustration of the system's performance on Navon stimuli. (a) A Navon stimulus (an arm formed from ells). (b) The attentional system initially identifies the stimulus as a single region and applies the default classification scale (indicated by the thick border); at this scale the classifier responds with **ARM**, indicating the presence of a single arm. (c) The attentional system then selects a finer classification scale for the same region (indicated by the thin border); at this scale the classifier responds with **ELL**, through group classification, indicating the presence of plural ells.

Note that our system's serial attentional algorithm processes global form before local form, in line with Navon's (1977) classic finding of global precedence. However, this precedence of global form over local form is simply stipulated in our algorithm; we do not want to draw any conclusions from this aspect of system performance. In fact, as noted in Section 5, global precedence can be modulated or reversed by priming particular spatial frequencies, as shown by Flevaris et al. (2011). When examining our system's performance on Navon stimuli, our main concern is not to show a particular ordering of local and global classification, but just to show that our system can classify a stimulus at two spatial scales, one after another. This capacity is not stipulated; it results from interactions between several different independently motivated system components: the saliency module's use of textural homogeneity in identifying salient regions, the attentional module's implementation of two competing classification scales, and the classifier's cardinality blindness.

The system's performance in classifying the local and global form of homogeneous stimuli (Navon stimuli) is somewhat worse than its performance in classifying group stimuli and large shapes by themselves (see Section 6.1). Its lower group classification performance is due to small shapes being packed closer together than is optimal in the current experiment. This creates artifactual shapes at positions in between the small shapes, which can confuse

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

the classifier. By contrast, the system's lower performance in classifying large shapes is due to small shapes being packed together more sparsely than is optimal for this task. In fact, the large shapes the system was tested on in the current experiment are quite different from the large shapes it was trained on. We expect that the system's performance on Navon stimuli would improve if it were explicitly trained on large shapes formed from smaller shapes. But our main purpose in the current paper is just to demonstrate that an ability to process such stimuli emerges reasonably naturally from the system without any specific training.

7.2. PERFORMANCE IN DISTINGUISHING BETWEEN SINGULAR AND PLURAL STIMULI

Our second experiment tested the system's ability to distinguish between single and plural visual stimuli within the subitization range more systematically. The system was presented with unseen stimuli consisting either of a single small shape or of a group of two, three, four, or five small shapes. If the system can distinguish between singular and plural, it should be able to classify individual small shapes at the default classification scale, thus recognizing these as singular instances of a given class, and to classify groups of size two to five at the finer-than-default classification scale, uniformly recognizing them as plural instances of a given class. (Its ability to classify the global form of group stimuli is not at issue here; this was assessed in the previous experiment.)

Each individual stimulus was a single small shape of some given category *C* presented at a random location on the retina. Each group stimulus was a homogeneous group of small shapes of category *C*, arranged in a random configuration, again placed at a random location on the retina. The density of group stimuli was again controlled, so that each group stimulus would be identified by the attentional system as a single salient region: in each group, no small shape was more than three pixels distant from some other shape, and each shape was at least two pixels distant from all other shapes. The category *C* varied over all the shapes the classifier was trained on. Plural stimuli ranged in size from two to five small shapes. There were twenty stimuli of each cardinality. Example stimuli are shown in Figure 7.

The system's ability to identify single *C*s and plural groups of *C*s in our test stimuli is charted in Table 2. As the table shows, the system reliably identifies single stimuli as singular; for groups of size two to five its performance varies from 95% to 83%. The limiting factor in identifying plural stimuli is the performance of group classification, which is markedly worse in the current experiment than when the classifier operates by itself, when it has an accuracy of 97–98% (see Section 6.1). This drop in performance is again because the shapes in the group stimuli in the current experiment are packed

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

TABLE 2. *Singular–plural attribution performance for a range of stimuli of different cardinalities*

Group size	Desired number judgement	Performance
1	singular	100%
2	plural	95%
3	plural	85%
4	plural	90%
5	plural	83%

more densely than is optimal for group classification. This is necessary in order for the attentional system to identify each group as a single salient region.

How to improve the classifier's performance on tightly packed groups is a matter for further research. It may be that tightly packed homogeneous groups are recognized as much by their visual texture as by the form of individual elements; indeed the boundary between texture classification and group classification is still unclear. Our present purpose is to demonstrate with some simple examples how our model of visual attention and classification delivers judgements about singular and plural visual stimuli. The key result in the current experiment is that the model has a qualitatively different response to individual *C*s and to groups of two to five *C*s: in the former case, the classifier outputs *C* when using the default classification scale, identifying the stimulus as singular; in the latter cases, it outputs *C* when using a finer-than-default scale, identifying these stimuli uniformly as plural.

8. Comparisons with existing visual models

There are many existing models of the relationship between visual attention and object classification. In this section, we discuss how our model compares to these.

A key point of variation among models of visual attention and object classification is whether these processes take place in separate pathways or a single pathway. Some models see attention and object classification as happening in two separate processing streams, so that attentional effects result from modulation of a 'classification stream' by an 'attentional stream'. Others see object classification and attention as happening in a single stream, so that attentional effects emerge naturally within the processing stream which computes the properties of objects. Our model clearly falls into the former camp. It is closely related to several other two-stream models in which a map of salient locations functions to bias processing in a separate object classification pathway. The closest model is probably that of Walther and Koch (2006), which uses a modified saliency map for the attentional processing stream and a variety of convolutional neural network (Riesenhuber & Poggio, 1999) for

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

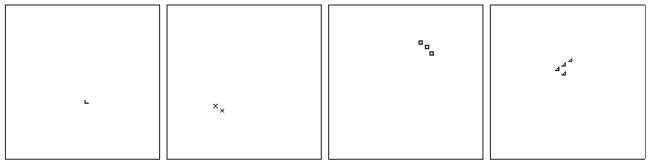

Fig. 7. Example stimuli used in assessing plurality and singularity.

the classification stream. Our modification of the saliency computation is also somewhat similar to that of Walther and Koch (2006), in that it computes salient regions of arbitrary shapes, intended to correspond to the shapes of the actual objects in the scene which are to be classified. Walther and Koch's method for establishing a salient region is to identify a salient point, determine the low-level visual features most responsible for its salience, and then spread activity to neighbouring points with similar low-level visual features. Our method achieves a similar effect by according salience to neighbouring points with similar low-level features.

Another type of two-stream model is the 'dynamic routing' model of Olshausen, Anderson, and van Essen (1993) (see Heinke & Humphreys, 2003, for an implementation and extension). While a convolutional neural network interleaves the computations involved in classification and abstraction over retinal location, the dynamic routing model sees these processes as more separable, with spatial abstraction happening at a particular stage relatively early in the processing pipeline, to create an 'object-centred' image representation which forms the input to a separate object classification system. Spatial abstraction is effected by a network with a very rich system of connections, which can 'route' information from any region of the retina to the object-centred medium. The attentional system selectively gates these connections, so that information from only one region is routed. In fact, the difference between a gated convolutional network and selective routing is qualitative rather than quantitative. While the implementations of Olshausen et al. (1993) and Heinke and Humphreys (2003) route low-level information about luminance on the retina, they both allow that the information which is routed could also be about higher-level visual features. In a convolutional neural network, each spatial abstraction layer implements something like a (very local) routing operation. However, the mechanism for selecting a salient region must be implemented differently in the two models. In our model, attention just modulates the input to the classifier, so that information only enters from a selected area. In selective routing, what is selected is not a retinal region but a whole mapping from retinal to object-centred coordinates. This involves modulation of a complex system of synaptic connections, which must be defined by their lateral relationships with one another, rather than just their retinotopic location. Selective routing is very good at modelling

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

some interactions between attention and object classification, in particular object-centred neglect (see Heinke & Humphreys, 2003). But it is not clear how it would model the processing of homogeneous group stimuli of the kind discussed in the current paper. Presumably, each stimulus in a homogeneous group would have to be mapped to the same object-centred representation. But in selective routing, alternative mappings normally inhibit one another, so that exactly one is picked. Some mechanism would have to be able to override this inhibition to allow identical tokens in a homogeneous group to be processed in parallel.

A representative single-stream model is the selective tuning model of Tsotsos and colleagues (see Tsotsos, Culhane, Wai, Lai, Davis, & Nuflo, 1995, for the original proposal, and Rothenstein & Tsotsos, 2008, for a recent exposition). This model uses a structure somewhat similar to a convolutional neural network: it features a hierarchy of processing layers, modelling processing in the ventral pathway, computing progressively more complex features, with increasingly coarse spatial resolution. However, in this system, each layer also functions as a saliency map in its own right. At each layer, beginning with the highest, a winner-take-all (WTA) mechanism operates to select a single region of the retina. WTA mechanisms at lower levels only operate over units which contribute to the selected region at the layer above, so that attentional effects percolate from the highest layer down. This model is attractive in that it allows attention to select visual features of different degrees of complexity, from very simple to very abstract. It also supports an interesting model of attentional grouping: the top-down constraints on WTA allow groups of stimuli to be selected in a given layer even if they are not spatially contiguous, if these stimuli all contribute to the winning units in the layer above. This allows a good account of how the discontiguous elements of an object can be selected if it is partially occluded. However, it is not clear how it should be extended to model the cardinality effects we discuss in the current paper. In our model, having two streams allows two distinct notions of spatial scale, which can vary independently: the dorsal stream represents the size of an attended region, while the ventral stream represents the size of the primitive visual features used to classify the contents of this region. We suggest that any account of the distinction between singular and plural cardinality must make reference to these two distinct notions of scale. If salient regions are identified internally to the object classification pathway, it is less easy to keep them separate. It may be possible to restrict the WTA process in given layers to units with a particular selected scale, but it is not clear at which layer this should apply. In any case, we have already reviewed evidence that cardinality is primarily computed in the dorsal visual pathway, so whatever mechanisms are responsible for computing cardinality seem likely to involve a separate processing stream.

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

9. Summary and further work

In this paper we proposed that the semantics of concrete noun stems and their number inflections can be read directly from the perceptual system, from the representations generated by the visual object categorization system and the visual attention system. We expressed this hypothesis in a model of visual object categorization and visual attention with two innovative features: first, the categorization system can classify homogeneous groups as well as individual objects; second, the attentional system can specify the scale at which the categorization system operates on a given salient region. Both these features are well motivated empirically in their own right; in combination, they result in a system that economically explains the dissociation of type and number information in language, and provides the first embodied account of the semantics of NUMBER features. However, there are many issues still to be explored; we conclude by briefly noting some of these.

First, our hypothesis about a link between linguistic number and local/global visual attention generates several predictions that could be tested empirically. Domahs et al.'s (2012) finding that the left TPJ is activated both during interpretation of PLURAL nouns and during attention to the local form of visual stimuli (see Section 4.2.2) is consistent with our hypothesis, but its involvement in these two processes may just be a coincidence, and the TPJ has many subregions, with many functions. An fMRI study explicitly comparing activity due to PLURAL nouns and to local visual attention (in the same subjects) would provide a better indication of whether the same neural region is involved in the two processes. A priming paradigm may also be able to test our proposal: if the semantic representations of SINGULAR and PLURAL nouns endure in time, our model predicts that interpreting SINGULAR and PLURAL nouns will bias visual attention towards the global and local form of subsequently presented Navon stimuli.

Second, the visual model presented in this paper is quite simple: several improvements are required. An important extension is to enable processing of more naturalistic images, both in the saliency and classification pathways, and to simulate more spatial scales, simulating the spatial frequency channels found in actual human vision (see, e.g., Hughes, George, & Kitterle, 1996). Another important extension is to include a mechanism for representing subitizable numbers as individuals, so that our account of the perceptual origin of the NUMBER feature can include the values DUAL and TRIAL. A further extension is to provide an account of texture classification and its relationship with group classification, to improve performance of the classifier on densely packed groups, and on groups with large cardinalities.

Third, the present account of noun stems and their number inflections must be extended to cover perceptual modalities other than vision – and also

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

to cover abstract uses of nouns. The classification of homogeneous groups of objects in other modalities has not yet been studied, to our knowledge. Hearing is a modality where group classification seems possible. Agents can use hearing to classify a single dog, for instance by its bark: Is the sound of many dogs barking simultaneously classified any differently? How is the distinction between one and many made here? Abstract nouns require further extensions to the theory, but here there are many accounts to draw on; the basic proposal is likely to echo the general proposal espoused by embodied semanticists, that the meanings of abstract words are grounded in concrete sensorimotor domains (see Lakoff & Johnson, 1980, and much subsequent work).

Finally, our embodied model of NUMBER features needs to consider the syntactic role that these features play, as well as their semantics. As discussed in Section 2, the syntax of a language often requires agreement between the NUMBER features of words within a noun phrase or clause, indicating that number features have a syntactic domain which extends beyond the words they appear in. Is there anything in our perceptually grounded account of NUMBER features which helps to explain their extended syntactic domain? If there is, this would supply interesting additional evidence for our proposal.

In summary, there are many directions for further development of our implemented model, and of the embodied theory of nouns which it expresses. The model nonetheless provides a useful platform on which more detailed and comprehensive hypotheses about the perceptual origins of nouns can be developed and quantitatively tested.

REFERENCES

Ansari, D., Lyons, I., van Eimeren, L., & Xu, F. (2007). Linking visual attention and number processing in the brain: the role of the temporo-parietal junction in small and large symbolic and nonsymbolic number comparison. *Journal of Cognitive Neuroscience*, **19**(11), 1845–1853.

Barner, D., Wood, J., Hauser, M., & Carey, S. (2008). Evidence for a non-linguistic distinction between singular and plural sets in rhesus monkeys. *Cognition*, **107**, 603–622.

Barsalou, L. (2008). Grounded cognition. *Annual Review of Psychology*, **59**, 617–645.

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, **86**, 201–221.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, **4**(2), 159–219.

Baylis, G., Driver, J., & Rafal, R. (1993). Visual extinction and stimulus repetition. *Journal of Cognitive Neuroscience*, **5**(4), 453–466.

Bergen, B., & Chang, N. (2005). Embodied construction grammar in simulation-based language understanding. In J. O. Östman & M. Fried (Eds.), *Construction grammar(s): cognitive and cross-language dimensions* (pp. 147–190). Amsterdam: John Benjamins. [Reprinted in V. Evans, B. Bergen, & J. Zinken (Eds.), *The cognitive linguistics reader*. Equinox. 2007.]

Cantlon, J., Brannon, E., Carter, E., & Pelphrey, K. (2006). Functional imaging of numerical processing in adults and 4-y-old children. *PLoS Biology*, **4**(5), 844–854.

Chierchia, G. (1998). Plurality of mass nouns and the notion of 'semantic parameter'. In S. Rothstein (Ed.), *Events and grammar* (pp. 53–103). Dordrecht: Kluwer.

Connolly, A., Guntupalli, S., Gors, J., Hanke, M., Halchenko, Y., Wu, Y.-C., Abdi, H., & Haxby, J. (2012). The representation of biological classes in the human brain. *Journal of Neuroscience*, **32**(8), 2608–2616.
Conway, B., Moeller, S., & Tsao, D. (2007). Specialized color modules in macaque extrastriate cortex. *Neuron*, **56**, 560–573.
Corbett, G. (2000). *Number*. Cambridge/New York: Cambridge University Press.
Damasio, H., Grabowski, T., Tranel, D., Hichwa, R., & Damasio, A. (1996). A neural basis for lexical retrieval. *Nature*, **380**, 499–505.
Domahs, F., Nagels, A., Domahs, U., Whitney, C., Wiese, R., & Kircher, T. (2012). Where the mass counts: common cortical activation for different kinds of nonsingularity. *Journal of Cognitive Neuroscience*, **24**(4), 915–932.
Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: object-files versus analog magnitudes. *Psychological Science*, **13**, 150–156.
Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, **8**(7), 307–314.
Feldman, J., & Narayanan, S. (2004). Embodiment in a neural theory of language. *Brain and Language*, **89**(2), 385–392.
Fias, W., & Fischer, M. (2005). Spatial representation of numbers. In J. Campbell (Ed.), *Handbook of mathematical cognition* (pp. 43–54). New York: Psychology Press.
Fink, G., Halligan, P., Marshall, J., Frith, C., Frackowiak, R., & Dolan, R. (1996). Where in the brain does visual attention select the forest and the trees. *Nature*, **382**, 626–628.
Flevaris, A., Bentin, S., & Robertson, L. (2010). Local or global? Attentional selection of spatial frequencies binds shapes to hierarchical levels. *Psychological Science*, **21**(3), 424–431.
Flevaris, A., Bentin, S., & Robertson, L. (2011). Attention to hierarchical level influences attentional selection of spatial scale. *Journal of Experimental Psychology: Human Perception and Performance*, **37**(1), 12–22.
Gonzalez, R. C., & Woods, R. C. (1992). *Digital image processing*. Reading, MA: Addison-Wesley.
Gottlieb, J., Kusunoki, M., & Goldberg, M. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, **391**, 481–484.
Heinke, D., & Humphreys, G. (2003). Attention, spatial representation, and visual neglect: simulating emergent attention and spatial memory in the selective attention for identification model (SAIM). *Psychological Review*, **110**(1), 29–87.
Hughes, H. C., George, N., & Kitterle, F. (1996). Global precedence, spatial frequency channels and the statistics of natural images. *Journal of Cognitive Neuroscience*, **8**(3), 197–230.
Hurford, J. (2001). Languages treat 1–4 specially. *Mind and Language*, **16**, 69–75.
Hurford, J. (2003). The interaction between numerals and nouns. In F. Plank (Ed.), *Noun phrase structure in the languages of Europe* (pp. 561–620). Berlin: de Gruyter.
Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, **40**, 1489–1506.
Izard, V., Dehaene-Lambertz, G., & Dehaene, S. (2008). Distinct cerebral pathways for object identity and number in human infants. *PLoS Biology*, **6**(2), 275–285.
Just, M., Cherkassky, V., Aryal, S., & Mitchell, T. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS One*, **5**(1), e8622.
Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, **45**(2), 83–105.
Kadir, T., Hobson, P., & Brady, M. (2005). From salient features to scene description. In *Workshop on Image Analysis for Multimedia Interactive Services*.
Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Dordrecht: Kluwer Academic Publishers.
Kanwisher, N. (1991). Repetition blindness and illusory conjunctions: errors in binding visual types with visual tokens. *Journal of Experimental Psychology: Human Perception and Performance*, **17**, 414–421.
Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, **4**, 219–375.
Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, **3**(9), 946–953.

Kriegeskorte, N., Mur, M., Ruff, D., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, **60**, 1120–1141.

Lakoff, G. (1987). *Women, fire and dangerous things*. Chicago/London: University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago/London: University of Chicago Press.

Liu, X., & Wang, D. (2000). Texture classification using spectral histograms. Technical Report TR17, Department of Computer and Information Science, Ohio State University, Columbus, OH 43210–1277.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, **19**, 577–621.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, **58**, 25–45.

Moore, T., & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, **421**, 370–373.

Mozer, M. C., & Baldwin, D. S. (2008). Experience-guided search: a theory of attentional control. In J. Platt, , D. Koller, & Y. Singer (Eds.), *Advances in neural information processing* 20 (pp. 1033–1040). Cambridge, MA: MIT Press.

Mozer, M. C., & Sitton, M. (1998). Computational modeling of spatial attention. In H. E. Pashler (Ed.), *Attention* (pp. 341–393). Hove: Psychology Press.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, **45**(2), 205–231.

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cognitive Psychology*, **9**, 353–383.

Nieder, A., & Miller, E. K. (2004). A parieto-frontal network for visual numerical information in the monkey. *Proceedings of the National Academy of Sciences*, **101**(19), 7457–7462.

Olshausen, B., Anderson, C., & van Essen, D. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, **13**(11), 4700–4719.

Palomares, M., & Egeth, H. (2010). How element visibility affects visual enumeration. *Vision Research*, **50**, 2000–2007.

Peggy Li, P., Barner, D., Ogura, T., Yang, S., & Carey, S. (2009). Does the conceptual distinction between singular and plural sets depend on language? *Developmental Psychology*, **45**(6), 1644–1653.

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, **44**, 547–555.

Piazza, M., Mechelli, A., Price, C., & Butterworth, B. (2006). Exact and approximate judgements of visual and auditory numerosity: an fmri study. *Brain Research*, **1106**, 177–188.

Polyn, S., Natu, V., Cohen, J., & Norman, K. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, **310**, 1963–1966.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, **32**(1), 3–25.

Quinlan, P. T., & Wilton, R. N. (1998). Grouping by proximity or similarity? Competition between the gestalt principles in vision. *Perception*, **27**, 417–430.

Riedmiller, M. (1994). Rprop − description and implementation details. Technical report, Institut für Logik, Komplexität und Deduktionssyteme, University of Karlsruhe.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, **2**(11), 1019–1025.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, **12**, 162–168.

Robertson, L. (1996). Attentional persistence for features of hierarchical patterns. *Journal of Experimental Psychology: General*, **125**, 227–249.

Robertson, L., Lamb, M., & Knight, R. (1988). Effects of lesions of temporo-parietal junction on perceptual and attentional processing in humans. *Journal of Neuroscience*, **8**, 3757–3769.

Rolls, E. T., & Deco, G. (2006). Attention in natural scenes: neurophysiological and computational bases. *Neural Networks*, **19**, 1383–1394.
Rothenstein, L., & Tsotsos, J. (2008). Attention links sensing to recognition. *Image and Vision Computing*, **26**, 114–126.
Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, volume 1: *Foundations* (pp. 318–362). Cambridge, MA: MIT Press.
Sarnecka, B., Kamenskaya, V., Yamana, Y., Ogura, T., & Yudovina, Y. (2007). From grammatical number to exact numbers: early meanings of 'one,' 'two,' and 'three' in English, Russian, and Japanese. *Cognitive Psychology*, **55**(2), 136–168.
Schyns, P., & Oliva, A. (1999). Dr Angry and Mr Smile: when categorization flexibly modifies the perception of faces in rapid serial visual presentations. *Cognition*, **69**, 243–265.
Shinkareva, S., Mason, R., Malave, V., Wang, W., Mitchell, T., & Just, M. (2008). Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS One*, **3**(1), e1394.
Sowden, P., & Schyns, P. (2006). Channel surfing in the visual brain. *Trends in Cognitive Sciences*, **10**(12), 538–545.
Sudre, G., Pomerleau, D., Palatucci, M., Wehbe, L., Fyshe, A., Salmelin, R., & Mitchell, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, **62**, 51–463.
Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, **19**, 103–139.
Theeuwes, J. (1994). The effects of location cuing on redundant-target processing. *Psychological Research*, **57**, 15–19.
Thompson, K., & Bichot, N. (2005). A visual salience map in the primate frontal eye field. *Progress in Brain Research*, **147**, 251–262.
Tranel, D., Adolphs, R., Damasio, H., & Damasio, A. (2001). A neural basis for the retrieval of words for actions. *Cognitive Neuropsychology*, **18**(7), 655–674.
Trick, L., & Pylyshyn, Z. (1994). Why are small and large numbers enumerated differently? A limited capacity preattentive stage in vision. *Psychological Review*, **101**, 80–102.
Tsotsos, J., Culhane, S., Wai, W., Lai, Y, Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, **78**, 507–545.
Ullman, S. (1996). *High-level vision: object recognition and visual cognition*. Cambridge, MA: MIT Press.
Ungerleider, L. A., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.
Van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, **9**, 333–377.
Vigliocco, G., Vinson, D., Druks, J., Barber, H., & Cappa, S. (2011). Nouns and verbs in the brain: a review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience and Biobehavioral Reviews*, **35**, 407–426.
Walles, H., Knott, A., & Robins, A. (2008). A model of cardinality blindness in inferotemporal cortex. *Biological Cybernetics*, **98**(5), 427–437.
Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, **19**(9), 1395 – 1407.
Wolfe, J. M. (1994). Guided search 2.0 – a revised model of visual-search. *Psychonomic Bulletin & Review*, **1**(2), 202–238.
Wolfe, J. M. (2007). Guided search 4.0: current progress with a model of visual search. In W. Graw (Ed.), *Integrated models of cognitive systems* (pp. 99–119). New York: Oxford University Press.
Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J., & Poldrack, R. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science*, **330**, 97–101.
Yamaguchi, S., Yamagata, S., & Kobayashi, S. (2000). Cerebral asymmetry of the 'top-down' allocation of attention to global and local features. *Journal of Neuroscience*, **20**(9), RC72.

Zhang, Y., Meyers, E., Bichot, N., Serre, T., Poggio, T., & Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proceedings of the National Academy of Sciences of the USA*, **108**(21), 8850–8855.

Zwaan, R., & Taylor, L. (2006). Seeing, acting, understanding: motor resonance in language comprehension. *Journal of Experimental Psychology: General*, **135**(1), 1–11.

Appendix: technical details of the visual model

The model of the classification and attentional subsystems can be thought of as a collection of retinotopic map representations. We implement a map as a matrix. The input map **I** is a greyscale image measuring 128 × 128 pixels, with element values in the range 0 to 255. The input is read directly from bitmap image files. Other maps are computed from this using a variety of operations. Most of these maps also measure 128 × 128 pixels with the exception of some employed by the classifier (see Section A.1). Except where noted it is safe to assume that the output of a map operation has the same dimensions as its inputs. Because of this we sometimes refer to pixels in maps other than **I** even though they do not, strictly speaking, form an image.

The map operations used are convolution (matrix convolution, denoted * with pixels lying beyond the map edge assumed to be white unless noted otherwise), addition, subtraction, and scalar multiplication (computed as for their matrix equivalents), modulus (denoted $|\mathbf{X}|$, computed by taking the modulus of each element), and some more complicated operations which will be defined where they occur. The matrix or map element at the ith row and jth column of $\mathbf{X}$ is denoted $\mathbf{X}_{i,j}$.

Some special maps have additional information associated with them, such as regions. Regions are sets of contiguous pixels in a map and we implemented these as either maps with characteristic pixel values for each region or as sets of maps, one per region, depending on which was more convenient. The map itself can still be considered just a matrix, with this extra information represented separately and bound to the map.

A.1. The classifier

A.1.1. Classifier structure

The classifier used was a convolutional neural network which takes a set of input maps and activates a set of output category units via a series of layered plies which alternately combine visual features from the ply below into more complex features and abstract over the spatial location of visual features. The CNN was mostly as described in Walles et al. (2008), except that the number of features used in each ply was different and there was some additional input preprocessing.

Figure 8 illustrates the overall structure of our CNN.

The **units** of the network are arranged in a series of **plies**, with units in each ply connected to units in the one above by a **layer** of **weights**.

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

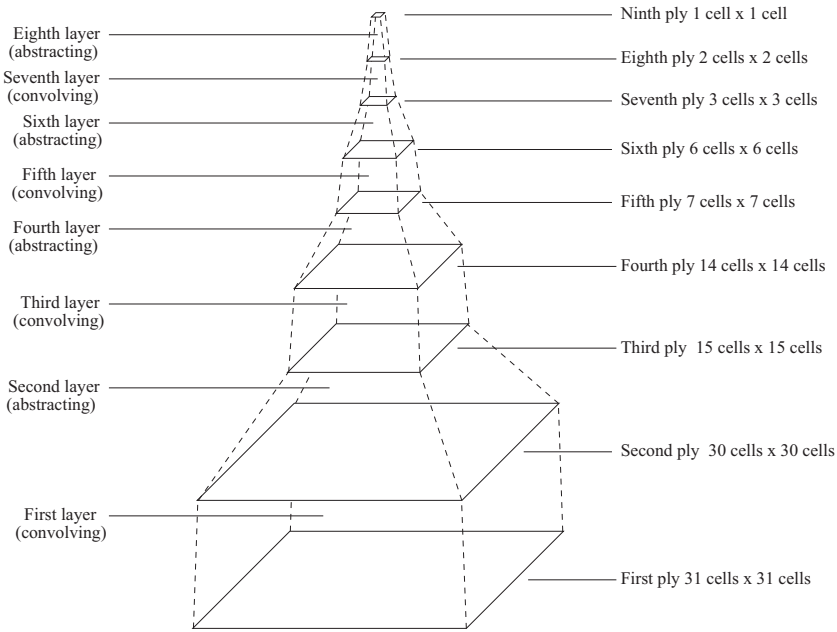

Fig. 8. The general structure of our convolutional neural network. Plies of cells are connected by layers of weights. Each cell contains one unit for every feature represented by that ply.

Our network had nine plies and eight layers. Units within each ply are clustered into **cells**, which are arranged retinotopically. Every cell in a particular ply contains the same number of units, one for each feature that the ply represents. Each unit in a cell represents the strength of its associated feature at the cell's location, and so each cell in a ply represents in parallel the presence of a set of features at the corresponding location in the input field. The successive plies of our network (going from input to output, measured in terms of cells) are 31×31, 30×30, 15×15, 14×14, 7×7, 6×6, 3×3, 2×2, and 1×1.

The features in the first (input) ply were divided into two groups. One feature was provided for high-frequency input which represented luminance directly. Four features represented low-frequency input: these were obtained with 9×9 convolution filters tuned to horizontal, vertical, and diagonal black-on-white lines.

Units receive input from a small square region of the ply beneath, the **integration window**, meaning they are connected locally and can only make use of local features. We used a window measuring 2×2 cells in all layers. All units in a cell have the same window, which can thus also be called the cell's window. The region of the retina that contributes to a unit's input is its

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

receptive field. In addition, the weights for corresponding units in different cells of a ply are constrained to be the same, effectively sharing the weights. This means that the response to activity inside a cell's window will be the same irrespective of where in a ply the cell is located.

Successive plies divide the visual field more and more coarsely, so contain fewer cells than their predecessors, each of which has a wider receptive field than those in earlier plies. However, later plies generally represent more features than earlier plies, and therefore contain more units per cell.

The function and structure of the weight layers alternates throughout the network between convolution and abstraction. Convolving layers compute combinations of features in the previous ply with little change in the number of cells between plies, while abstracting layers reduce the number of cells of the input ply without interaction between different features.

In convolving layers, an output unit receives input from every unit within its 2 × 2 integration window. A unit receiving input from a ply representing n features will have $4n + 1$ inputs (including a bias).

Weights in abstracting layers are simpler. Input and output plies contain the same number of features and there is no interaction between features. A unit receiving input from its 2 × 2 window will have five inputs (including a bias). The window of a cell in the output ply precisely abuts but does not overlap with the windows of neighbouring cells. The effect is that the integration windows of cells in the output ply tile the input ply. Weights are shared even further within abstracting layers, with all weights for a feature constrained to be identical. This means that each abstracting layer really has only two variable parameters per feature: one weight shared among all the inputs units, and the bias.

Apart from the varying structure of the layers, unit activation is computed in the same way throughout the network. For a unit with n inputs $p_1 \ldots p_n$ (excluding the bias) and $n + 1$ weights (including the bias) $w_1 \ldots w_{n+1}$ the unit's **activation**, a weighted sum, σ is computed:

$$\sigma = \sum_{i=1}^{n} p_i w_i + w_{n+1}$$

which, for an abstracting unit, can be simplified further to:

$$\sigma = w_{ply} \sum_{i=1}^{n} p_i + w_{bias}$$

because of weight sharing.

The output of the unit is then computed via the logistic function:

$$f = \frac{1}{1 + e^{-\sigma}}$$

This is conventional for feed-forward networks.

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

Going from the input ply to the output ply the number of features in each ply were 5, 25, 25, 32, 32, 32, 32, 7, and 7.

Although inputs to the system as a whole measure 128 × 128 pixels, inputs to the classifier always measure 31 × 31 pixels as in our original design. This is a practical limitation of the classifier to allow training in reasonable time, and the disparity is resolved by always centring the attended region in the classifier's input for classification purposes. This ensures that the bounding rectangle of the attended region is centred in the classifier's input.

A.1.2. Training regime

The network was trained using the RPROP algorithm (Riedmiller, 1994). This is a variation of the BACKPROP algorithm (Rumelhart, Hinton, & Williams, 1986). The training algorithm is described in more detail in Walles et al. (2008).

We trained with small (high-frequency) shapes, each presented at a randomly chosen third of possible retinal locations. We also trained with large (low-frequency) shapes each at a random third of all possible retinal locations for each of four densities. These included solid shapes as well as large shapes with pixels randomly ablated to the background colour with probabilities $\frac{1}{6}$, $\frac{1}{3}$, or $\frac{1}{2}$. Thus, for the low-frequency training, total spatial coverage was likely. The small shapes were presented at the high-frequency inputs only and the large shapes at the low-frequency inputs only. During operation only one of the sets of inputs is used at a time, the other being suppressed entirely. There were 1566 high-frequency training examples and 2152 low-frequency training examples. As in Walles et al. (2008), these included 371 noise examples which were each fed to the low- and high-frequency inputs in turn. New noise examples were generated on each cycle of training. In other respects the architecture and training of the CNN was as described in Walles et al. (2008).

A.2. Parallel attention component: saliency analysis

Saliency analysis is based on the model presented by Itti and Koch (2000) and Walther and Koch (2006), modified to fit the size constraints of the classifier and support scale-based attention in the selection mechanism.

A.2.1. Local contrast

Local contrast computation begins by taking the input image (a luminance image) and stretching the values into the range −128 (black) to 127 (white).

$$\mathbf{I}' = stretch(\mathbf{I}, -128, 127), \text{where} \tag{5}$$

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

$$stretch(\mathbf{X},L,U)_{i,j}=\frac{\left(\mathbf{X}_{i,j}-min(\mathbf{X})\right)(U-L)}{max(\mathbf{X})-min(\mathbf{X})}+L \tag{6}$$

and *min* and *max* are functions that produce the minimum and maximum element values, respectively, of a matrix or map. Local contrast is then computed by convolving with two normalized Laplacian of Gaussian filters, one for each spatial frequency ($\sigma = 1$ and $\sigma = 15$, chosen by trial and error to produce strong response to shapes of the relevant scale while trying to minimize response to shapes at the other scale). The absolute value of these results is then taken. Given

$$LoG(\sigma)_{i,j}=\left(1-\frac{r^2}{\sigma^2}\right)e^{-\frac{r^2}{\sigma^2}}, \text{where} \tag{7}$$

$$0\le i,j<5\sigma$$

$$o=\frac{5\sigma}{2}$$

$$r^2=(i-o)^2+(j-o)^2 \tag{8}$$

and normalization was achieved using

$$norm(\mathbf{X})_{i,j}=\frac{\mathbf{X}_{i,j}}{s}, \text{where} \tag{9}$$

$$s=\sum_i\sum_j|X_{i,j}|$$

we compute the high-frequency local contrast $\mathbf{C}_{\text{hi}}$ and low-frequency local contrast $\mathbf{C}_{\text{lo}}$ using

$$\mathbf{C}_{\text{hi}}=|norm(LoG(\sigma=1))*\mathbf{I}'| \tag{10}$$

$$\mathbf{C}_{\text{lo}}=|norm(LoG(\sigma=15))*\mathbf{I}'| \tag{11}$$

We use LoG filters here rather than the orientation-specific filters used in the classifier for two reasons. First, the orientation-specific filters used in the classifier grew out of the existing orientation-specific features used by Mozer and Sitton (1998), which our classifier is based on. Second, while one of the purposes of filtering the classifier inputs is to provide directed information (orientation) to aid classification, here we are only interested in contrast of suitably-sized shapes whatever their orientation. Having said that, it would be desirable in future to find a way to use the classifier's filters to produce these contrast maps instead of the LoG.

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

A.2.2. Homogeneity

The similarity measure is computed by the procedure described in Liu and Wang (2000). This procedure samples a small 7 × 7 pixel region around each pixel in the input image, computing its spectral histogram which can be thought of as a high dimensional feature vector, and finally finds the closest match to this histogram among those belonging to a set of texture templates derived from images of the small shapes used in the experiment both closely packed and sparsely scattered.

The spectral histogram is constructed by first convolving the 7 × 7 window with each of seven normalized filter matrices. The first three are the Kronecker $\boldsymbol{\delta}$ filter, which constitutes an identity operation in this instance and the D_{xx} and D_{yy} filters:

$$\delta = [1] \tag{12}$$

$$D_{xx} = [-1 \quad 2 \quad -1] \tag{13}$$

$$D_{yy} = \begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix} \tag{14}$$

There are also two Laplacian of Gaussian filters (see equation (7)) $LoG(\sigma = 1)$ and $LoG(\sigma = 2)$.

Finally there are three Gabor filters $G(\sigma = 2, \theta = \frac{\pi}{6})$, $G(\sigma = 2, \theta = \frac{\pi}{2})$ and $G(\sigma = 2, \theta = \frac{5\pi}{6})$ where

$$G(\sigma,\theta)_{i,j} = e^{\frac{-1}{2\sigma^2 r}} \cos\left(\frac{-2\pi}{\sigma}((j-o)\cos\theta + (i-o)\sin\theta)\right) \tag{15}$$

$$0 \le i, j < w = \frac{8\sigma}{\sqrt{2}}$$

$$o = \frac{w}{2}$$

$$r = \left((j-o)\cos\theta + (i-o)\sin\theta\right)^2 + \left((o-j)\sin\theta + (i-o)\cos\theta\right)^2$$

The filter matrices are each normalized with the *norm* function given in equation (9). Their choice is justified by Liu and Wang (2000). The window is convolved with each normalized filter with pixels at the edge of the map replicated to infinity to ensure a result for every pixel in the input. The histograms of the resulting maps (with unit-sized bins) are concatenated to produce the spectral histogram. Spectral histograms are compared using the

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

χ^2 value. If H_1 and H_2 are two spectral histograms, and $H(i)$ is the ith element of the histogram H then this is computed as follows.

$$\chi^2 = \sum_i \frac{\left(H_1(i) - H_2(i)\right)^2}{H_1(i) + H_2(i)} \tag{16}$$

For each pixel's associated histogram, the template histogram which has the lowest χ^2 value relative to it determines the category assigned to the pixel.

Once each pixel is assigned a category (square, ell, etc. or background), boundaries are determined by comparing each pixel with its four-neighbours. The four-neighbours of a pixel at coordinates (i,j) are the pixels at coordinates $(i-1, j)$, $(i+1, j)$, $(i, j-1)$, and $(i, j+1)$. Whenever a pair of pixels differs in category, the pixel that was least certainly classified (measured by the χ^2 of its histogram relative to its category's template) is marked as a texture boundary. In the resulting boundary map **B** homogeneous regions are marked with zero, boundaries with one.

For the experiments presented here we wanted some stimuli to be considered similar enough for saliency analysis to group them even though they were distinct. To this end we defined that boundaries between ells and squares, crosses and arrows, arrows and arms, arrows and triangles and triangles and arms would not be marked in the boundary map.

This confusion of types was based on the confusion patterns of the CNN but is effectively arbitrary. It is intended to model the Gestalt principle of similarity between types. We would have preferred to model this confusion using comparisons between the histograms of neighbouring pixels directly but the small size of the retina made this impractical (we consider this to be just an implementation detail).

A.2.3. Partial saliency maps

The boundary map **B** is combined with the low-frequency local contrast map $\mathbf{C}_{\text{lo}}$ by a weighted sum and thresholded to produce the low-frequency saliency map $\mathbf{S}_{\text{lo}}$:

$$\mathbf{S}_{\text{lo}} = H\left(\tau_{\text{lo}}; \alpha\mathbf{C}_{\text{lo}} - \beta\mathbf{B}\right) \tag{17}$$

where

$$H\left(\tau; \mathbf{X}\right)_{i,j} = \begin{cases} 1 & \text{if } \mathbf{X}_{i,j} > \tau \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

$$\tau_{\text{lo}} = 0.060 \tag{19}$$

$$\alpha = 15.5 \tag{20}$$

$$\beta = 1.45 \tag{21}$$

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

These scalings were chosen by trial and error so that contrast and homogeneity would interact without one dominating all the time. The high-frequency saliency map $\mathbf{S}_{\text{hi}}$ is just the same as high-frequency local contrast map, thresholded:

$$\mathbf{S}_{\text{hi}} = H\left(\tau_{\text{hi}}; \mathbf{C}_{\text{hi}}\right) \tag{22}$$

where

$$\tau_{\text{hi}} = 0.4 \tag{23}$$

The threshold values were chosen by trial and error so that regions of both frequencies at a reasonable contrast would become salient.

The ratio between contrast and homogeneity weights (α and β in equation (17)) determines the relative contributions of contrast and homogeneity to overall salience. Table 3 shows the effect on grouping behaviour of changing β while keeping α constant. Column 3 shows the maximum separation between heterogeneous stimuli for which they are grouped together, and column 4 shows the minimum separation between homogeneous stimuli for which they are treated as separate regions, for a range of different weight ratios. Distances are measured in pixels. The second row shows the parameter values used in the experiments in the current paper.

A.2.4. Combination of partial saliency maps

Regions which are four-neighbour contiguous are next identified and labelled by region merging (Gonzalez & Woods, 1992, see Section LABEL:sec:homogeneity for a definition of four-neighbouring pixels). Any labelled region in the low-frequency map containing fewer than 55 pixels is discarded, yielding a new low-frequency saliency map $\mathbf{S}'_{\text{lo}}$ which is used for further operations:

$$\mathbf{S}'_{\text{lo}} = F\left(\mathbf{S}_{\text{lo}}\right) \tag{24}$$

where F is a function that just sets pixels belonging to such regions in the input to zero.

This was done to remove high-frequency objects strong enough to stimulate the low-frequency saliency map as well as occasional artefacts between objects, both of which we consider to be noise. It acts as a kind of low-pass filter, removing regions too small to be of interest to the low-frequency map. The point-wise sum of these maps yields the master saliency map in which contiguous regions are also identified and labelled.

$$\mathbf{S} = \mathbf{S}'_{\text{lo}} + \mathbf{S}_{\text{hi}} \tag{25}$$

A.2.5. Inhibition and suppression

The preceding operations have been all bottom-up, but further computation relies on some top-down influence in the form of *inhibition*. A map is inhibited

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

TABLE 3. *The effect of changing homogeneity weight on grouping behaviour*

Contrast weight (α)	Homogeneity weight (β)	Maximum separation between grouped heterogeneous stimuli	Minimum separation between separate homogeneous stimuli
15.5	0.8	3	4
15.5	1.45	1	3
15.5	2.5	1	1

by combining a top-down *inhibition map* with its bottom-up activation. It can be thought of as an additional factor in the computation of the map. If $\mathbf{X}$ is a map and $\mathbf{X}^I$ its corresponding inhibition map then the inhibited version of the map $\mathbf{X}'$ (its effective value, used by operations which depend on the map) is given by

$$\mathbf{X}'_{i,j} = \begin{cases} \mathbf{X}_{i,j} & \text{if } \mathbf{X}^I_{i,j} = 0 \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

It is also possible to inhibit an entire map at once, equivalent to inhibiting with a map containing no zero elements.

In this paper we use the term **suppression** where inhibition is only temporary as part of a computation. Where applicable, inhibition and suppression are governed by independent inhibition maps associated with the primary map.

A.2.6. Computation of salient regions

The final stage of saliency analysis is the extraction of a well-defined set of salient regions, each tagged with a default classification scale based on its size. In our implementation these are represented in a series of maps, one for each salient region – though a single map with an appropriate coding could be used because the regions do not overlap. First, any of the most strongly activated pixels in the master saliency map is chosen (we used the left- and top-most such point but this is arbitrary). If there is a low-frequency salient region at that point, the low frequency is selected as the salient scale, otherwise the high frequency is selected. Standard morphological dilation (Gonzalez & Woods, 1992) is then applied to the corresponding region (radius 2 pixels for high frequency, 4 pixels for low frequency). Finally, pixels are removed from the region if they overlap salient regions that have already been computed, other active pixels in the master saliency map or pixels inhibited by attention operations (for which the associated inhibition map will be active).

The resulting region is added to the set of salient regions, tagged with its associated scale. The region is suppressed in the corresponding scale saliency

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press

map and the master saliency map and any overlapping regions in the non-selected scale saliency map are also suppressed. The above process is repeated until all activity in the master saliency map has been suppressed.

Once the set of salient regions is computed, one is chosen at random by the selection mechanism and the associated region and scale become the subjects of attention. We do not select salient regions by decreasing order of saliency, as is typically done, because our stimuli are very simple and the standard measure of 'degree of saliency' does not really apply. The randomization of selection can be viewed as the addition of noise to simulate the variation of saliency found in real-world stimuli.

After the winner is selected, suppression of the saliency maps introduced during computation of salient regions is then removed. Salient regions are recomputed whenever there is a change to the maps that the computation depends on, which happens when the selection mechanism inhibits the saliency maps.

A.3. Attentional selection operations

There are two kinds of gating operation: gating by location and gating by scale. A map $\mathbf{Y}$ containing a salient location spatially gates another map $\mathbf{X}$ with the result given by the *gate* function:

$$gate\left(\mathbf{X},\mathbf{Y}\right)_{i,j}=\begin{cases}\mathbf{X}_{i,j} & \text{if } \mathbf{Y}_{i,j}\neq 0\\ 0 & \text{otherwise}\end{cases} \tag{27}$$

Gating by scale is achieved by entirely inhibiting a scale-specific set of classifier input maps (equivalent to gating the maps with a map containing only zero elements). When the low frequency is selected, the high-frequency maps are entirely gated off and vice versa.

https://doi.org/10.1017/langcog.2014.9 Published online by Cambridge University Press