Flexible Spatial Language Behaviors: Developing a Neural Dynamic Theoretical Framework

John Lipinski (2johnlipinski@gmail.com)
Yulia Sandamirskaya (yulia.sandamirskaya@neuroinformatik.rub.de)
Gregor Schöner (gregor.schoener@neuroinformatik.rub.de)
Institut für Neuroninformatik, Ruhr-Universität Bochum, 44780 Bochum, Germany

Abstract

To date, spatial language models have tended to overlook process-based accounts of building scene representations and their role in generating flexible spatial language behaviors. To address this theoretical gap, we implemented a model that combines spatial and color semantic terms with neurally-grounded scene representations. Tests of this model using real-world camera input support its viability as a theoretical framework for behaviorally flexible spatial language.

Keywords: dynamical systems; neural networks; spatial cognition; spatial language.

Introduction

Spatial language is an incredibly flexible tool whose capabilities range from generating and comprehending directions (Tom & Denis, 2004) to facilitating coordinated action (Bangerter, 2004). Yet, despite this broad behavioral scope, implemented spatial language models which seek to uncover processes underlying basic spatial communication (e.g. object location description) have tended to focus on a limited range of behaviors, namely relational judgment tasks. These models have successfully accounted for a complex array of empirical data including the influence of landmark shape (Regier & Carlson, 2001) and functional object features (Coventry et al., 2005). The neural processing aspects underlying these accounts, however, remain underdeveloped. Consequently, a number of critical questions that bear directly on spatial language and its linkage to supporting sensory-motor processes have gone unaddressed. For example, how does a neural scene representation evolve on the basis of sensory information? How might complex higher-level behaviors like spatial language emerge from these lower-level dynamic processes? How are the time courses of spatial language behaviors structured by their roots in scene representations?

Behavioral flexibility in the spatial language system becomes a central issue once one addresses the neural processes that link spatial language to the sensory-motor system. Fundamentally, we do not yet understand how the sensory-motor foundations of scene representations and spatial language work to support the broad array of spatial language behaviors. The absence of process-based accounts for the generation of spatial scene representations and the behaviors derived from these representations is a significant barrier to developing a more comprehensive, integrative spatial language model.

As a step to overcoming this barrier, we were led to consider three elements underlying behavioral flexibility in spatial language. First, the spatial language system uses both spatial and non-spatial characteristics. Second, it integrates the graded sensory-motor representations with symbolic, linguistic terms. Finally, the spatial language system combines these numerous elements continuously in time according to the specific behavioral context.

To develop a behaviorally flexible theoretical framework for spatial language that satifies these constraints, one needs a representational language that links to both the sensory-motor and linguistic worlds. The Dynamic Field Theory (Erlhagen & Schöner, 2002), a neuronally based theoretical language emphasizing attractor states and their instabilities, is one viable approach. Recent applications of the DFT have extended beyond spatial working memory development (Spencer, Simmering, Schutte, & Schöner, 2007) to include a theoretically generative account of signature landmark effects in spatial language (Lipinski, Spencer, & Samuelson, in press). Critically, this latter work integrated a connectionist-style localist spatial term network into the model. This suggests that the DFT can provide the requisite, integrative representational language.

The present work incorporates this hybrid approach to implement a new model integrating spatial language semantics with real-world visual input. Our goal is to qualitatively test the model's core functionality and, thus, its viability as an initial theoretical framework for flexible spatial language behaviors. To rigorously test our model, we implement it on a robotic platform continously linked to real-world visual images of everyday items on a tabletop workspace. Our model extracts the categorical, cognitive information from the low-level sensory input through the system dynamics, not through neurally ungrounded preprocessing of the visual input. Models which do not directly link cognitive behavior to lower-level perceptual dynamics risk side-stepping this difficult issue. Our demonstrations specifically combine visual space, a selected subset of basic English spatial semantic terms, and color. These demonstrations serve as an initial proof of concept that takes an early step towards modeling more complex, natural spatial language behaviors.

Modeling neurons and dynamical neural fields

This section briefly reviews the mathematics of our model (see also (Erlhagen & Schöner, 2002)).

Dynamical fields

The dynamical neural fields are mathematical models first used to describe cortical and subcortical neural activation dynamics (Amari, 1977). The dynamic field equation Eq. (1) is a differential equation describing the evolution of activation u defined over a neural variable(s) $\mathbf{x}$. These neural variables represent continuous perceptual (e.g. color) or behavioral (e.g. reaching amplitude) dimensions of interest that can be naturally defined along a continuous metric.

$$\begin{aligned}\tau\dot{u}(\mathbf{x},t) &= -u(\mathbf{x},t)+h+\int f(u(\mathbf{x}',t))\omega(\Delta\mathbf{x})d\mathbf{x}'+ \\ &+I(\mathbf{x},t)\end{aligned} \qquad (1)$$

Here, $h<0$ is the resting level of the field; the sigmoid non-linearity $f(u)=1/(1+e^{-\beta u})$ determines the field's output at suprathreshold sites with $f(u)>0$. The field is quiescent at subthreshold sites with $f(u)<0$. The homogeneous interaction kernel $\omega(\Delta x)=c_{exc}e^{\frac{-(\Delta x)^2}{2\sigma^2}}-c_{inh}$ depends only on the distance between the interacting sites $\Delta x=\mathbf{x}-\mathbf{x}'$. This interaction kernel is a Bell-shaped (Gaussian), local excitation/global inhibition function. The short-range excitation is of amplitude c_{exc} and spread σ. The inhibition is global, as we are not interested in multipeak solutions here, and has an amplitude c_{inh}. $I(\mathbf{x},t)$ is the summed external input to the field; τ is the time constant.

If a localized input activates the neural field at a certain location, the interaction pattern ω stabilizes a localized ”peak”, or ”bump” solution of the field's dynamics. These activation peaks represent the particular value of the neural variable coded by the field and thus provide the representational units in the DFT (Spencer & Schöner, 2003).

In our model, all entities having ”field” in their name evolve according to Eq. (1), where $\mathbf{x}$ is a vector representing the two-dimensional visual space in Cartesian coordinates. The links between the fields are realized via the input term $I(\mathbf{x},t)$, where only sites with $f(u)>0$ propagate activation to other fields or neurons.

Discrete nodes

The discrete (localist) neural nodes in the model representing spatial and color semantic terms can be flexibly used for either user input or response output. Their activation evolves according to the dynamic equation (2).

$$\tau_d\dot{d}(t)=-d(t)+h_d+f(d(t))+I(t). \qquad (2)$$

Here, d is the activity level of a node; the sigmoidal nonlinearity term $f(d)$ shapes the self-excitatory connection for each discrete node and provides for self-stabilizing activation. The negative resting level is defined by h_d. The $I(t)$ term represents the sum of all external inputs into the given node. This summed input is determined by the input coming from the connected neural field, the user interface specifying the language input, and the competitive, inhibitory inputs from the other discrete nodes defined for that same feature group (color or space); τ is the time constant of the dynamics.

The spatial language framework

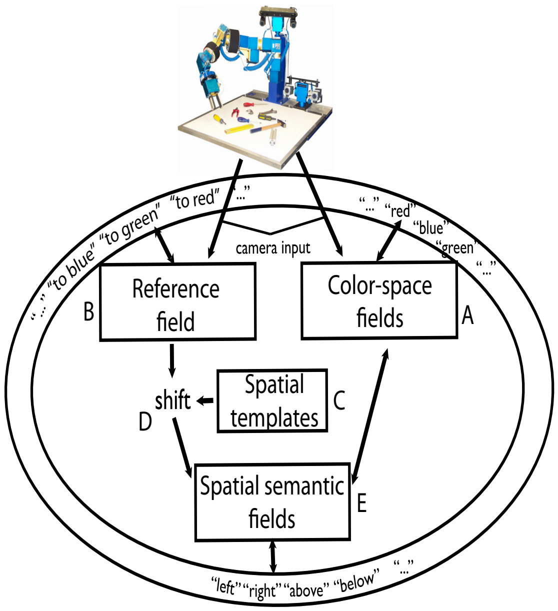

"..." "to blue" "to green" "to red" "..."
"..." "red" "blue" "green" "..."
camera input
B
Reference field
Color-space fields
A
shift
D
Spatial templates
C
Spatial semantic fields
E
"left" "right" "above" "below" "..."

Figure 1: Overview of the architecture

This section outlines the overall structure (see Fig. 1) of our integrative model and explains how it operates in two scenarios fundamental to spatial language: describing *where* an object is (Demonstration 1) and describing *which* object is in a specified spatial relation (Demonstration 2).

Color-space fields

The color-space fields (Fig. 1A) are an array of several dynamical fields representing the visual scene. Each of the fields is sensitive to a hue range which corresponds to a basic color. The resolution of color was low in the presented examples because only a few colors were needed to represent the used objects. In principle, the color (hue) is a continuous variable and can be resolved more finely. The stack of color-space fields is therefore a three-dimensional dynamic field that represents colors and locations on the sensor surface. The camera provides visual input to the color-space field, which is below the activation threshold before the task is defined. The field is thus quiescent to this point.

Once the language input specifying the *color* of the object activates the respective color-term node, however, the resting levels of all sites of the corresponding color-space field are raised homogeneously. Because the color-space fields receive localized camera input, this uniform activation increase is summed with that input to enable the development of an instability and, ultimately, the formation of a single-peak solution. This peak is centered over the position of the object with that specified color.

The *spatial* language input also influences the color-space field's dynamics through the aligned spatial semantic fields (see below).

Reference field

The reference field (Fig. 1B) is a spatially-tuned dynamic field which also receives visual input (Fig 1B). When the user specifies the reference object color, the corresponding "reference-color" node becomes active and specifies the color in the camera image that provides input into the reference field. A peak of activation in the reference field evolves at the location of the reference object. The reference field continuously tracks the position of the reference object. Its dynamics also filters out irrelevant inputs and camera noise and thus stabilizes the reference object representation. Having a stable, but updatable reference object representation allows the spatial semantics to be continuously aligned with the visual scene.

Spatial semantic templates

The spatial semantic templates (Fig. 1C) are represented as a set of synaptic weights that connect spatial terms to an abstract, "retinotopic" space. The particular functions defining "left", "right", "below", and "above" here were two-dimensional Gaussians in polar coordinates and are based on a neurally-inspired approach to English spatial semantic representation (O'Keefe, 2003). When viewed in Cartesian coordinates, they take on a tear-drop shape for these terms.

Shift

The shift mechanism (Fig. 1D) aligns these retinotopically defined spatial semantics with the current task space. The shift is done by convolving the "egocentric" weight matrices with the outcome of the reference field. Because the single reference object is represented as a localized activation peak in the reference field, the convolution simply centers the semantics over the reference object. The spatial terms thus become defined relative to the specified reference object location (for related method see (Pouget & Sejnowski, 1995)).

Aligned spatial semantic fields

The aligned spatial semantic fields (Fig. 1E) are arrays of dynamical neurons with weak lateral interaction. They receive input from the spatial alignment or "shift" mechanism which maps the spatial semantics onto the current scene by "shifting" the semantic representation of the spatial terms to the reference object position. The aligned spatial semantic fields integrate the spatial semantic input with the summed outcome of the color-space fields and interact reciprocally with the spatial-term nodes. Thus, a positive activation in an aligned spatial semantic field increases the activation of the associated spatial-term node and vice versa.

Demonstrations

We here detail two exemplar demonstrations (from a set of thirty conducted) which address two behaviors fundamental to spatial language. In the presented scenarios, three objects were placed in front of the robot: a green stack of blocks, a yellow plastic apple, and a blue tube of sunscreen. The visual input was formed from the camera image and sent to the reference and color-space fields. The color-space field input was formed by extracting hue value ("color") for each pixel in the image and assigning that pixel's intensity value to the corresponding location in the matching color-space field. The input for the reference field was formed in an analogous fashion according to the user-specified reference object color. When the objects are present in the camera image, the reference and color-space fields receive localized inputs, corresponding to the three objects in view (marked with arrows, see Fig. 2 and Fig. 3). This was the state of the system before the particular task was set.

In Demonstration 1 we ask "Where is the yellow object relative to the green one?" and the robot must select the correct descriptive spatial term. In Demonstration 2 we ask "Which object is to the right of the yellow one?" and the robot must select the color term that describes the target object. Both examples were performed with exactly the same visual scene and parameter set. Thus, the only difference for the system was the user-specified task input. If our model functions properly, the interactive dynamics should select the correct spatial or color term according to the task details.

Due to the graded representation of space and color in the neural fields, being able to solve these two tasks means accessing hundreds of scenarios with multiple objects and object positions in the image. More fundamentally, these different tasks both require the integration of visual and symbolic input as well as the autonomous selection of a descriptive spatial term. Such integration and decision processes are a core capacity of the human spatial language system and underlie the full range of real-world spatial language behaviors. Accounting for these core processes in different tasks in a single, neurally-grounded model provides a strong foundation for scaling up to more complex spatial language scenarios.

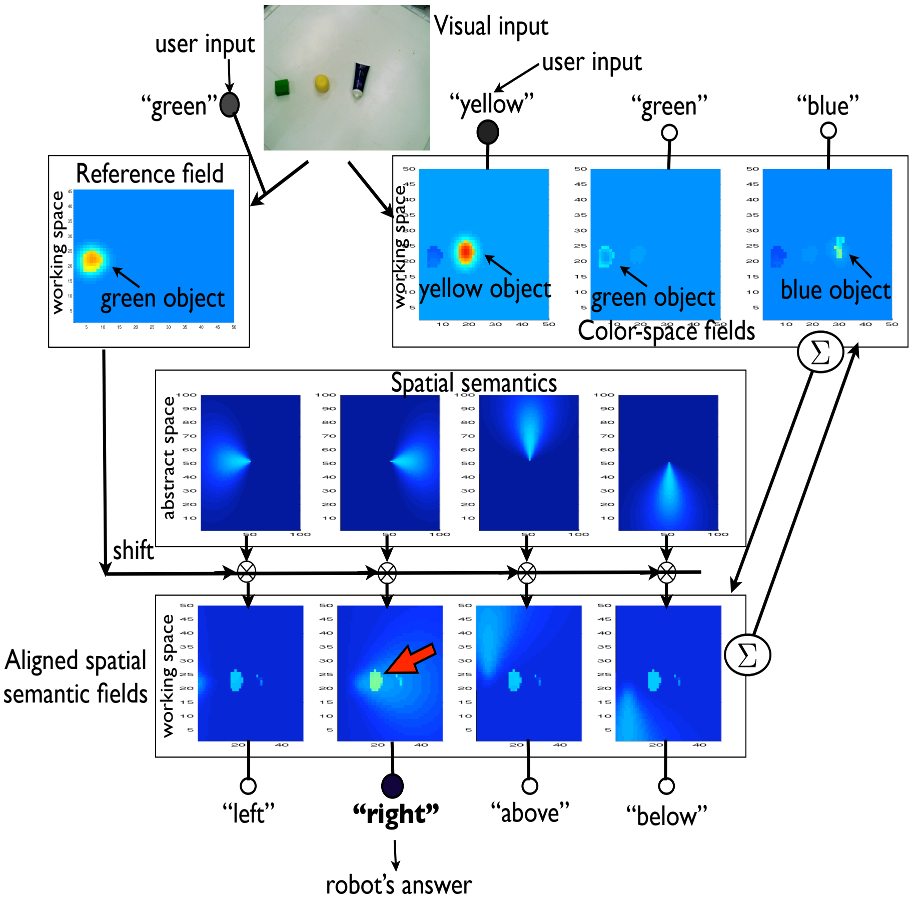


Figure 2: Demonstration 1 activations just before answering ”Where”.

Demonstration 1: Describing ”Where”

Demonstration 1 asks ”Where is the yellow object relative to the green one?” To respond correctly, the robot must select ”Right”. Fig. 2 shows the neural fields' activation just before the answer is given. The task input first activates two discrete neurons, one representing ”green” for the user-specified reference object color and the other ”yellow” for the user-specified object color (see user inputs, top Fig. 2). The reference object specification ”green” leads to the propagation of the green camera input into the reference field, creating an activation bump in the reference field at the location of the green item (see Reference field, Fig. 2). The specification of the target color ”yellow” increases the activation for the ”yellow” node linked to the ”yellow” color-space field (see yellow activation time course line, top Fig. 4a), which raises the resting level of the associated ”yellow” color-space field. This uniform activation boost coupled with the camera input from the yellow object induces an activation peak in the field (see ”yellow” Color-space field, Fig. 2).

This localized target object activation is then transfered to the aligned semantic fields. In addition to receiving this target-specific input, the aligned semantic fields also receive input from spatial term semantic nodes. Critically, these semantic profiles are shifted to align with the reference object position. In the current case, the yellow target object activation therefore overlaps with the aligned ”right” semantic field (see red arrow in the ”right” Aligned spatial semantic field, Fig. 2). This overlap ultimately drives the activation and selection of the ”right” node (see spatial-term neuron activation time course, bottom Fig. 4a).

Demonstration 2: Describing ”Which”

Demonstration 2 asks ”Which object is to the right of the yellow one?”. To respond correctly, the robot must select ”Blue”. As indicated in Fig. 3, the task input first activates two discrete nodes, one representing the reference object color ”yellow” and the other representing ”right”.

The reference object specification ”yellow” creates an activation bump in the reference field location matching that of the yellow item (see Reference field, Fig. 3). The specification of ”right”, in its turn, increases the activation for

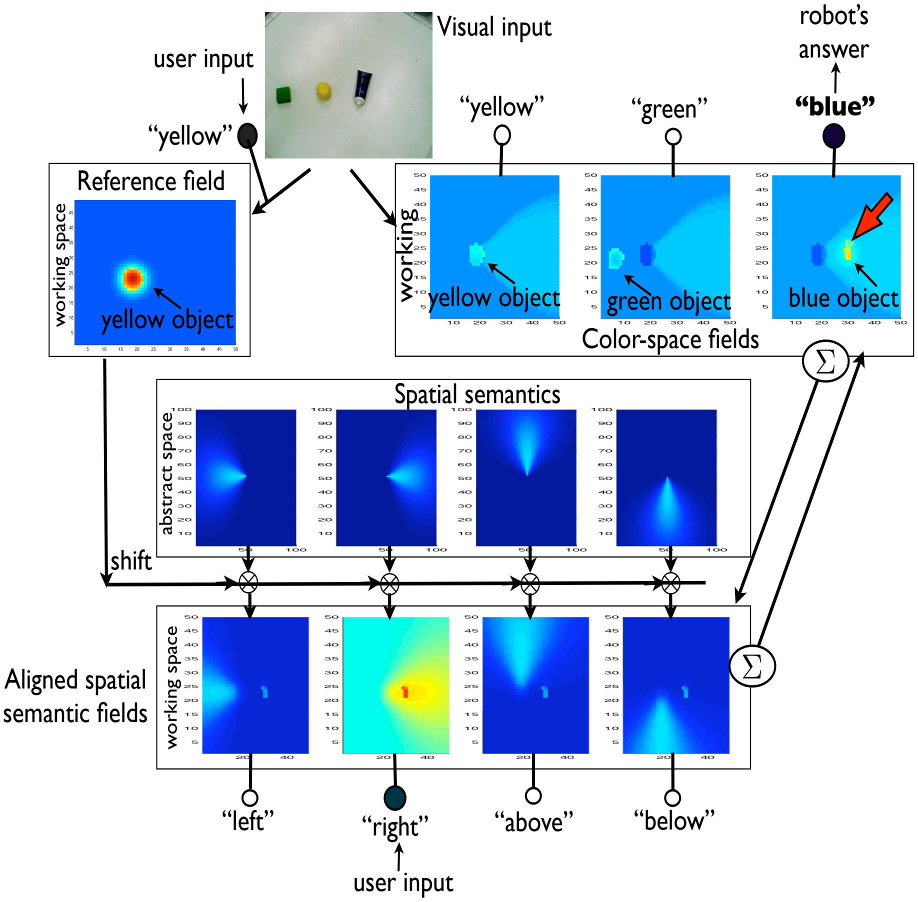
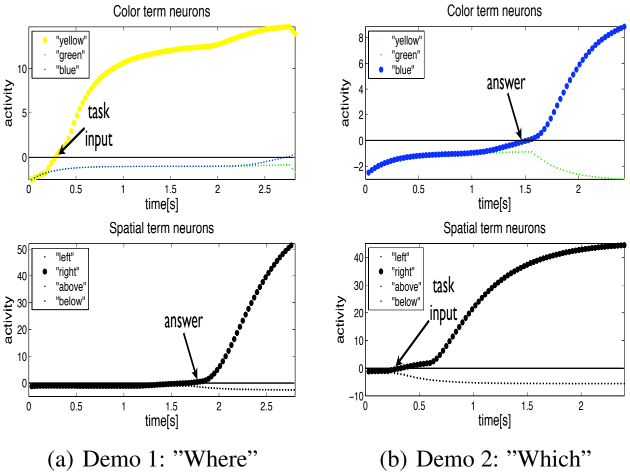


Figure 3: Demonstration 2 activations just before answering ”Which”.

that spatial-term node (see activation time course, bottom Fig. 4b), creating a homogeneous activation boost to the ”right” semantic field. This activation boost creates a positive activation in the field to the right of the yellow reference object (see ”right” Aligned spatial semantic field, Fig. 3). This spatially-specific activation is then input into the color-space fields and subsequently raises activation at all those color-space field locations to the right of the reference object (see lighter-blue Color-space fields' regions, Fig. 3). This region overlaps with the localized input of the blue object in the ”blue” color-space field and an activation peak develops in that field (see red arrow in the ”blue” Color-space field, Fig. 3). This increases the activation of the associated ”blue” color-term node, triggering selection of the correct answer, ”blue” (see color-term node's activation profile, top Fig. 4b).

Discussion

Together, these demonstrations reveal the model's ability to localize the specified target object in the visual scene and to extract the required spatial or non-spatial target infor-



(a) Demo 1: ”Where”

(b) Demo 2: ”Which”

Figure 4: Activation time courses for the spatial and color terms neurons

mation. These different behaviors emerged from the autonomous dynamics integrating the low-level camera input and the categorical user input and are thus truly context-dependent. In assessing this framework it is also important to note that precisely the same parameter setting was used in all tasks; only the context input changed. Thus, the behaviors are autonomously structured simply by the symbolic and visual input. Even with our initially limited range of spatial and color terms, the framework can be immediately applied to a broad range of real-world objects and locations without modification. This novel system therefore provides a contextually adaptive framework for the flexible application of spatial semantics. More fundamentally, because of its focus on integrative dynamic processes modelled in accordance with neural principles, it also provides a foundation for modeling more complex human spatial language behaviors.

References

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*, 77-87.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention. *Psychological Science*, *15*, 415-419.

Coventry, K., Cangelosi, A., Rajapakse, R., Bacon, A., Newstead, S., Joyce, D., et al. (2005). Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In C. Freksa (Ed.), *Spatial cognition iv* (Vol. LNAI 3343, p. 98-110). Heidelberg: Springer-Verlag.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*, 545-572.

Lipinski, J., Spencer, J. P., & Samuelson, L. (in press). It's in the eye of the beholder: Spatial language and spatial memory use the same perceptual reference frames. In L. B. Smith, M. Gasser, & K. Mix (Eds.), *The spatial foundations of language.* Oxford University Press.

O'Keefe, J. (2003). *Vector grammar, places, and the functional role of the spatial prepositions in english* (E. van der Zee & J. Slack, Eds.). Oxford: Oxford University Press.

Pouget, A., & Sejnowski, T. J. (1995). Spatial representations in the parietal cortex may use basis functions. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems 7.* MIT Pres, Cambridge MA.

Regier, T., & Carlson, L. (2001). Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, *130*(2), 273-298.

Spencer, J. P., & Schöner, G. (2003). Bridging the representational gap in the dynamical systems approach to development. *Developmental Science*, *6*, 392-412.

Spencer, J. P., Simmering, V. R., Schutte, A. R., & Schöner, G. (2007). What does theoretical neuroscience have to offer the study of behavioral development? insights from a dynamic field theory of spatial cognition. In J. Plumert & J. P. Spencer (Eds.), *The emerging spatial mind.* Oxford University Press.

Tom, A., & Denis, M. (2004). Language and spatial cognition: Comparing the roles of landmarks and street names in route instructions. *Applied Cognitive Psychology*, *18*, 1213-1230.