Glossos
Issue 7, Spring 2006
http://seelrc.org/glossos/
glossos@seelrc.org

Steven J. Clancy
University of Chicago

The Topology of Slavic Case: Semantic Maps and Multidimensional Scaling

Introduction

Language typologists and cognitive linguists have put the notion of the semantic map to great use over the last decade, particularly in such works as Haspelmath (1997) on indefinite constructions. Croft (2001, 2003) has also developed the notion of conceptual space in his typological and construction grammar work. Thus far, semantic maps have been the result of empirical research involving laborious consideration of cross-linguistic data in order to identify the relevant categories and then to arrange those categories into a conceptual space. The arrangement of these categories reflects the actual overlapping polysemy found in the data, so that connections between concepts accord with Croft's (2001, 2003) Semantic Map Connectivity Hypothesis. A conceptual space such as that found in Haspelmath (1997, 2003) emerges (Figure 1) and the indefinite constructions of all languages can subsequently be mapped onto this space (Figures 2 and 3) in accord with Croft's hypothesis. If exceptions are found as new languages are added, the semantic map can be further refined. The connections between categories in the conceptual space have validity, but the specific geometrical arrangement and distance between categories lack theoretical import. However, the recent work of Croft and Poole (forthcoming)[1] revolutionizes the semantic map and introduces a meaningful notion of quantitative semantic distance as well as a precisely defined geometric arrangement, through the use of a mathematically well-defined model, Multidimensional Scaling (MDS), more specifically utilizing Poole's *Optimal Classifcation* (OC) method. MDS techniques have long been used by researchers in psychology, economics, and political science among other disciplines. Additionally, MDS analysis allows the linguist interested in semantic maps to consider much larger conceptual spaces, where the

[1] I would like to express my gratitude to Bill Croft and Keith Poole for sharing their manuscript with me and for commentary and assistance in setting up this project. I am also grateful to Keith Poole for making

necessary permutations of possible category arrangements would be onerous if not impossible if undertaken by hand. Whereas Haspelmath's (1997, 2003) work was manageable due to the limited number of indefinite categories, MDS analysis makes possible the consideration of such topics as lexical aspect and spatial adpositions (Croft and Poole forthcoming). The study described here represents the first steps in applying MDS to the questions of case semantics with attention given to Russian, Polish, and Czech.

Conceptual Spaces and Semantic Maps: Haspelmath's study of indefinite constructions

The notions of conceptual space and the semantic map are arguably best represented in the work of Haspelmath. His cross-linguistic study of nine types of indefinite constructions in 40 languages has not only proved useful to our understanding of how these items are structured across languages, but also has provided valuable insight into the nature of conceptual space, the methodology for identifying conceptual spaces, and the explanatory and theoretical power of semantic maps. With regard to the empirical data required and the categorial manipulation involved in drawing the semantic maps, Haspelmath's work also provides a sobering statement on the difficulties, tedium, and limitations involved in the process of identifying conceptual spaces and in applying this theoretical tool to more expansive sets of data.

Haspelmath's conceptual space for indefinite constructions consists of a geometrical arrangement of the nine indefinite categories with connections between certain categories (Figure 1).

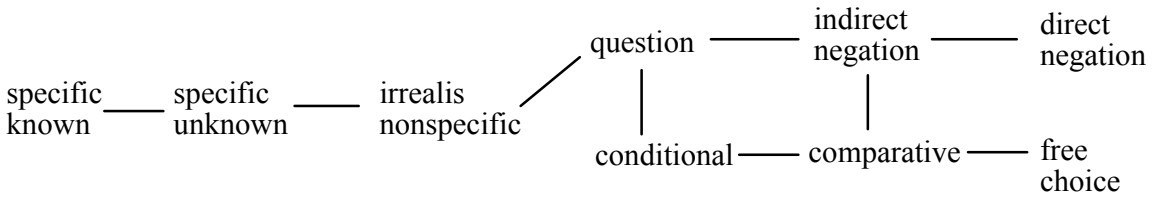

Figure 1
Haspelmath's Conceptual Space for indefinite pronouns

his Optimal Classification MDS program available, for running numerous data sets and returning many fine plots, and for walking me through the setup so that I could begin running my own data.

Haspelmath (1997), adapted by Croft & Poole (forthcoming)[2]

The nine categories he identified are *specific known, specific unknown, irrealis non-specific, question, conditional, indirect negation, comparative, free choice,* and *direct negation* as exhibited for Russian *-koe*, *-to*, *-nibud'*, *-libo*, X *by to ni byl*(*o*), X *ugodno*, *ljuboj*, and *ni-* constructions and by translation for English *some-*, *any-*, and *no-* constructions in examples (1)-(9).

(1) Specific known (to the speaker, not to the hearer)
*Maša vstretilas' **koe** s **kem** okolo universiteta.*
*Masha met with **someone/somebody** near the university.*
(Haspelmath 1997:46)

(2) Specific unknown (to neither the speaker nor the hearer)
*Maša vstretilas' s **kem-to** okolo universiteta.*
*Masha met with **someone/somebody** near the university.*
(Haspelmath 1997:46)

(3) Irrealis non-specific
*Kupi mne **kakuju-nibud'** gazetu.*
*Buy me **some** newspaper.* (Haspelmath 1997:42)

(4) Question (polar question)
*Zvonil li mne **kto-nibud'/kto-libo**?*
*Did **anyone** call me?* (Haspelmath 1997:274)

(5) Conditional (protasis)
*Esli **čto-nibud'/čto-libo** slučitsja, ja skažu mame.*
*If **anything** happens, I'll tell mom.* (Haspelmath 1997:274)

(6) Indirect negation
*bez **kakoj-libo/kakoj by to ni bylo** pomošči*
*without **any** help* (Haspelmath 1997:33)

(7) Comparative
*Zdes' prijatnee žit' čem **gde-libo/gde by to ni bylo** v mire.*
*It is more pleasant to live here than **anywhere** in the world.*
(Haspelmath 1997:35)

[2] The conceptual space identified in Haspelmath (1997) has been adapted slightly by Haspelmath (2003) as well as by Croft and Poole (forthcoming). The version used here is from Croft and Poole (forthcoming), where the direct link between *irrealis nonspecific* and *conditional* (Haspelmath 1997, 2003) has been eliminated in favor of a link between *irrealis nonspecific* and *conditional* through the *question* node.

(8) Free choice
Ty možeš kupit' ***ljubuju/kakuju ugodno*** *knigu.*
You can buy ***any*** *book.* (Haspelmath 1997:274)

(9) Direct negation
My ***ničego*** *ne znaem.*
We don't know ***anything****./We know* ***nothing****.*

By studying the overlapping functions across languages, Haspelmath was able to identify the connections between concepts and then arrange them into a conceptual space (Figure 1), such that the indefinite constructions of individual languages could map onto the geometrical arrangement as shown here for Russian (Figure 2).

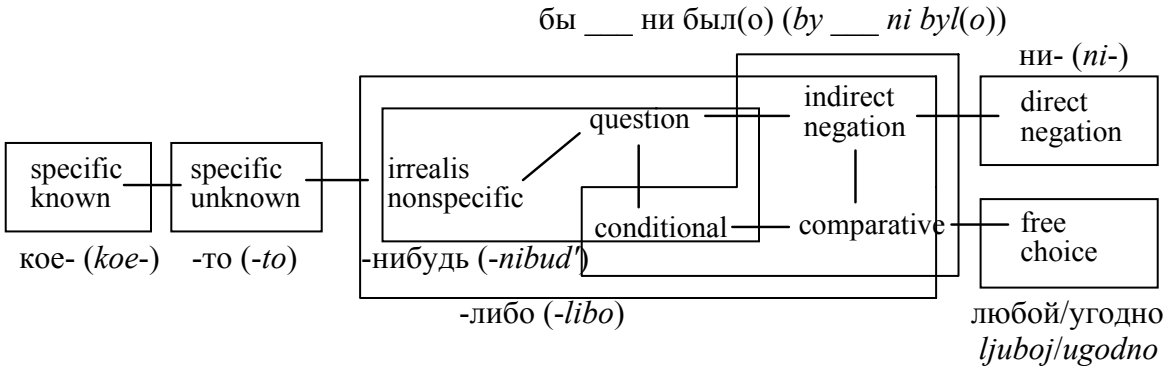

Figure 2
Russian indefinites adapted from Haspelmath (1997:273, 2003:223)

The conceptual space of Figure 1 becomes a semantic map for an individual language once the available constructions in that language are mapped onto it as exemplified in Figure 2. A single boxed area may encompass one or more nodes in the conceptual space such that there are no discontinuities, e.g. a lexical item for *conditional* cannot also express *irrealis nonspecific* without also encompassing *question* along the way. If such connections between functions across discontinuous areas of the conceptual space were found as new languages were added to Haspelmath's data set, he would have adjusted the connections and the arrangement of his conceptual space accordingly. The conceptual space presented in Haspelmath's work provides for the connections found in the data in a geometrical arrangement that makes it possible to draw the semantic maps for the languages in the study. The Semantic Map Connectivity Hypothesis (Croft 2001, 2003) proposes that we should find no discontinuities.

> *Semantic Map Connectivity Hypothesis*: any relevant language-specific and construction-specific category should map onto a connected region in conceptual space.
>
> (Croft 2001:96)

Taken together, these ideas provide powerful tools for cross-linguistic comparison and for identifying language universals.

Multidimensional Scaling: General applications and Poole's study of legislatures

Haspelmath's (1997) foray into conceptual space and the coherent connections between the functions therein as well as the explanatory power of the language-specific semantic maps suggests no end of fascinating studies for future linguistic research. However, one must soon face up to the labor intensive process of carrying out the permutations of possible arrangements of the identified functions in order to reveal a given conceptual space, effectively putting a limit on the types of linguistic problems that can be subjected to semantic map analysis. However, Croft and Poole (forthcoming) identify a powerful mathematical tool for identifying conceptual spaces that is surely as ground-breaking as Haspelmath's research over the past decade. Multidimensional Scaling (MDS) is just the tool we need to allow us to consider the conceptual space of much larger linguistic regions. MDS has long been used in the social sciences, particularly in psychology, economics, and political science, e.g. Poole's studies of parliamentary voting patterns[3]. Croft and Poole (forthcoming) presents an initial example of MDS with driving distances between US cities. The tables of cities and driving distances we commonly encounter in road atlases provide a set of cities that differ from each other in how many miles apart they are, but the tables provide no information about the specific locations of those cities. However, an MDS analysis of that data is able to form a 2 dimensional space from that data that we easily recognize as an approximate map of the United States in its north-south-east-west dimensions. In one problem dealing with a symmetric matrix, we have data on a group of cities, in which we know the distances between every pair of cities in the matrix. An MDS analysis based on similarity (small distances) and dissimilarity (long distances) produces a spatial map. However, MDS analysis is also useful even when we do not have a symmetric matrix. An unfolding analysis of a table of driving distances, in

[3] See Croft and Poole (forthcoming) for a brief, yet thorough, introduction to MDS and various applications.

which we have one group of cities running vertically and another group of cities running horizontally, with distances specified between, but not within, the two groups, also yields a rough spatial map of the United States. This revelation of a spatial pattern within a body of data for driving distances is intriguing, but the usefulness of MDS analysis only begins with such examples. Poole's work on parliamentary voting patterns analyzes large bodies of roll call votes in legislatures. In these studies it is possible to map individual legislators as points in a multidimensional space based on their voting patterns in a series of yes/no roll call votes. For each roll call, the analysis proceeds to arrange all the legislators voting for the measure on one side of a cutting point (one dimension), a cutting line (two dimensions), or a cutting plane (three dimensions) such that all of those voting for the measure are on one side and all voting against are on the other side. With a higher number of votes and some diversity of positions among the legislators, an ideal point emerges for each legislator, such that, in two-dimensions, for instance, a cutting line may be drawn for any vote and that legislator is placed in the appropriate position for any outcome.[4] So as not to lose sight of our goal, we can here consider the smaller voting body of the US Supreme Court as shown in Figure 3.

As opposed to the problem of driving distances where the X-Y-axes were readily identified as compass points on a physical map, a map of legislators requires further analysis. In this map of Supreme Court justices, we can see a liberal-conservative dimension running from left to right on the X-axis corresponding to conventional wisdom and journalistic writing about the Supreme Court. This spatial arrangement of justices allows us to quantify certain suspected relationships between justices (e.g., an ideological affinity between Justices Scalia and Thomas) or general traits about justices (e.g., O'Connor was a "swing vote").

However, we are still left scratching our heads as to the nature of the Y-axis. Poole suggests that there is basically one dimension in this data, with Breyer and O'Connor the worst-fitting justices in a one dimensional liberal-conservative model.

[4] "At the heart of OC are two algorithms -- *the cutting plane procedure* and *the legislative procedure*. Both of these procedures are unique and stable. In particular, Monte-Carlo tests show that when the number of legislators is 100 or greater and the number of roll calls is on the order of 500 – typical of national legislatures like the U.S. Senate – then the recovery of the legislators and cutting lines/planes in one to ten dimensions at high levels of error and missing data is very precise. Even with very small data sets OC

O'Connor must be further "north" than her fellow justices so that the cutting lines will sometimes include her with one side of a vote on a particular issue and sometimes with the other side on another issue. In effect the Y-axis is a measure of being a swing vote. However, the spatial maps from Poole's study of the US House and Senate reveal a liberal-conservative X-axis and a meaningful Y-axis representing social issues. From such studies, one can see the usefulness of MDS for eliciting the spatial structure, but the necessity of further interpretation once the mathematics has done its work.

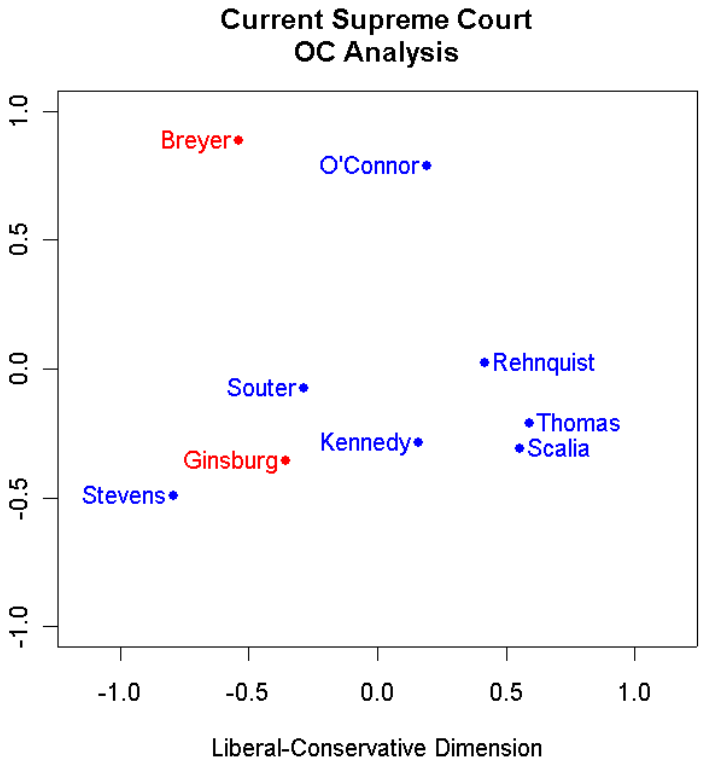

Figure 3
Spatial Map of the Supreme Court[5]

Multidimensional Scaling: points, polytopes and cutting lines

Using Poole's Optimal Classification nonparametric unfolding algorithm, Croft and Poole (forthcoming) present a number of applications of MDS to linguistic issues. At this point, one may be wondering what is multidimensional about MDS? An MDS

produces reliable results. It is a stable building block upon which more complex parametric scaling methods can be constructed" (Poole 2005:46; see Poole 2005:46 for further references.)

analysis may use one, two, three, or any number more dimensions in the analysis, but one rarely finds that more than two dimensions is necessary for the type of linguistic applications considered here. A lower number of dimensions is also of greater utility in the production of a spatial representation that reveals something meaningful to us about the data in question. Some linguistic problems are one-dimensional as have been identified in hierarchies or rankings of grammatical relations such as the hierarchy of relative clauses presented in Keenan and Comrie (1977) and discussed in Croft and Poole (forthcoming). Unruly data in a one-dimensional model can usually find a better fit within a two-dimensional model, but further increases in dimensionality result in only marginal increases in fit (see below for further discussion of Correct Classification and APRE, the two measures of fit provided in the OC analysis).

Using Haspelmath's data on 40 languages, Croft and Poole replicated Haspelmath's conceptual space (Figure 1) through use of MDS.

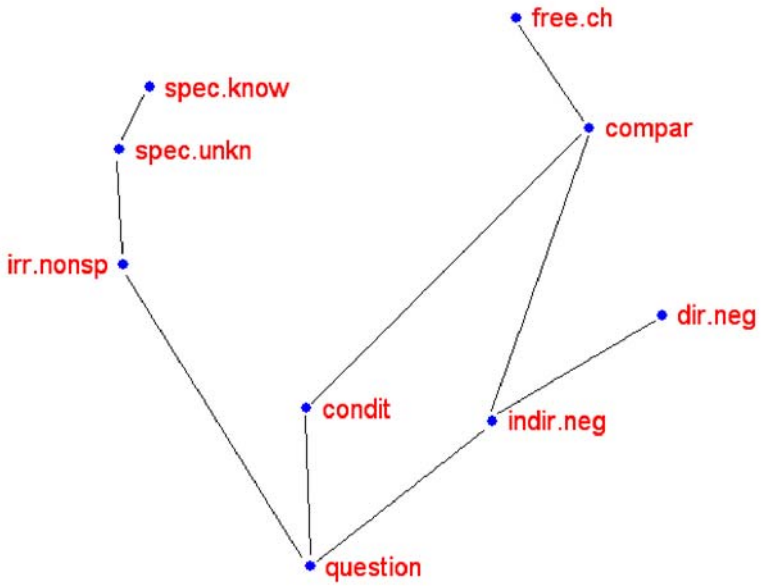

Figure 4
MDS analysis of Haspelmath's Data in Croft and Poole (forthcoming)

What was a basically linear conceptual space in Haspelmath (1997) is now a horseshoe shaped, curvilinear conceptual space. The procedure of using straight cutting lines results in this artifact in the spatial map from the MDS analysis. The MDS plot of the indefinite

[5] http://voteview.com/images/current_supreme_court_oc.gif

construction data provides a clear picture of the roles of points, polytopes, and cutting lines in defining the conceptual space. The ideal points should not be interpreted as precise locations for the various functions. Rather, these points are located in a bounded (or open) space called a polytope. The polytope is the bounded space formed around the ideal point by the various cutting lines. The location of the point for any given function could actually be plotted anywhere within that polytope.

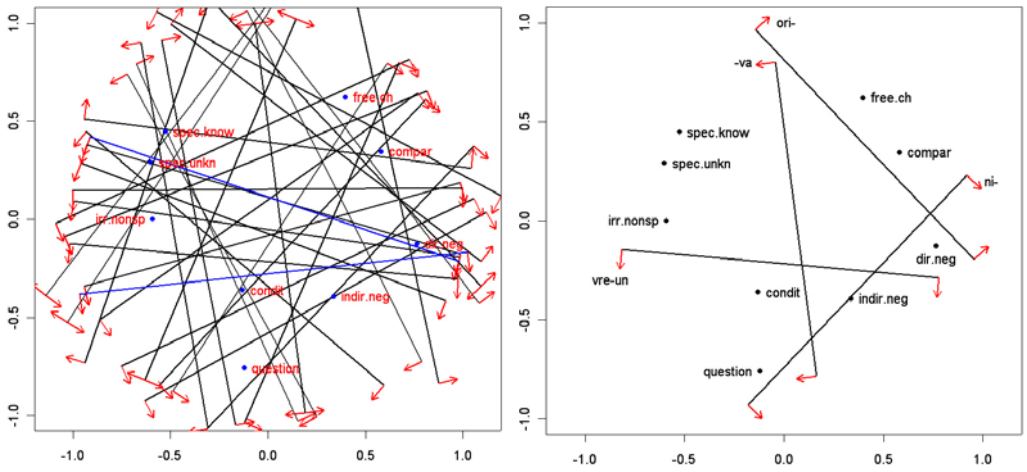

Figure 5a
Cutting lines for full 40 language set

Figure 5b
Cutting lines for Romanian only
(Croft and Poole forthcoming)

When we look at the plot of the 40 language set in Figure 5a, we see that many of the functions are limited to a very small polytope (*specific known*, *specific unknown*, *conditional*, and *direct negation*), others to a relatively well-bounded polytope (*free choice*, *comparative*), others to a slightly larger, but still well defined polytope (*irrealis*, *question*), and only one to a large, unbounded polytope (*indirect negation*). When we look at the cutting lines for one language only, such as Romanian in Figure 5b, we see a language-specific semantic map drawn onto the conceptual space defined from the 40 language set. The arrows on the cutting lines indicate the side of the cutting line that includes the functions in that roll call element. We further see that if we only included one language, the polytopes formed by the cutting lines for the four relevant constructions would be much larger and less well defined. The strength of the MDS analysis lies in revealing language universals in a large, diverse body of data, but just

such a body of data is required as well. The inclusion of additional, unrelated languages reduces the size of the polytopes and adds a greater degree of precision to the conceptual space shown in Figure 5a. Haspelmath (2002:217) indicates that it is generally sufficient to examine "a dozen genealogically diverse languages to arrive at a stable map that does not undergo significant changes as more languages are considered". These strengths and weaknesses of the MDS method should be kept in mind when considering the analysis of the Slavic case constructions discussed in this paper. As a pilot study, I am only considering data from three closely related languages, yet I am looking at areas where the languages in question show some measure of diversity.

Recall that in Haspelmath's conceptual space, only the connections between functions had theoretical import. The distance between functions was not significant and the specific geometrical arrangement could have been different as long as it remained possible to draw boxes around connected functions in the language-specific semantic maps. However, Croft and Poole have introduced a tool for creating semantic maps in which the distances between functions is quantified and in which the geometric arrangement, artifacts of the analysis aside, are also significant. We would expect points located closely together in the conceptual space to be more frequently encompassed by a single lexical item or morphological construction. We would also expect semantic development to proceed from one node to a closely related node as specific languages change over time. Equipped with this powerful cartographic tool, let us now turn our attention to the issue of case semantics in Slavic.

Conceptual Spaces and Semantic Maps: Applications to Slavic Case Semantics

In this paper, I consider two pilot studies in the use of Poole's Optimal Classification nonparametric unfolding algorithm to analyze the semantics of the Slavic case systems. As a sample conceptual region, let us consider the prepositions and case uses involved with the expression of DESTINATION, LOCATION, and SOURCE in Russian, Czech, and Polish. Figure 6 shows what I typically present to students of Slavic languages when I teach these topics in language courses. There is a certain logic to laying out these constructions in terms of going to a DESTINATION, being in a LOCATION, and coming from a SOURCE location. There is also a certain logic to considering the three

classes involved: the majority class (so called *v*-words), the minority class (so called *na*-words), and the human class. For Russian, one can see how the accusative case is associated with DESTINATION, the locative case is associated with LOCATION, and the genitive case is strongly associated with SOURCE. Furthermore, one can see the connection between R *v* 'to, in' in the accusative and locative cases and R *na* 'to, on' in those cases. On the other hand, one can see certain frustrating tendencies as well, such as in Czech where the three classes of DESTINATION are marked by not only three different prepositions, but by three different cases, or in Polish, where a single case, the genitive, is associated with DESTINATION, LOCATION, and SOURCE.

Russian	DESTINATION	LOCATION	SOURCE
majority	*v* + ACC	*v* + LOC	*iz* + GEN
na-words	*na* + ACC	*na* + LOC	*s* + GEN
human	*k* + DAT	*u* + GEN	*ot* + GEN

Czech	DESTINATION	LOCATION	SOURCE
majority	*do* + GEN	*v* + LOC	*z* + GEN
na-words	*na* + ACC	*na* + LOC	
human	*k* + DAT	*u* + GEN	*od* + GEN

Polish	DESTINATION	LOCATION	SOURCE
majority	*do* + GEN	*w* + LOC	*z* + GEN
na-words	*na* + ACC	*na* + LOC	
human	*do* + GEN	*u* + GEN	*od* + GEN

Figure 6
Charts for DESTINATION-LOCATION-SOURCE Constructions

When considered as conceptual spaces and semantic maps, these charts in Figure 6 fail to conform to Croft's Semantic Map Connectivity Hypothesis for Czech and Polish and must be subjected to further rearrangement. Figure 7 emerges as a possible arrangement of the conceptual space, one in which the individual cases form contiguous regions, but in which the overlapping use of a single preposition, e.g., R *v* or R *na*, is no longer contiguous. Further rearrangement is necessary.

Russian	DESTINATION	SOURCE	LOCATION
majority	*v* + ACC	*iz* + GEN	*v* + LOC
na-words	*na* + ACC	*s* + GEN	*na* + LOC
human	*k* + DAT	*od* + GEN	*u* + GEN

Polish	DESTINATION	SOURCE	LOCATION
majority	*do* + GEN	*z* + GEN	*w* + LOC
na-words	*na* + ACC	*z* + GEN	*na* + LOC
human	*do* + GEN	*od* + GEN	*u* + GEN

Czech	DESTINATION	SOURCE	LOCATION
majority	*do* + GEN	*z* + GEN	*v* + LOC
na-words	*na* + ACC	*z* + GEN	*na* + LOC
human	*k* + DAT	*od* + GEN	*u* + GEN

Figure 7
Charts for DESTINATION-LOCATION-SOURCE Constructions
as semantic maps

Again, we see how even a small set of functions may cause problems for the achievement of a stable conceptual space. Fortunately, we may set our permutations aside and subject the data to an MDS analysis in hopes of settling the issue.

Multidimensional Scaling: Poole's Optimal Classification method

Poole makes much of his material and software available at his voteview.com website. The program for performing the Optimal Classification algorithm may be found there and runs with data in plain text files on a Windows computer (see below for more information). Poole's software is set up to deal with legislative roll call votes, but it may just as fruitfully be applied to linguistic problems and the legislative metaphor is useful in considering how to structure the database. The functions/meanings/constructions are considered as "legislators" and the roll calls are lexical items, case endings, etc. in specific languages. Just as the number of senators in the US Senate is finite, the set of functions considered may be defined at the outset of a particular study, but once consituted, those "legislators" can then go on to "vote" on the data from innumerable languages with additional languages adding to the diversity, and thus, specificity, of the resulting conceptual space.

When faced with a new roll call, say the introduction of a preposition, we may enter a "vote" of yes/1 or no/0 for each function, depending on whether or not that function is expressed by the preposition in question.[6] However, the question remains, how ought we to code the data in question? To answer this question, I considered three coding schemes for the DESTINATION-LOCATION-SOURCE constructions: overspecified, underspecified, and correctly specified. Table 1 shows an example for dealing with R *na*+ACC and R *na*+LOC. Coding **destination (*na*-words)** in the database as a "1" for **R *na*** associates this function with all uses of R *na*, including both accusative and locative uses. Coding as **Russ ACC** associates this function with all other uses of the accusative. Coding as **R *na*+ACC** associates this function only with other instances of R *na*+ACC.

Construction/Function:	destination (*na*-words) R *na*+ACC	location (*na*-words): R *na*+LOC
Overspecified	R *na* Russ ACC R *na*+ACC	R *na* Russ LOC R *na*+LOC
Underspecified	R *na*+ACC	R *na*+LOC
Correctly specified	R *na* Russ ACC	R *na* Russ LOC

Table 1
Coding Schemes for DESTINATION-LOCATION-SOURCE

The underspecified coding model fails to establish a connection between R *na* in its dual case governance and also fails to connect R *na* with either the accusative or locative cases in Russian. In effect, we have coded two prepositions R na_1 and R na_2. The overspecified coding model captures the identity of a single preposition R *na* associated with both accusative and locative cases and also associates this function with the full set of accusative constructions in Russian, but the additional coding as R *na*+ACC or R *na*+LOC also effectively creates a R na_2 and R na_3, each of which is only associated with one case and neither of which is associated with the full range of uses of those cases in Russian. However, it may be that either the overspecified or underspecified coding schemes would elicit patterns leading to a viable semantic map if the overall data sample

[6] Data is actually encoded as 1 for Yes and 6 for No in Poole's optimal classification application PERFL.EXE, but the 1/0 principle still stands.

were larger, both in terms of functions considered and greater linguistic diversity entered into the database.

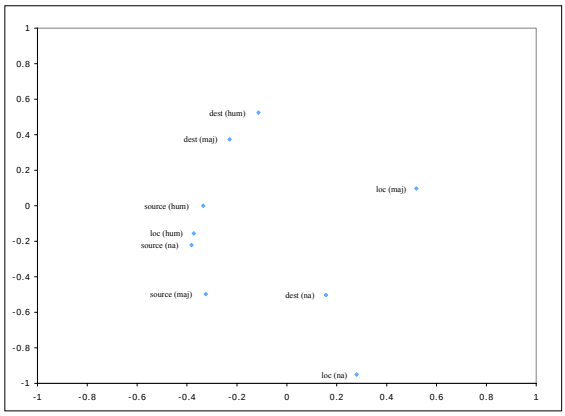

Figure 8
DESTINATION—LOCATION—SOURCE
Overspecified Case Marking

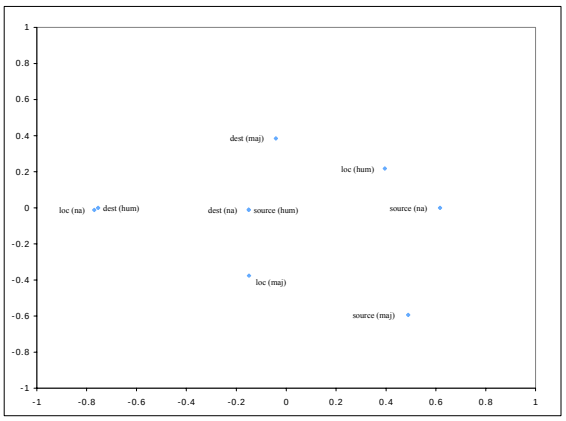

Figure 9
DESTINATION—LOCATION—SOURCE
Underspecified Case Marking

The MDS analysis of DESTINATION-LOCATION-SOURCE constructions uses data from Russian, Polish and Literary as well as Colloquial Czech[7]. The nine functions considered are **destination (majority), destination (human), destination (na-words subset), source (majority), source (human), source (na-words subset), location (majority), location (human), location (na-words subset)**. The resulting conceptual space for the Overspecified and Underspecified models are presented in Figures 8 and 9. At this point, the computational phase has played its role and the interpretation of the linguist must enter the picture. The Underspecified model can be rejected based on the clustering of obviously unrelated functions such as **location (na-words subset)** — exclusively locative case — grouped with **destination (human)** — associated with either the dative or genitive cases. We also see a single set of coordinates for **destination (na-words subset)** — all *na*+ACC — and **source (human)** — all *od*/*ot*+GEN — precisely because these two clusters have uniform results in the data set, not because of any affinity between DESTINATION and SOURCE in this instance. The Underspecified model simply fails to make meaningful connections within the data set. The Overspecified model is more difficult to dismiss, but we can here turn to two additional indicators in order to assess the

[7] There are actually no differences between Literary Czech and Colloquial Czech for this set of constructions, but this data was included because it was all part of a larger data set used below in the analysis of 46 case constructions.

quality of the MDS analysis: Correct Classification and Aggregate Proportional Reduction in Error (APRE). Both of these numbers should be reported for any spatial map (Poole 2005:129). Correct Classification is a measure of fit for the ideal points found for each "legislator" such that these points are located on the appropriate side of the cutting line for each roll call in the data. The MDS analysis works to adjust the ideal points and cutting lines in order to maximize the correct classification. APRE, defined in Figure 10, is a measure of how well the minority position is accounted for in the model. This is especially important in linguistic applications where the "vote", i.e., the participation of a given lexical item or case for each function considered, may be quite lopsided. For instance, the roll call for Cz *k*+DAT only applies to the **destination (human)** function, so we are left with a vote of 8 against, 1 for. For this reason, it is also useful to consider the Average Majority Margin for a data set.

$$\text{APRE} = \frac{(\text{total number of choices cast on minority side of all roll calls} - \text{total classification error})}{(\text{total number of choices cast on minority side of all roll calls})}$$

Figure 10
Aggregate Proportional Reduction in Error (APRE)

APRE ranges from 0 to 1. An APRE of 0 means the analysis is no better than in a random spatial map and an APRE of 1 means a perfect fit for the data (Poole 2005:129).

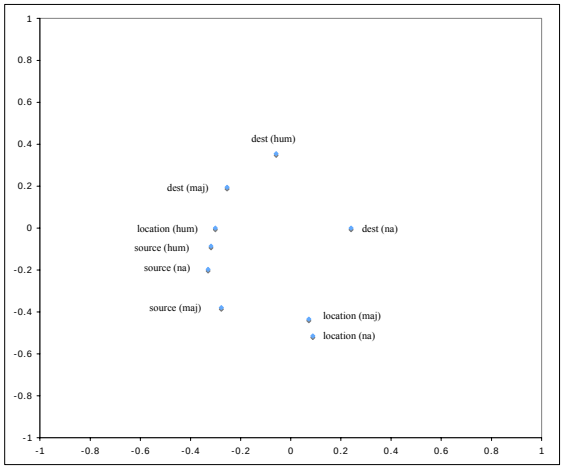

Figure 11
DESTINATION—LOCATION—SOURCE
Correctly Specified Case Marking

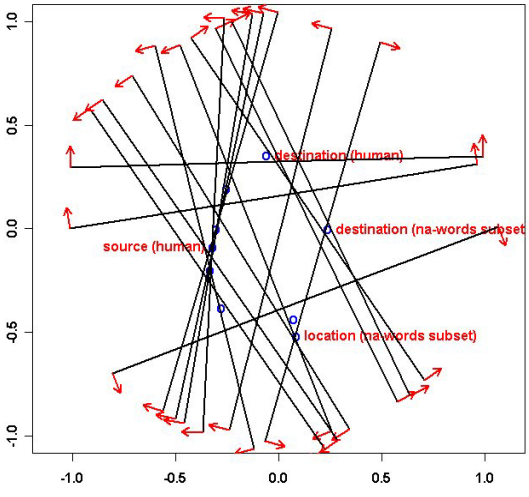

Figure 12
DESTINATION—LOCATION—SOURCE
Correctly Specified with Cutting Lines

The "Correctly specified" coding scheme (Figure 11) produces the best visual clustering of functions for the present set of data in the DESTINATION-LOCATION-SOURCE map as well as better results for the two measures of fit in the analysis. Figure 12 shows the cutting lines and polytopes for the nine functions. For such a small data set, the functions are remarkably small with only **location (na-words subset)** appearing in an open polytope. Table 2 compares the Correct Classification and APRE results for the three coding schemes used here. For the "Correctly Specified" scheme, I have also considered one vs. two dimensions.

	Correct Classification (1-D) 2-D	APRE (1-D) 2-D	Average Majority Margin
Underspecified	97.2%	0.778	87.5%
Overspecified	98.2%	0.884	84.5%
Correctly Specified	(91.3%) 98.4%	(0.507) 0.910	82.3%

Table 2
DESTINATION-LOCATION-SOURCE
Correct Classification and APRE Results

The increase in dimensionality from one dimension to two yields a significant increase in Correct Classification and APRE scores. These two fit indicators can aid one in deciding the appropriate number of dimensions for a given data set. Once the correct number of dimensions has been reached, the improvements in Correct Classification and APRE will increase only marginally. In general, a lower number of dimensions provides a better image of the structure of the data (Croft and Poole forthcoming).

The goal of establishing the conceptual space for a given data set is to reveal a universal space onto which the semantic maps for specific languages may be drawn. The DESTINATION-LOCATION-SOURCE data patterns nicely for establishing locative, genitive, accusative, and dative zones. Figures 13-15 show the semantic maps for Russian, Czech, and Polish. The conceptual space from the MDS analysis shows nice clustering effects for locative, genitive, and the single dative function. The accusative space is somewhat spread out, but the **destination (majority)** point has to account for overlap between genitive and accusative cases. Although prepositions are not shown with connections in these plots, we could also consider the semantic space of prepositions and of the notions of DESTINATION, LOCATION, and SOURCE. In the semantic maps, we could connect the

Russian, Czech, and Polish *na*+ACC and *na*+LOC functions and for Russian the *v*+ACC and *v*+LOC functions. For Polish, we could connect the P *do*+GEN functions and for Polish and Czech the *z*+GEN functions.

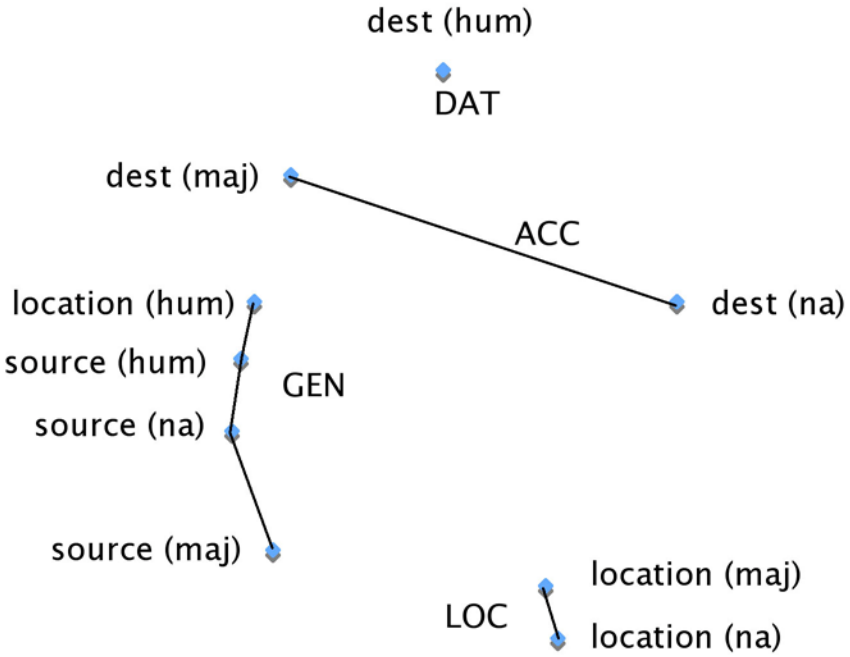

Figure 13
DESTINATION—LOCATION—SOURCE
Russian

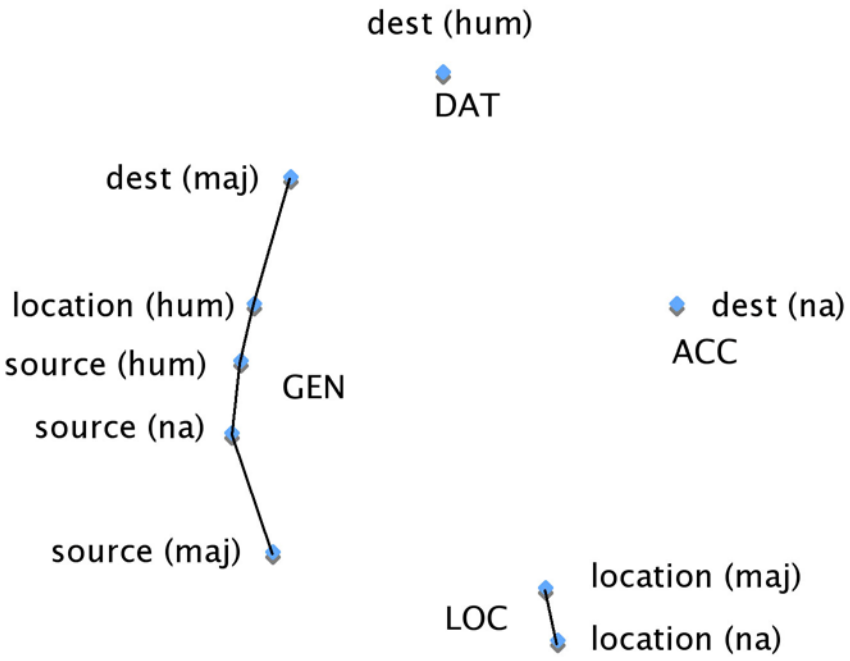

Figure 14
DESTINATION—LOCATION—SOURCE
Czech

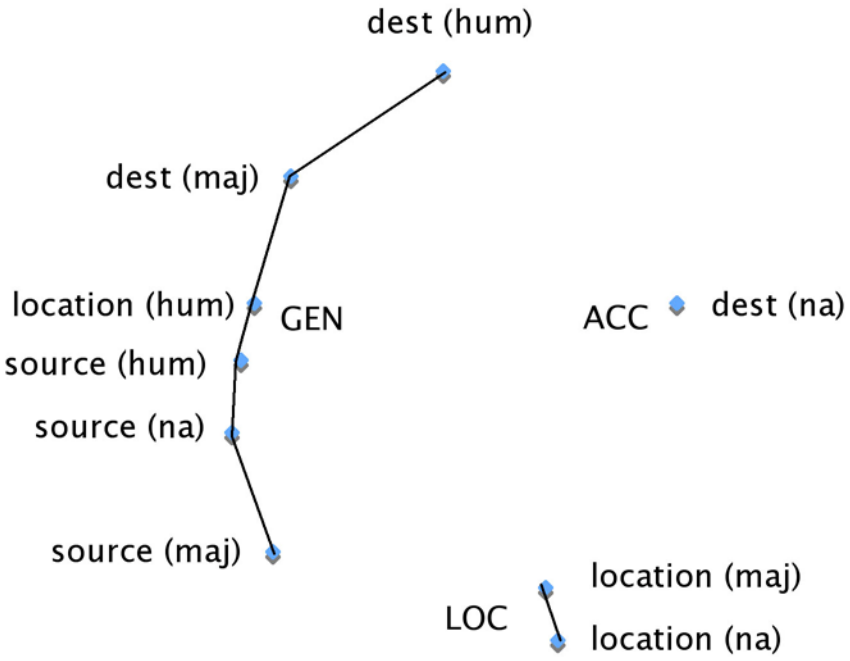

Figure 15
DESTINATION—LOCATION—SOURCE
Polish

Multidimensional Scaling: Using Poole's Optimal Classification software

One application file and two properly formatted plain text files are required to use Poole's Optimal Classification program. PERFL.EXE is a compiled Fortran program developed by Poole to perform the Optimal Classification MDS analysis on Windows computers. PERFSTRT.DAT is a control card file containing the information the program needs to analyze the data such as the name of the data file to analyze ((*title)*.ORD), the title for the data, the number of dimensions, how many columns of data (=number of roll call votes, i.e., the number of case and preposition categories) are being analyzed, and some other information the program needs, most of which does not change from data set to data set. A sample PERFSTRT.DAT file is shown in Figure 16.

```
MOT002B.ORD[1]
NON-PARAMETRIC MULTIDIMENSIONAL UNFOLDING OF SLAVCASE SEMANTICS[2]
     2[3]   42[4]    20    30[5]    01    01    01 0.005
(30[5]A1,3900I1)
(I5,1X,30[5]A1,2I5,50F8.3)
```

Figure 16
Sample PERFSTRT.DAT File[8]

[1]The name of the file containing the data.
[2]A title for the analysis.
[3]The number of dimensions in the analysis. Adjust accordingly.
[4]The number of roll calls. Adjust according to your data sample.
[5]The number of characters for function names. Can be adjusted up to 99 characters.

The .ORD is a plain text file containing only the rows of data with function names no longer than the number of characters specified in PERFSTRT.DAT followed by single digit columns of roll call data. Figure 17 presents a sample .ORD file processed from a more informative, reader-friendly spreadsheet database, a sample of which is shown in Table 3 (see note 6 above for an explanation of 1/yes and 6/no used in the database and .ORD file.). Whenever PERFL.EXE is run, it produces three output files:

PERF21.DAT	contains the Correct Classification and APRE values
PERF23.DAT	contains the Average Majority Margin and is used for diagnostics and debugging
PERF25.DAT	contains the ideal point coordinates and other useful figures

[8] See http://pooleandrosenthal.com/Optimal_Classification.htm for more details on the control card and data files.

```
destination (majority)         16666666111666616666666666666666661116666666
destination (human)            66661116166666666666666666666666661661116666
destination (na-words subset)  11116666666666666661111666666666666666666666
source (majority)              66666661111666666666666611116666666666666666
source (human)                 66666661111666666666666666666666666666661111
source (na-words subset)       66666661111666666666661611166666666666666666
location (majority)            66666666666111111116666666666666666666666666
location (human)               66666661111666666666666666661111666666666666
location (na-words subset)     66666666666111166661111666666666666666666666
```

Figure 17
Sample .ORD File

Expression, Construction, etc.	RUSSacc	PLSHacc	LCZacc	CCZacc	RUSSdat	LCZdat	CCZdat	RUSSgen	PLSHgen	LCZgen	CCZgen	⋮
destination (majority)	1	6	6	6	6	6	6	6	1	1	1	...
destination (human)	6	6	6	6	1	1	1	6	1	6	6	⋯
destination (na-words subset)	1	1	1	1	6	6	6	6	6	6	6	⋯
source (majority)	6	6	6	6	6	6	6	1	1	1	1	⋯
source (human)	6	6	6	6	6	6	6	1	1	1	1	⋯
source (na-words subset)	6	6	6	6	6	6	6	1	1	1	1	⋯
location (majority)	6	6	6	6	6	6	6	6	6	6	6	⋯
location (human)	6	6	6	6	6	6	6	1	1	1	1	⋯
location (na-words subset)	6	6	6	6	6	6	6	6	6	6	6	⋯

Table 3
Portion of sample database

Construction	Classi-fication Errors	Total # of Choices	Proportion Correctly Classified	Maximum Distance to Polytope	X-axis	Y-axis
destination (majority)	2	42	0.952	0.082	-0.254	0.194
destination (human)	0	42	1	0.184	-0.059	0.355
destination (na-words subset)	0	42	1	0.255	0.241	0
source (majority)	0	42	1	0.117	-0.277	-0.379
source (human)	0	42	1	0.031	-0.318	-0.086
source (na-words subset)	0	42	1	0.026	-0.33	-0.197
location (majority)	4	42	0.905	0.096	0.073	-0.434
location (human)	0	42	1	0.041	-0.301	0
location (na-words subset)	0	42	1	0.501	0.088	-0.515
Correct Classification	98.4%					
APRE	0.910					
Average Majority Margin	82.3%					

Table 4
DESTINATION-LOCATION-SOURCE
Correctly Specified
X-Y Coordinates and other information

The relevant information from the three output files can be compiled into a spreadsheet and used to plot the points and consider the measure of dimensionality and fit. Under "Maximum Distance to Polytope", 0.501 is the default maximum distance to a polytope indicating an open polytope, here shown for **location (na-words subset)**. The additional software components necessary to produce the cutting line plots has not yet been made available and I am grateful to Poole for producing the cutting line plots included in this paper.

Multidimensional Scaling: Towards a conceptual space of Slavic case semantics

Based on the success of the DESTINATION-LOCATION-SOURCE set of data, I expanded the data set to begin to account for the entire case system in Russian, Polish, Literary Czech, and Colloquial Czech. For an initial pilot project, I chose a small sample of 46 functions and constructions (Table 5) to generate a sample data set including areas where there is uniform agreement across these languages as well as areas of diversity. The data was entered according to the Correctly Specified coding scheme discussed above.

Functions/Constructions	Russian	Polish	Lit Czech	Coll Czech
give X sth.	DAT	DAT	DAT	DAT
help X	DAT	DAT	DAT	DAT
before, in front of X	INST	INST	INST	INST
without X	*bez* GEN	*bez* GEN	*bez* GEN	*bez* GEN
on X (day of week)	*v* ACC	*w* ACC	*v* ACC	*v* ACC
time (duration)	ACC	ACC	ACC	ACC
time (for an amount)	*na* ACC	*na* ACC	*na* ACC	*na* ACC
source (majority)	*iz* GEN	*z* GEN	*z* GEN	*z* GEN
source (na-words subset)	*s* GEN	*z* GEN	*z* GEN	*z* GEN
source (human)	*ot* GEN	*od* GEN	*od* GEN	*od* GEN
location (majority)	*v* LOC	*w* LOC	*v* LOC	*v* LOC
location (na-words subset)	*na* LOC	*na* LOC	*na* LOC	*na* LOC
location (human)	*u* GEN	*u* GEN	*u* GEN	*u* GEN
on a date (calendar)	GEN	GEN	GEN	GEN
control, govern X	INST	INST, *nad* INST	DAT	DAT
destination (majority)	*v* ACC	*do* GEN	*do* GEN	*do* GEN
destination (na-words subset)	*na* ACC	*na* ACC	*na* ACC	*na* ACC
destination (human)	*k* DAT	*do* GEN	*k* DAT	*k* DAT

Functions/Constructions	**Russian**	**Polish**	**Lit Czech**	**Coll Czech**
understand X	ACC	ACC	DAT	DAT/ACC
negation (subject)	GEN	GEN	NOM	NOM
negation (object)	ACC	GEN	ACC	ACC
negation (object), strong	GEN	GEN	GEN	ACC
BE predicate (pres)	NOM	INST	INST	NOM
BE predicate (past)	INST	INST	INST	NOM
BE predicate (future)	INST	INST	INST	NOM
at X (time of day, X:00)	*v* ACC	*o* LOC	*v* ACC	*v* ACC
ago	ACC *nazad*	ACC *temu*; *przed* INST	*před* INST	*před* INST
comparison, than	GEN, *čem* NOM	*od* GEN, *niż* NOM	*než* NOM	*než* NOM
comparison, amount by which	*na* ACC	ACC	*o* ACC	*o* ACC
every other X	*čerez* ACC	*co* ACC	*ob* ACC	*ob* ACC
the date is X	NOM	NOM	GEN	GEN
after X	*posle* GEN	*po* LOC	*po* LOC	*po* LOC
wish X something	GEN	GEN	ACC	ACC
take something from X	*u* GEN	DAT	DAT	DAT
around an area	*po* DAT	*po* LOC	*po* LOC	*po* LOC
ask X something	ACC, *u* GEN	ACC	GEN	GEN
ask someone X	*o* LOC	*o* ACC	*na* ACC	*na* ACC
distance from X	*ot* GEN	*od* GEN	*od* GEN	*od* GEN
close to X	*ot* GEN, *k* DAT	GEN	GEN	GEN
be interested in X	INST	INST	*o* ACC	*o* ACC
be afraid of X	GEN	GEN	GEN	GEN
become X	INST	iNST	INST	INST
in the fall	INST	INST, *w* LOC	*na* ACC	*na* ACC
in the spring	INST	*na* ACC, INST	*na* LOC	*na* LOC
in the summer	INST	INST, *w* LOC	*v* LOC	*v* LOC
in the winter	INST	INST, *w* LOC	*v* LOC	*v* LOC

Table 5
Slavic Case Semantics
Pilot Project Functions/Constructions

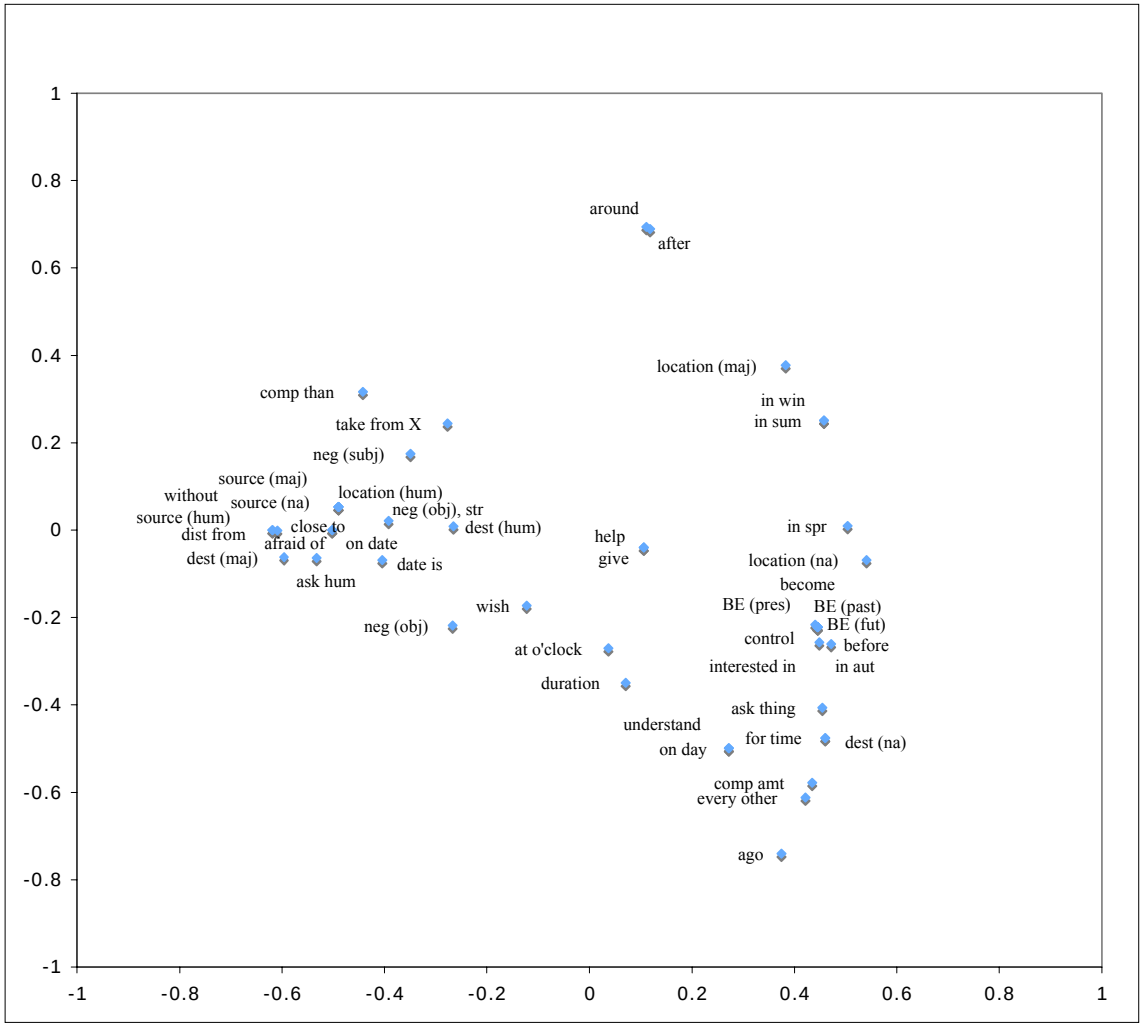

Figure 18
Slavic Case Semantics
Pilot Project Functions/Constructions

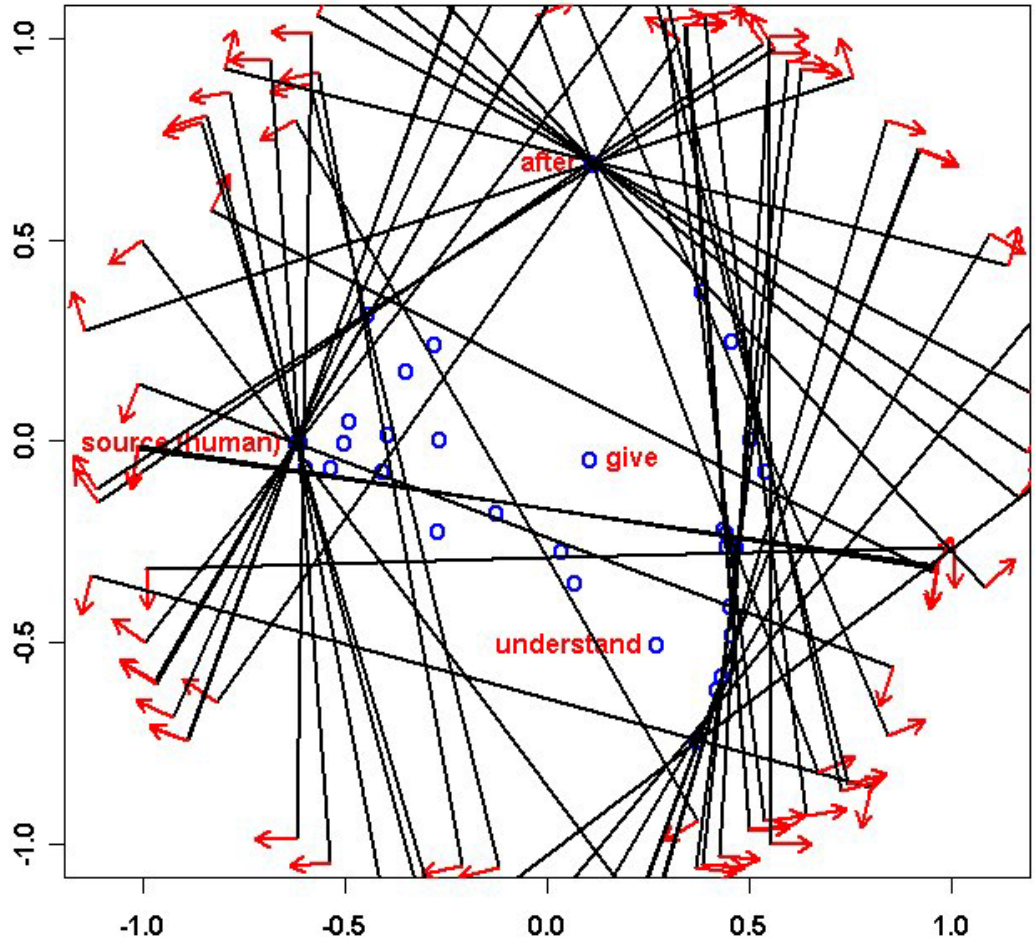

Figure 19
Slavic Case Semantics
Correctly Specified Case Marking
with cutting lines

	Correct Classification 2-D (3-D)	APRE 2-D (3-D)	Average Majority Margin
Correct Specification	97.0% (98.6%)	0.667 (0.845)	91.0%

Table 6
Slavic Case Semantics
Correct Classification and APRE Results

The resulting conceptual space with and without cutting lines is shown in Figures 18-19 and the measures of fit in Table 6. A cogent analysis of Figure 18 will not be attempted in the current paper, but it is sufficient to see that the MDS analysis is capable of clustering case functions together as well as considering "border zones" of overlapping case function. Genitive functions in the "west" transition into accusative functions to the

"southeast". The dative forms a central region with an instrumental belt to the "northeast". The locative case runs in a narrow band along from "north" to "south" on the far "east" side of the map. In addition, comparing the fit indicators for two and three dimensions, there may indeed be an argument to be made in favor of three dimensions for the full case map. However, all of this should be taken with a grain of salt. The data set here is imbalanced and incomplete. The conceptual space changes in sometimes subtle, sometimes dramatic ways when additional data is included. The data from the pilot project also underrepresents and misrepresents the bare case functions and prepositional functions that are simply not included at this point, resulting in overly lopsided "vote" margins.

However, the conceptual space in Figure 18 should be seen as a stepping stone to a larger goal. For future projects, I am developing a database for Slavic Case including 1200-1400 functions and constructions, intended to account for all case use in Slavic. This database is partially complete and will include the languages considered here as a minimum and then would be expanded to include Bosnian/Croatian/Serbian as well as remnants of case constructions and prepositional constructions in Bulgarian. I am interested in establishing a conceptual space for the types of relationships expressed by case within Slavic and then in expanding this data set to include the proper linguistic diversity to achieve a universal conceptual space through MDS analysis. In addition to uncovering this conceptual space for case relations, it is also hoped that the MDS analysis will shed new light on Janda and Clancy's analysis of Slavic case semantics in *The Case Book for Russian* (2002), *The Case Book for Czech* (2006), and *The Case Book for Polish* (Forthcoming) in order to confirm, refine, or challenge those analyses.

Conclusion

The conceptual spaces identified through empirical consideration and MDS analysis are understood to reflect something of the topology of linguistic concepts and how these concepts are structured mentally. These semantic maps also provide predictive power for language change, identifying the meanings that are more or less likely to be encompassed by polysemous words and morphemes under diachronic development. This paper extends the work of Croft and Pool (forthcoming) using MDS analysis to reveal the

semantic space of Slavic case, giving consideration to case semantics in Russian, Czech, and Polish. It is hoped that MDS analysis in linguistic research will provide a far-reaching tool for analyzing large samples of linguistic data while also providing a rigorously defined mathematical method that gives teeth to the powerful insights of cognitive linguistics.

Bibliography

Clancy, Steven J., and Laura A. Janda. Forthcoming. *The Case Book for Polish*.

Croft, William. 2001. *Radical construction grammar: syntactic theory in typological perspective.* Oxford: Oxford University Press.

_____. 2003. *Typology and universals*, 2nd edition. Cambridge: Cambridge University Press.

Croft, William, and Keith T. Poole. Forthcoming. "Inferring universals from grammatical variation: multidimensional scaling for typological analysis". http://lings.ln.man.ac.uk/Info/staff/WAC/Papers/MDSpaper.pdf http://lings.ln.man.ac.uk/Info/staff/WAC/Papers/MDSfigures.pdf

Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Oxford University Press.

_____. 2003. The geometry of grammatical meaning: semantic maps and crosslinguistic comparison. *The new psychology of language*, *vol. 2*, ed. Michael Tomasello, 211-42. Mahwah, NJ: Lawrence Erlbaum Associates.

Jakobson, R.O. 1936/1984. Contribution to the general theory of case: general meanings of the Russian cases [translation of Beitrag zur allgemeinen Kasuslehre: Gesamtbedeutung der russischen Kasus, originally in *TCLP* 4]. In: Linda R. Waugh and Morris Halle (eds.) *Roman Jakobson. Russian and Slavic grammar: Studies 1931-1981*. Berlin: Mouton de Gruyter, 59-103.

_____. 1958/1984. Morphological observations on Slavic declension (the structure of Russian case forms) [translation of "Morfologičeskie nabljudenija nad slavjanskim skloneniem (sostav russkix padežnix form)", originally in *American contributions to the Fourth International Congress of Slavists, Moscow*]. In: Linda R. Waugh and Morris Halle (eds.) *Roman Jakobson. Russian and Slavic grammar: Studies 1931-1981*. Berlin: Mouton de Gruyter, 105-133.

Janda, Laura A., and Steven J. Clancy. 2002. *The Case Book for Russian*. Bloomington, IN: Slavica.

Janda, Laura A., and Steven J. Clancy. In press 2006. *The Case Book for Czech*. Bloomington, IN: Slavica.

Keenan, Edward L. & Bernard Comrie 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8:63–99.

Poole, Keith T. 2005. *Spatial Models of Parliamentary Voting* (*Analytical Methods for Social Research*). Cambridge: Cambridge University Press.

_____. [Software for Optimal Classification Method.]
http://pooleandrosenthal.com/Optimal_Classification.htm