EPISTEMIC GAMES GROUP

FORMATTING DATA FOR EPISTEMIC NETWORK ANALYSIS

TECHNICAL REPORT 2014-1

DRAFT

DAVID WILLIAMSON SHAFFER

DWS@EDUCATION.WISC.EDU

ABSTRACT

Quantitative analysis of network data requires that the data be represented in a machine-interpretable format. The purpose of this report is to describe data formatting standards for importing data into the online Epistemic Network Analysis (ENA) tool. Key definitions are provided for how to represent *stanza-based interaction data* in ENA format in terms of *code, stanza, and unit columns*, and worked examples are provided for how to format data sets. Other guidelines for data formatting are provided, as well as information about preparing non-standard data sets.

GAMES AND PROFESSIONAL SIMULATIONS (GAPS)
TECHNICAL REPORT SERIES

FORMATTING DATA FOR EPISTEMIC NETWORK ANALYSIS

TECHNICAL REPORT 2014-1

STANZA-BASED INTERACTION DATA

There are a number of existing data formats for networks, including SIF (Simple Interaction Format), GML, XGMML, BioPAX, PSI-MI, GraphML, KGML (KEGG XML), SBML, OBO, and Gene Association, many of them optimized around specific kinds of network models. (See http://wiki.cytoscape.org/Cytoscape_User_Manual/Network_Formats for a summary of these formats.)

However, none of these extant formats are ideal for representing ENA data because the ENA modeling environment actually performs two separate but related functions on network data:

1. ENA *creates network models* from *(symmetrical) stanza-based interaction* data.[1]

2. ENA creates an *analytic space* in which these models can be compared.

Therefore, rather than importing network models directly, ENA uses data formatting specifications that are designed to accept stanza-based interaction data and convert them into models with accompanying analytical spaces.

DEFINITION: SYMMETRICAL STANZA-BASED INTERACTION DATA

Stanza-based interaction data refers to information about a set of *objects*, the way they *relate* to one another, and a series of *stanzas* which reveal *evidence* about the relations between the objects:

1. *Objects* can refer to people, concepts, or anything whose network of connections is being modeled.

2. *Relations between objects* can refer to associations like strength of social tie, conceptual similarity, any connection, interaction, or association that links one object to another.

3. *Stanzas* can be units of time, or steps in a process, or any way of identifying a unit in the data for quantifying relations between objects.

4. *Evidence* refers to any specific elements of the data that can be used to identify the relations being modeled.

EXAMPLE: MODEL 1

To use a concrete example, consider 3 students working on a problem in a group. There are a number of ways we could model this data. For instance we could consider *students* as the objects, *agreement* as the relation, turns of talk or *utterances* as the stanzas, and *reference to prior statements* as

[1] ENA can model *asymmetrical* or *directional* interaction data, which is why the term symmetrical appears in parentheses. This more advanced feature of the tool will be covered later in the report.

evidence of agreement. This would provide a network model of the pattern of agreement among students in the group.

Imagine the following turns of talk occurred in the data:

A: I feel the Cadmium Battery was the best choice because it is inexpensive and is relatively reliable.
B: I agree with that, but I am concerned about its weight.
C: Yes weight is a problem, but I still think it is the best choice.
B: Reliability is key. Definitely.

We could represent this in terms of our model in the following data table:

STANZA NUMBER	STUDENT (OBJECT)	UTTERANCE (STANZA)	REFERENCE TO PRIOR STATEMENTS (EVIDENCE)	AGREEMENT WITH ANOTHER OBJECT… (RELATION)
1	A	I feel the Cadmium Battery was the best choice because it is inexpensive and is relatively reliable.		No
2	B	I agree with that, but I am concerned about its weight.	"I agree with that"	Between A and B
3	C	Yes weight is a problem, but I still think it is the best choice.	"Yes weight is a problem" reiterates "I am concerned about its weight"	Between C and A
4	B	Reliability is key. Definitely.	"reliability"	Between A and B

Figure 1: Student talk represented in Model 1.

Note that here we make an assertion about the *evidence* for each *stanza* in the data. That assertion, in turn, allows us to infer the *relation* of interest between one or more objects. In Stanza 4, for example, Student B repeats the word "reliably" from Student A's earlier contribution. This is taken as an assertion of agreement between Students A and B.

If we were to convert this short excerpt from the data into a network model of agreement, it would look like this:

MEANING SHARED WITH	A	B	C
A	**0**	2	1
B	2	**0**	0
C	1	0	**0**

Figure 2: Adjacency Matrix for Model 1

We read the table across the rows: in this portion of the data, B had two utterances in which there was evidence of agreement with A. C had one utterance in which there was evidence of agreement with A.

Notice that the network is shown here as an *adjacency matrix*, which is a way of summarizing the strength of the relations between the objects. The objects are indicated as the row and columns of the matrix. The cell $m_{i,j}$ shows the strength of the relation object i→object j. The diagonal of the matrix (with the **bold numbers**) contains all zeros, which indicates that objects are not (in this model, and in general in ENA) related to themselves; that is object i→object i is always 0.

Notice also that this network *symmetrical*. When B is in agreement with A about something, then A is in agreement with B. That is, for any pair of objects in the network, the strength of the relationship A→B is the same as the relationship B→A. It would be possible to model this data asymmetrically, such that B's assertion that s/he agrees with A does not imply that A also agrees with B. Setting up asymmetrical models will be covered later in this report.

CODE COLUMNS

Figure 1 above can be converted easily into the ENA format for stanza-based interaction data. The first key change is in the representation of the Relation column in the data—or, more specifically, the replacement of the single Relation column by a *set of code columns* that specifies the objects being related.

STANZA NUMBER	STUDENT	UTTERANCE	RELATION	A	B	C
1	A	I feel the Cadmium Battery was the best choice because it is inexpensive and is relatively reliable.	No	1	0	0
2	B	I agree with that, but I am concerned about its weight.	Between A and B	1	1	0
3	C	Yes weight is a problem, but I still think it is the best choice.	Between C and A	1	0	1
4	B	Reliability is key. Definitely.	Between A and B	1	1	0

Figure 3: Student talk from Model 1 in ENA format stanza-based interaction data.

Figure 3 illustrates this transformation. In the original data, relations were represented in each line of the data by reading one indicated in the Relation column. In ENA format, the relation information is represented in a series of columns that are referred to as *codes* such that:

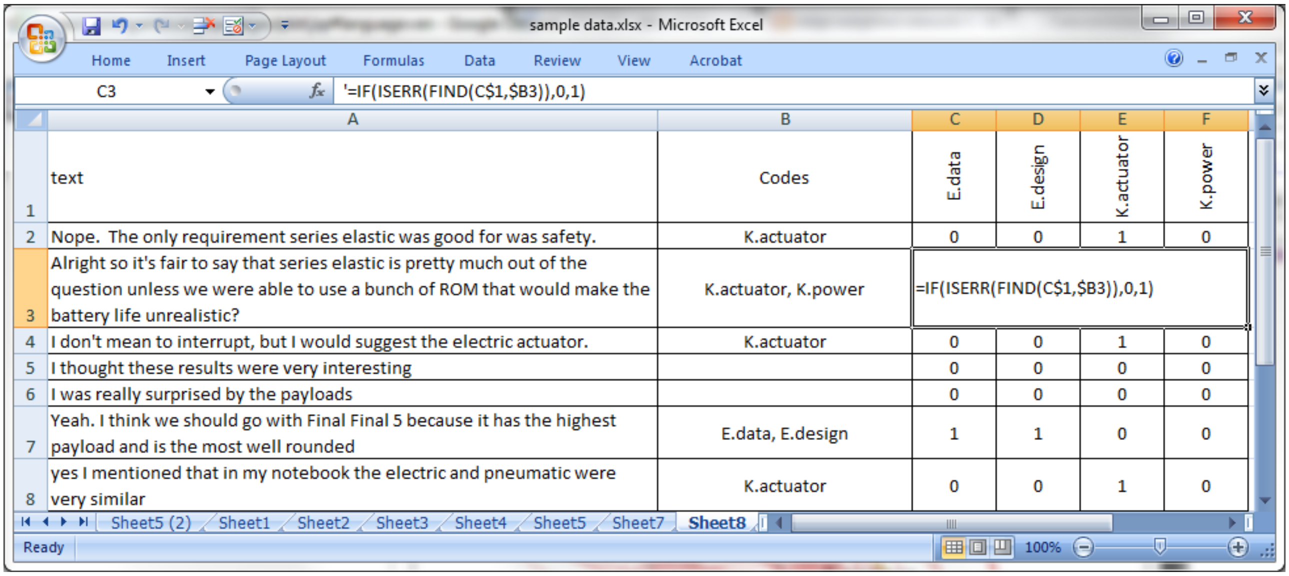
1. Each object in the model is assigned a single code column.

2. The value in the code column for object A, B, or C indicates, for each line in the data, whether that object is being *related* to some other object.

3. Every line in the data has a value for each code.

Typically, the values are binary (0 or 1). Fractional values or other weights can be used, but that is discussed later in this report.

For binary codes a simple Excel formula can often translate from a single column indicating codes in a line of data to the required code columns for each object:



C3 '=IF(ISERR(FIND(C$1,$B3)),0,1)

	A	B	C	D	E	F
1	text	Codes	E.data	E.design	K.actuator	K.power
2	Nope. The only requirement series elastic was good for was safety.	K.actuator	0	0	1	0
3	Alright so it's fair to say that series elastic is pretty much out of the question unless we were able to use a bunch of ROM that would make the battery life unrealistic?	K.actuator, K.power	=IF(ISERR(FIND(C$1,$B3)),0,1)			
4	I don't mean to interrupt, but I would suggest the electric actuator.	K.actuator	0	0	1	0
5	I thought these results were very interesting		0	0	0	0
6	I was really surprised by the payloads		0	0	0	0
7	Yeah. I think we should go with Final Final 5 because it has the highest payload and is the most well rounded	E.data, E.design	1	1	0	0
8	yes I mentioned that in my notebook the electric and pneumatic were very similar	K.actuator	0	0	1	0

Figure 3b: Sample Excel sheet showing an Excel formula to convert a single column of codes into binary code columns for ENA formatted data.

HOW ENA CONSTRUCTS A NETWORK MODEL FROM DATA IN ENA FORMAT

To construct a network model from data in ENA format:

1. ENA converts the code columns into an adjacency matrix for each stanza in the data; then

2. ENA adds the adjacency matrices together into a *cumulative adjacency matrix* for the network.

For example, the data above from Model 1 produces 4 adjacency matrices, one for each stanza:

<table>
<tr><td colspan="16">ADJACENCY MATRICIES BY STANZA</td></tr>
<tr><td colspan="4">1</td><td colspan="4">2</td><td colspan="4">3</td><td colspan="4">4</td></tr>
<tr><td></td><td>A</td><td>B</td><td>C</td><td></td><td>A</td><td>B</td><td>C</td><td></td><td>A</td><td>B</td><td>C</td><td></td><td>A</td><td>B</td><td>C</td></tr>
<tr><td>A</td><td>0</td><td>0</td><td>0</td><td>A</td><td>0</td><td>1</td><td>0</td><td>A</td><td>0</td><td>0</td><td>1</td><td>A</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>B</td><td>0</td><td>0</td><td>0</td><td>B</td><td>1</td><td>0</td><td>0</td><td>B</td><td>0</td><td>0</td><td>0</td><td>B</td><td>1</td><td>0</td><td>0</td></tr>
<tr><td>C</td><td>0</td><td>0</td><td>0</td><td>C</td><td>0</td><td>0</td><td>0</td><td>C</td><td>1</td><td>0</td><td>0</td><td>C</td><td>0</td><td>0</td><td>0</td></tr>
</table>

Figure 4: Student talk from Model 1 in as adjacency matrices by stanza

So, for example, the adjacency matrix for stanza 1 is all zeros because there is only one code in stanza 1, so there is no relation *between* objects indicated in the stanza. Stanzas 2 and 4 are identical, since the codes in the data are identical, and in both cases indicate a relation between A and B. And finally Stanza 3 shows agreement between A and C.

If we sum these matrices, we get the following cumulative adjacency matrix:

MEANING SHARED WITH→	A	B	C
A	**0**	2	1
B	2	**0**	0
C	1	0	**0**

Figure 5: Cumulative Adjacency Matrix for Model 1 from ENA format data

Careful readers will quickly note that this is, in fact, the same matrix as Figure 2 above.

STANZA COLUMNS

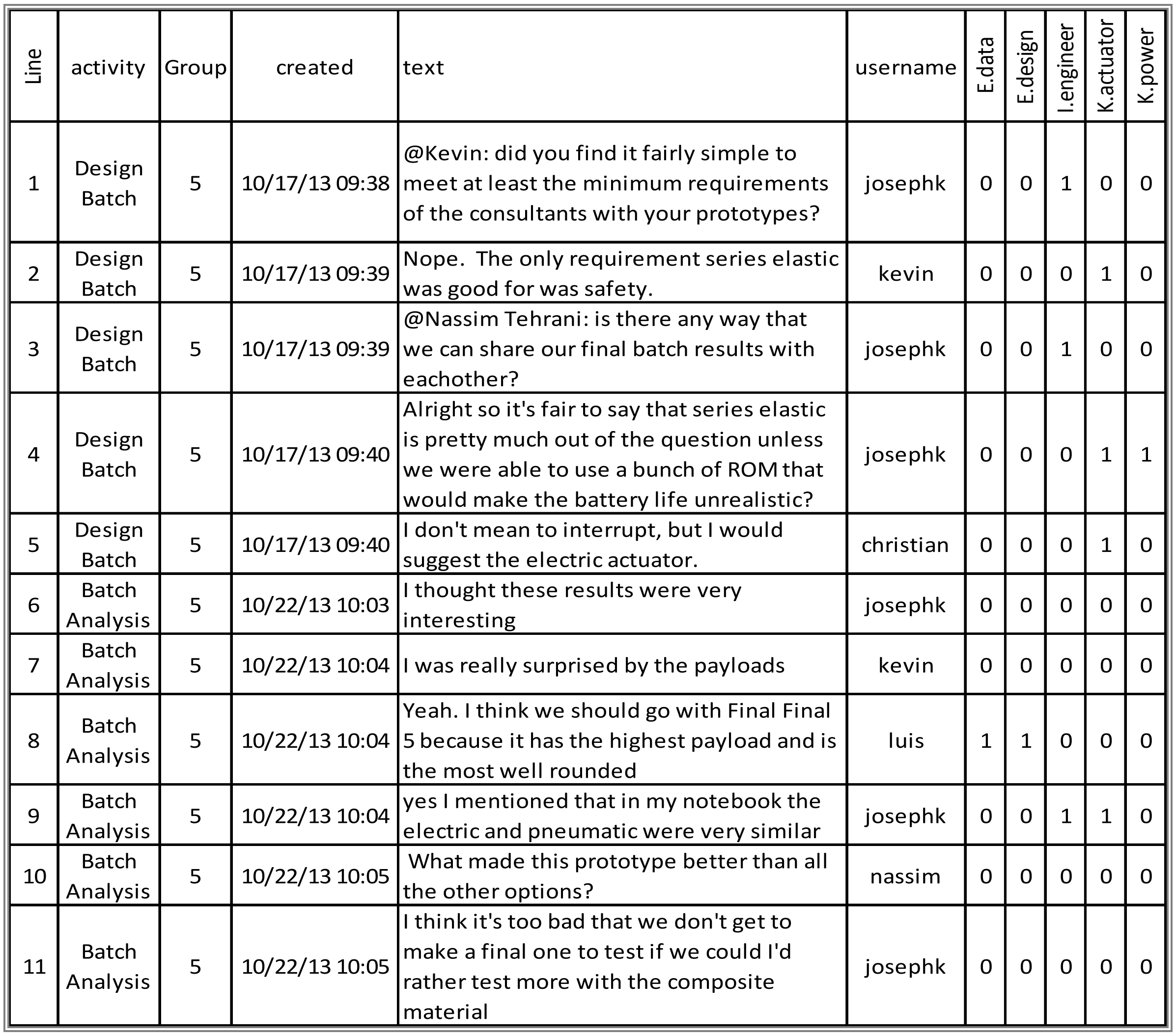
The second key feature is identifying stanzas. In Model 1 each turn of talk (or line in the data table) corresponded to a stanza: meaning, an adjacency matrix was computed for each line of talk in the model. The data was coded or tagged such that each line indicated agreement with previous lines.

But in many (indeed, probably most) cases where relations arise *among, across, or between* the lines of talk, these relations need to be *inferred from the data*, rather than having the relations coded directly.

In such cases, rather than taking a single line of data as a stanza, we need to identify stanzas with *collections of lines of talk* rather than with individual lines, and then infer or compute the relations among objects within each stanza (collection of lines).

MODEL 2

To see how this works, let's use a second example, this time excerpted from real data from a logfile of student work in an engineering simulation. (See Figure 6.) The source of the data is chat messages between members of the group.



Line	activity	Group	created	text	username	E.data	E.design	I.engineer	K.actuator	K.power
1	Design Batch	5	10/17/13 09:38	@Kevin: did you find it fairly simple to meet at least the minimum requirements of the consultants with your prototypes?	josephk	0	0	1	0	0
2	Design Batch	5	10/17/13 09:39	Nope. The only requirement series elastic was good for was safety.	kevin	0	0	0	1	0
3	Design Batch	5	10/17/13 09:39	@Nassim Tehrani: is there any way that we can share our final batch results with eachother?	josephk	0	0	1	0	0
4	Design Batch	5	10/17/13 09:40	Alright so it's fair to say that series elastic is pretty much out of the question unless we were able to use a bunch of ROM that would make the battery life unrealistic?	josephk	0	0	0	1	1
5	Design Batch	5	10/17/13 09:40	I don't mean to interrupt, but I would suggest the electric actuator.	christian	0	0	0	1	0
6	Batch Analysis	5	10/22/13 10:03	I thought these results were very interesting	josephk	0	0	0	0	0
7	Batch Analysis	5	10/22/13 10:04	I was really surprised by the payloads	kevin	0	0	0	0	0
8	Batch Analysis	5	10/22/13 10:04	Yeah. I think we should go with Final Final 5 because it has the highest payload and is the most well rounded	luis	1	1	0	0	0
9	Batch Analysis	5	10/22/13 10:04	yes I mentioned that in my notebook the electric and pneumatic were very similar	josephk	0	0	1	1	0
10	Batch Analysis	5	10/22/13 10:05	What made this prototype better than all the other options?	nassim	0	0	0	0	0
11	Batch Analysis	5	10/22/13 10:05	I think it's too bad that we don't get to make a final one to test if we could I'd rather test more with the composite material	josephk	0	0	0	0	0

Figure 6: Logfile data coded in ENA format

The research question of interest in this case (or one possible research question of interest) is how these students are making connections between different aspects of engineering in this design space.

In this case the *objects* in the network model can be seen in the code columns (the 5 rightmost columns with vertical headers). For those interested, the codes in this case represent KNOWLEDGE (K) of actuators and power in the object being designed, IDENTITY (I) of an engineer, and the EPISTEMOLOGY (E), or decision-making and justification processes of engineering. These are just a subset of the engineering issues relevant to this particular problem and its solution. The actual data had both more lines of talk, and more codes.

There are two important points to consider in this data. The first is that this is a case where sometimes we can see connections among the elements of the design space (the codes) in a single line of data: for example, in Line 4 JosephK references knowledge of both actuators and power:

> *Alright so it's fair to say that series elastic is pretty much out of the question unless we were able to use a bunch of* ROM *that would make the battery life unrealistic?*

But in general in this data it does not make sense for each *line* to be a *stanza*. Each line is coded for the elements of engineering design that are referenced *in that line*. But the lines of talk are aksi related to one another.

In Line 9, for example, JosephK talks about recording actuator properties in his own notebook:

> *yes I mentioned that in my notebook the electric and pneumatic were very similar*

But he is actually making a connection between those elements of the design space, and the data and design considerations Luis used in his decision making process in the preceding Line 8:

> *Yeah. I think we should go with [the design labeled] Final Final 5 because it has the highest payload and is the most well rounded*

We thus want to look at relations among objects not just *within* specific turns of talk, but *among* or *across* turns of talk.

In this sense in Model 2 we want the stanzas to encompass multiple lines of the data.

Of course, (and this is the second important point) we don't necessarily want to include *all* of the lines of the data in a single stanza.

For example, an examination of the columns labeled Activity and Created in Figure 6 show that there are actually *two separate conversations* here. Both involved the same group of students (Group 5), but the conversations took place on two different days, while the students were working on two different activities. On the first day they designed prototypes and were deciding which ones to send to a fabrication lab for testing. On the second day, they were reviewing the results of the tests.

So in this case, although JosephK's comment in Line 9 should be connected to Luis's comment in Line 8, it does not make as much sense for it to be linked to the comments in Lines 1-5, which took place in a conversation from the previous week.

Of course, precisely which lines should be included together in stanzas is ultimately an analytical question. From the point of view of *data formatting*, the key issue is that there has to be some *column or columns in the data that indicates how to group lines into stanzas*.

Conceptually, the key idea behind a stanza is that:

1. Objects in lines *anywhere within the same stanza* are *related* to one another in the model.

2. Objects in lines that are *not in the same stanza* are *not related* to one another in the model.

Technically, the ENA tool treats stanzas are *categorical*: all lines with the same value in the stanza column are treated as being in the same stanza. Note that means in this case if the Activity column were used to indicate stanzas, it would result in two stanzas, one from that conversation during the "Design Batch" activity, and one from the conversation during the "Batch Analysis" activity. If Created was chosen as a Stanza column it would result in a different stanza for each minute of the conversation.

HOW ENA CONSTRUCTS A NETWORK MODEL FROM MULTI-LINED STANZAS

To construct a network model from data in which stanzas are composed of multiple lines, ENA first *collapses* the stanzas into single data entries.

Usually it does this as a *binary OR*, meaning if any line of data in the stanza contains code A, then the stanza contains code A. There are circumstances when it makes more sense to sum the lines for each code across the stanza, but that is a question of analysis rather than data formatting and so is beyond the scope of this report.

Using this method Model 2 with stanzas by Activity would be reduced to 2 stanzas:

Activity	Group	E.data	E.design	I.engineer	K.actuator	K.power
Design Batch	5	0	0	1	1	1
Batch Analysis	5	1	1	1	1	0

Figure 7: Stanzas by Activity for Model 2

Note that in the data, all of the columns (other than the codes) that *change across the stanza* are removed at this point, because there is no sensible way to associate, for example, a single username with the accumulated lines from the group.

ENA then proceeds to create adjacency matrices for each stanza as before with Model 1, and sums them across all stanzas into a cumulative adjacency matrix. The Unit column(s) thus identify the final network models, which become the *units of analysis* in the analysis of the networks.

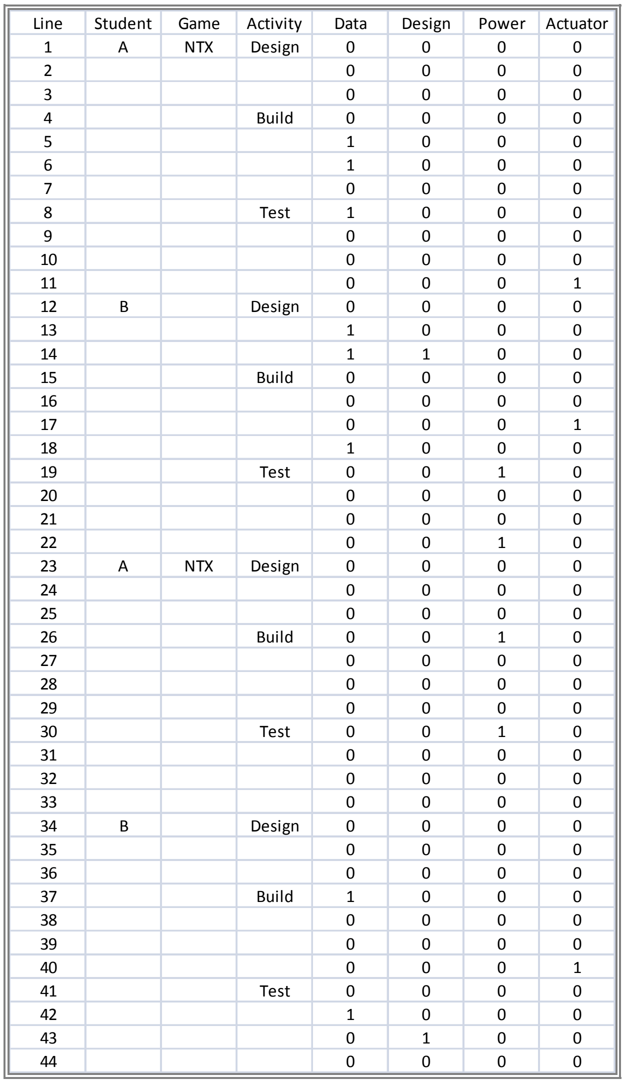
UNIT COLUMNS

While Code Columns and Stanza Columns are sufficient for ENA to *create network models*, in order to analyze these models (that is, to *construct an analytic space*), ENA needs more than one network.

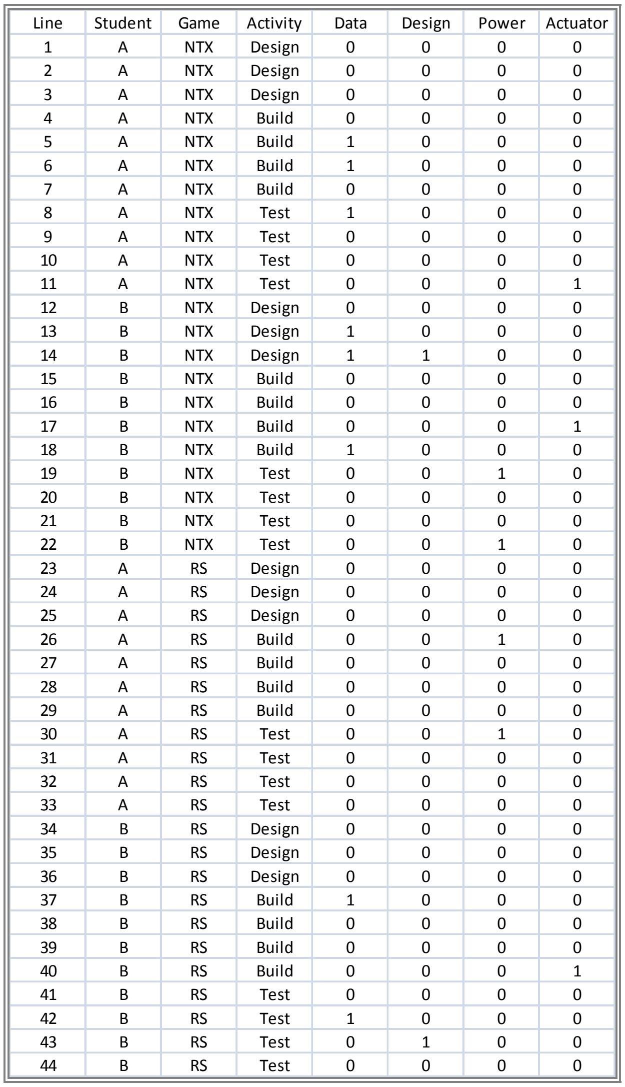
Just as collections of lines belonging to the same stanza are indicated by one or more stanza columns in the data, the lines belonging to different networks (which might be networks from different people, or from people in different conditions, for example) are indicated by Unit Columns.

As in the example above with Stanza Columns, ENA looks for an entry in the Unit column or columns for each line, and assigns the line to a Unit of analysis (a network) based on the value in those columns. This means that each line must have a value associated with each Stanza and Unit column.

So, for example, even though to a human reader it is reasonable clear which lines are associated with which networks in the data on the left in Figure 8, it is not correctly formatted. Because there are entries for every line in every column, the data set on the right in Figure 8 is.



Line	Student	Game	Activity	Data	Design	Power	Actuator
1	A	NTX	Design	0	0	0	0
2				0	0	0	0
3				0	0	0	0
4			Build	0	0	0	0
5				1	0	0	0
6				1	0	0	0
7				0	0	0	0
8			Test	1	0	0	0
9				0	0	0	0
10				0	0	0	0
11				0	0	0	1
12	B		Design	0	0	0	0
13				1	0	0	0
14				1	1	0	0
15			Build	0	0	0	0
16				0	0	0	0
17				0	0	0	1
18				1	0	0	0
19			Test	0	0	1	0
20				0	0	0	0
21				0	0	0	0
22				0	0	1	0
23	A	NTX	Design	0	0	0	0
24				0	0	0	0
25				0	0	0	0
26			Build	0	0	1	0
27				0	0	0	0
28				0	0	0	0
29				0	0	0	0
30			Test	0	0	1	0
31				0	0	0	0
32				0	0	0	0
33				0	0	0	0
34	B		Design	0	0	0	0
35				0	0	0	0
36				0	0	0	0
37			Build	1	0	0	0
38				0	0	0	0
39				0	0	0	0
40				0	0	0	1
41			Test	0	0	0	0
42				1	0	0	0
43				0	1	0	0
44				0	0	0	0

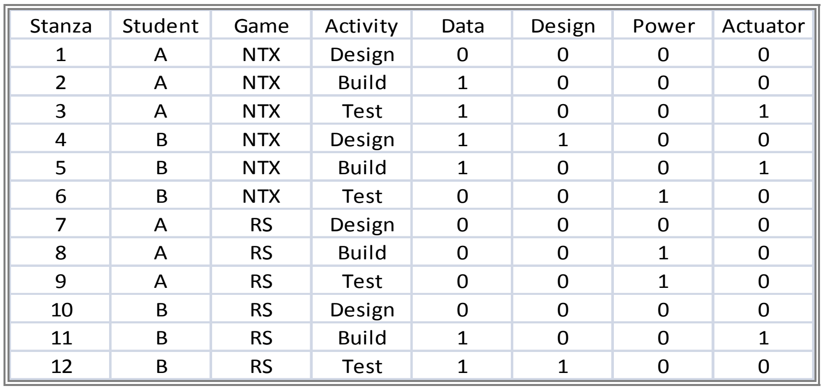


Line	Student	Game	Activity	Data	Design	Power	Actuator
1	A	NTX	Design	0	0	0	0
2	A	NTX	Design	0	0	0	0
3	A	NTX	Design	0	0	0	0
4	A	NTX	Build	0	0	0	0
5	A	NTX	Build	1	0	0	0
6	A	NTX	Build	1	0	0	0
7	A	NTX	Build	0	0	0	0
8	A	NTX	Test	1	0	0	0
9	A	NTX	Test	0	0	0	0
10	A	NTX	Test	0	0	0	0
11	A	NTX	Test	0	0	0	1
12	B	NTX	Design	0	0	0	0
13	B	NTX	Design	1	0	0	0
14	B	NTX	Design	1	1	0	0
15	B	NTX	Build	0	0	0	0
16	B	NTX	Build	0	0	0	0
17	B	NTX	Build	0	0	0	1
18	B	NTX	Build	1	0	0	0
19	B	NTX	Test	0	0	1	0
20	B	NTX	Test	0	0	0	0
21	B	NTX	Test	0	0	0	0
22	B	NTX	Test	0	0	1	0
23	A	RS	Design	0	0	0	0
24	A	RS	Design	0	0	0	0
25	A	RS	Design	0	0	0	0
26	A	RS	Build	0	0	1	0
27	A	RS	Build	0	0	0	0
28	A	RS	Build	0	0	0	0
29	A	RS	Build	0	0	0	0
30	A	RS	Test	0	0	1	0
31	A	RS	Test	0	0	0	0
32	A	RS	Test	0	0	0	0
33	A	RS	Test	0	0	0	0
34	B	RS	Design	0	0	0	0
35	B	RS	Design	0	0	0	0
36	B	RS	Design	0	0	0	0
37	B	RS	Build	1	0	0	0
38	B	RS	Build	0	0	0	0
39	B	RS	Build	0	0	0	0
40	B	RS	Build	0	0	0	1
41	B	RS	Test	0	0	0	0
42	B	RS	Test	1	0	0	0
43	B	RS	Test	0	1	0	0
44	B	RS	Test	0	0	0	0

Figure 8: Two data tables showing Unit, Stanza, and Code columns. Table on the left is not properly formatted because values for Unit and Stanza are not indicated for all lines (rows) in the data.

CHECK YOUR UNDERSTANDING

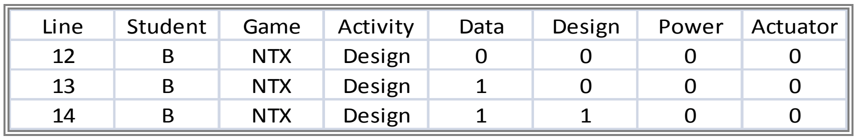
If you want to check your understanding of how ENA accumulates stanzas and Units, consider the set in Figure 8. If the Unit Columns are Student and Game, and the Stanza Column is Activity, then the stanzas for the data set on the right are:



Stanza	Student	Game	Activity	Data	Design	Power	Actuator
1	A	NTX	Design	0	0	0	0
2	A	NTX	Build	1	0	0	0
3	A	NTX	Test	1	0	0	1
4	B	NTX	Design	1	1	0	0
5	B	NTX	Build	1	0	0	1
6	B	NTX	Test	0	0	1	0
7	A	RS	Design	0	0	0	0
8	A	RS	Build	0	0	1	0
9	A	RS	Test	0	0	1	0
10	B	RS	Design	0	0	0	0
11	B	RS	Build	1	0	0	1
12	B	RS	Test	1	1	0	0

Figure 9: Stanzas from the data in Right Hand side of Figure 8, with Student and Game as Unit Columns and Activity as the Stanza Column

To see why, consider Student B while he or she was playing the game NTX. In the activity marked Design, there are 3 lines:



Line	Student	Game	Activity	Data	Design	Power	Actuator
12	B	NTX	Design	0	0	0	0
13	B	NTX	Design	1	0	0	0
14	B	NTX	Design	1	1	0	0

Figure 10: Lines in the Stanza "Build" for Student A in Game NTX from the data in Right Hand side of Figure 8

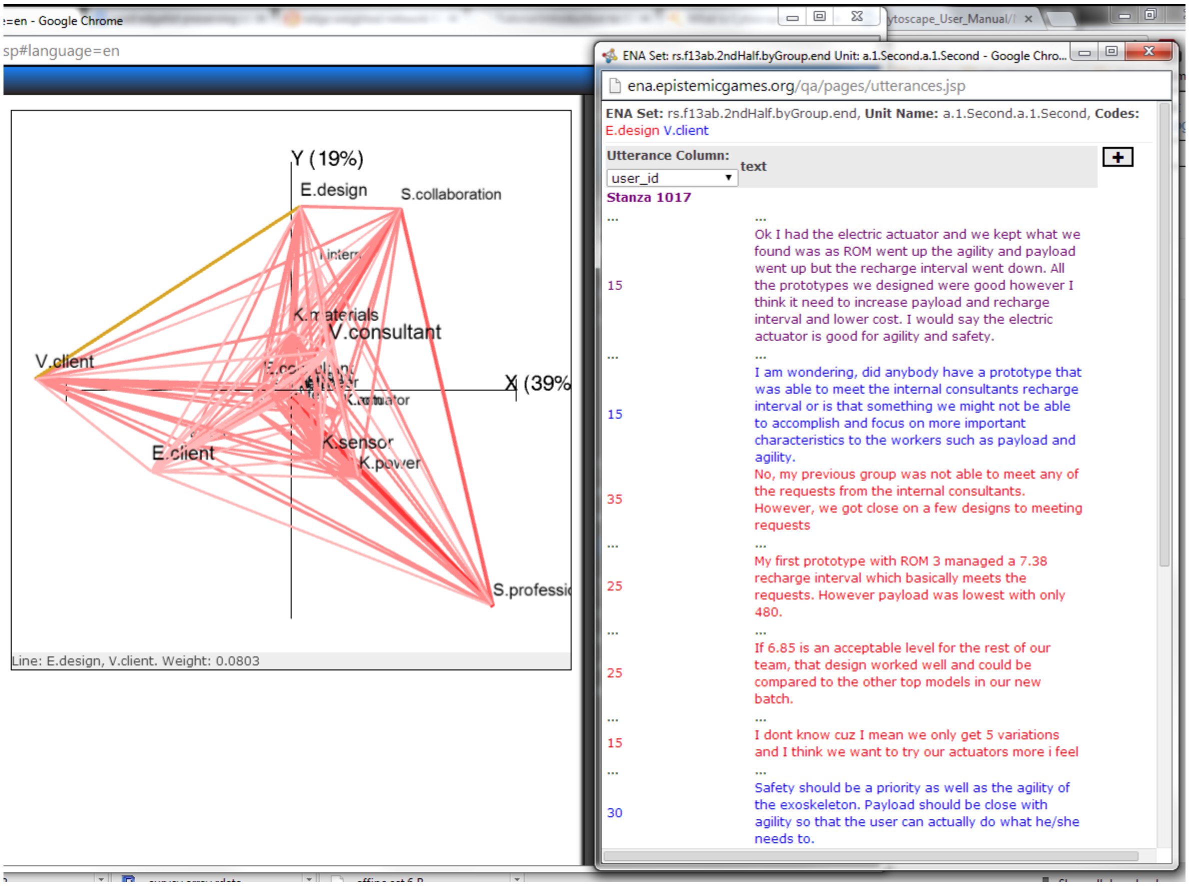
Across all the lines of this stanza, the objects Data and Design are coded. Thus Stanza 4 of the stanza list—which corresponds to Student B in NTX in the Design activity has codes 1,1,0,0 for Data, Design, Power, and Actuator.

There would be 4 cumulative adjacencies in this set: Student A in NTX, Student B in NTX, Student A in RS, Student B in RS.

RAW DATA COLUMNS

In addition to Unit, Stanza, and Code columns, an ENA formatted data set may contain any other columns of identifying information about the data. In particular, the Raw Data that was coded with the objects of interest in the model can be included in the data set. (See, for example, Figure 6.) This might include a column for the raw data excerpts, for the speaker (in the case of discourse data), time of the action coded, and so on.

When ENA analyses are computed, additional data can be linked to the model to show the precise activities or actions that support the network model. (See Figure 11 for an example.)



Line: E.design, V.client. Weight: 0.0803

ENA Set: rs.f13ab.2ndHalf.byGroup.end Unit: a.1.Second.a.1.Second - Google Chro...

ena.epistemicgames.org/qa/pages/utterances.jsp

ENA Set: rs.f13ab.2ndHalf.byGroup.end, **Unit Name:** a.1.Second.a.1.Second, **Codes:** E.design V.client

Utterance Column: user_id text

Stanza 1017

...

15 Ok I had the electric actuator and we kept what we found was as ROM went up the agility and payload went up but the recharge interval went down. All the prototypes we designed were good however I think it need to increase payload and recharge interval and lower cost. I would say the electric actuator is good for agility and safety.

...

15 I am wondering, did anybody have a prototype that was able to meet the internal consultants recharge interval or is that something we might not be able to accomplish and focus on more important characteristics to the workers such as payload and agility.

35 No, my previous group was not able to meet any of the requests from the internal consultants. However, we got close on a few designs to meeting requests

...

25 My first prototype with ROM 3 managed a 7.38 recharge interval which basically meets the requests. However payload was lowest with only 480.

...

25 If 6.85 is an acceptable level for the rest of our team, that design worked well and could be compared to the other top models in our new batch.

...

15 I dont know cuz I mean we only get 5 variations and I think we want to try our actuators more i feel

...

30 Safety should be a priority as well as the agility of the exoskeleton. Payload should be close with agility so that the user can actually do what he/she needs to.

Figure 11: Sample ENA network showing raw data supporting connections between codes for Valuing Client (V.Client) and the Epistemology of Design (E.Design)

TYPES OF COLUMNS

In general, then, ENA formatted data has two types of column:

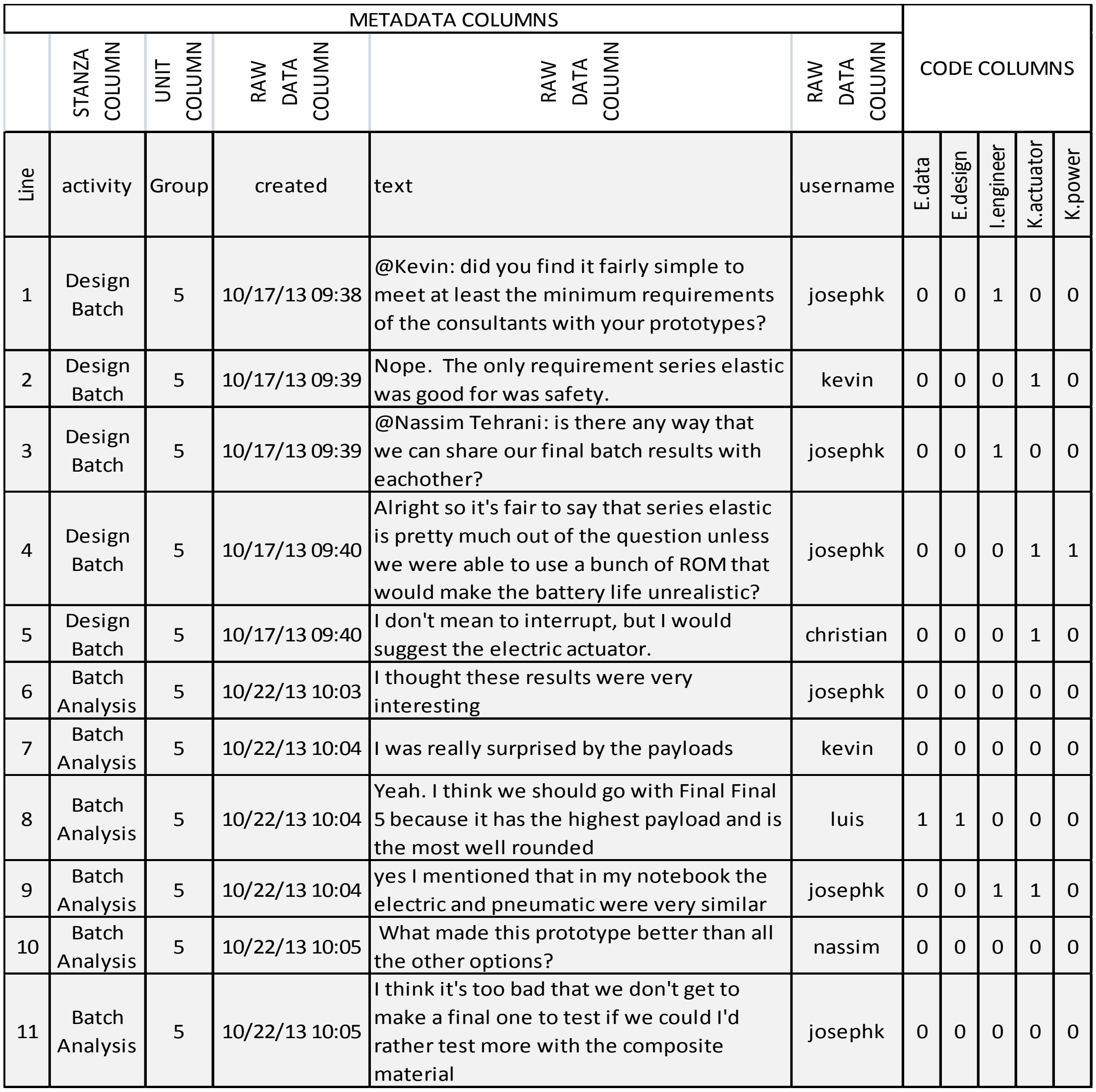
1. Code Columns, containing binary data indicating the presence or absence of each object in the model within each line of data.

2. Metadata Columns, containing any other information relevant to the data set, but at a minimum containing data necessary to associate each line of data with:

    a. Its appropriate *unit* (ie, unit of analysis)

    b. Its appropriate *stanza* (ie, collection of lines within which objects might be connected to one another)

These columns can appear in any convenient order in the data file.



METADATA COLUMNS						CODE COLUMNS				
	STANZA COLUMN	UNIT COLUMN	RAW DATA COLUMN	RAW DATA COLUMN	RAW DATA COLUMN					
Line	activity	Group	created	text	username	E.data	E.design	I.engineer	K.actuator	K.power
1	Design Batch	5	10/17/13 09:38	@Kevin: did you find it fairly simple to meet at least the minimum requirements of the consultants with your prototypes?	josephk	0	0	1	0	0
2	Design Batch	5	10/17/13 09:39	Nope. The only requirement series elastic was good for was safety.	kevin	0	0	0	1	0
3	Design Batch	5	10/17/13 09:39	@Nassim Tehrani: is there any way that we can share our final batch results with eachother?	josephk	0	0	1	0	0
4	Design Batch	5	10/17/13 09:40	Alright so it's fair to say that series elastic is pretty much out of the question unless we were able to use a bunch of ROM that would make the battery life unrealistic?	josephk	0	0	0	1	1
5	Design Batch	5	10/17/13 09:40	I don't mean to interrupt, but I would suggest the electric actuator.	christian	0	0	0	1	0
6	Batch Analysis	5	10/22/13 10:03	I thought these results were very interesting	josephk	0	0	0	0	0
7	Batch Analysis	5	10/22/13 10:04	I was really surprised by the payloads	kevin	0	0	0	0	0
8	Batch Analysis	5	10/22/13 10:04	Yeah. I think we should go with Final Final 5 because it has the highest payload and is the most well rounded	luis	1	1	0	0	0
9	Batch Analysis	5	10/22/13 10:04	yes I mentioned that in my notebook the electric and pneumatic were very similar	josephk	0	0	1	1	0
10	Batch Analysis	5	10/22/13 10:05	What made this prototype better than all the other options?	nassim	0	0	0	0	0
11	Batch Analysis	5	10/22/13 10:05	I think it's too bad that we don't get to make a final one to test if we could I'd rather test more with the composite material	josephk	0	0	0	0	0

Figure 12: ENA formatted data, showing types of data columns

OTHER CONSIDERATIONS

In addition to the above conceptual issues in the ENA format, there are some other structural issues to keep in mind when formatting you data:

1. The best format for data to be uploaded is .csv; however, different countries have different default delimiters for .csv files. ENA works best with comma delimiters used in US .csv files.

2. Data in ENA format is represented as a series of rows, all with equivalent format.

3. There must be a header row indicating column names. Blank column names can cause erratic behavior.

4. Empty cells can cause erratic behavior during upload. It is better to use NA or some other identifier for missing data.

5. Missing data in the file can also cause erratic behavior during analysis. If the intent is to "ignore" rows of data with missing elements, it is best to remove these from the data set before uploading the data.

ADVANCED FORMATTING FOR SPECIAL CASES

While ENA has been optimized for analyzing symmetric data with binary codes, it is possible to examine data sets where objects can have fractional or other weights associated with each line of data, as well as to model non-symmetric relations.

WEIGHTED CODES

The ENA format for weighted codes is the same as for binary codes: simply put the weight for each code for each line of data in place of the 0 or 1 for the binary code in the code column. During network model construction ENA provides several options for accumulating weighted codes.

Weighted codes are multiplied by one another in the construction of adjacency matrices. This works well for fractional codes that indicate a probability that a particular object is present in the raw data. However, when integer or other large values are used, there are conceptual issues in the analysis that arise from the process of constructing adjacency matrices by multiplication. These are beyond the scope of this report on formatting, however.

NON-SYMMETRIC RELATIONS

Non-symmetric relations require more sophisticated transformation of the data for representation in ENA format. However, the basic principle for representing non-symmetric or directional data in ENA format is to create *two code columns* for each object identified in the data: one that is identified as a *sender* and the other as a *receiver*. Each line of data is then represented in the code columns by indicating a sender and receiver object.

For example, the code columns for a directional or asymmetric network in Model 1 would look like this:

STUDENT	UTTERANCE	RELATION	A SEND	B SEND	C SEND	A REC	B REC	C REC
A	I feel the Cadmium Battery was the best choice because it is inexpensive and is relatively reliable.	No	1	0	0	0	0	0
B	I agree with that, but I am concerned about its weight.	A→B	0	1	0	1	0	0
C	Yes weight is a problem, but I still think it is the best choice.	C→A	0	0	1	1	0	0
B	Reliability is key. Definitely.	A→B	0	1	0	1	0	0

Figure 13: Code columns for Model 1 as an asymmetrical network in ENA format stanza-based interaction data.

When ENA models the data it represents sending and receiving as separate functions of each object. This has both limitations and useful properties from an analytic point of view, both of which are beyond the scope of this report on data formatting.

ACKNOWLEDGMENTS

This work was funded in part by the Macarthur Foundation and the National Science Foundation through grants REC-0347000, DUE-0919347, DUE-1225885, EEC-1232656, DRL-0918409, DRL-0946372, and DRL-1247262. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.