Army Research Laboratory

Application of Social Network Analysis Techniques to Machine Translated Documents

by Ann E. M. Bornstein, John H. Brand, Michelle C. McVey,
and Sean Murray

ARL-MR-741 April 2010

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5067

ARL-MR-741 **April 2010**

Application of Social Network Analysis Techniques to Machine Translated Documents

Ann E. M. Bornstein and John H. Brand
Computational and Information Sciences Directorate, ARL

Michelle C. McVey
University of Maryland

Sean Murray
University of Delaware

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* April 2010	**2. REPORT TYPE** Final	**3. DATES COVERED (From - To)** June 2007–October 2009

4. TITLE AND SUBTITLE

Application of Social Network Analysis Techniques to Machine Translated Documents

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Ann E. M. Bornstein, John H. Brand, Michelle C. McVey,* and Sean Murray*

5d. PROJECT NUMBER

0TEDTC

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

U.S. Army Research Laboratory
ATTN: RDRL-CII-C
Aberdeen Proving Ground, MD 21005-5067

8. PERFORMING ORGANIZATION REPORT NUMBER

ARL-MR-741

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

*Students working under the auspices of the George Washington University Science and Engineering Apprentice Program-College Qualified Level, 1776 G St., Ste. 171, NW, Washington, DC 20052

14. ABSTRACT

This report documents and closes out the investigation into use of machine translation (MT) to provide material to generate actionable intelligence. A capability for automatic, real, or near real-time extraction of information on the local and external social networks will help produce actionable intelligence while the forces are at the site and able to act on the intelligence. Freeware tools exist to perform extraction of social network information from textual material. Using a social network analysis (SNA) toolset on textual material may provide the insights into social and, for irregular forces, command and action hierarchies while the area and its inhabitants are under friendly control. On-site exploitation of textual material surfaced during the SNA process may allow disruption of enemy cellular organization and logistic support assets and networks. SNA of this textual material thus constitutes an application of data mining in support of tactical operations.

The SNA and information exploitation enablers teams conducted an investigation to determine the degree to which MT preserves concept maps, including social network maps. A freeware SNA toolset was applied to a series of texts translated by various MT engines. The SNA tools were exercised on a set of MT documents provided by the U.S. Army Research Laboratory Multilingual Computing Research Branch and on documents acquired from the Internet. This report documents results of the investigation.

15. SUBJECT TERMS

social network analysis, machine translation, concept map, precision, recall

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Ann Bornstein
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	60	**19b. TELEPHONE NUMBER** (*Include area code*) 410-278-8947

Contents

List of Figures

List of Tables

Acknowledgments

Thanks to Dr. Michelle Vanni of the U.S. Army Research Laboratory Multilingual Computing Research Branch for the Spanish documents and translations.

INTENTIONALLY LEFT BLANK.

1. Introduction

This report documents and closes out an investigation into using social network analysis and textual network analysis (SNA/TNA) tools to extract social and other network information from machine translated documents. The focus of the investigation centered not on machine translation (MT) but on the utilization of a popular analysis suite on translated documents. Several machine translation engines were employed, with human translations used in parallel as baseline or ground truth documents. The report discusses benefits and problems encountered in the application of SNA/TNA tools to MT documents in both qualitative and quantitative terms.

A short outline of the utilization of MT in current operations is presented, followed by an outline of the functionality of the SNA/TNA tool set chosen for this investigation. The language data sets are then described. The results of the qualitative assessment of using the toolset on an ensemble of translated documents are presented, followed by a quantitative assessment of the information extracted from a different set of translations.

MT algorithms have been in use for many years, with wide use for military applications, including the Forward Area Language Converter (FALCON). FALCON has given Soldiers in the field in combat theaters the ability to translate or, more often, triage for later translation, documents found or captured during operations. One goal of applying SNA capability to MT texts is to generate actionable intelligence on command, support, and action networks. This information may be highly perishable and should be generated as soon as possible.

The material found and translated during operations can be extremely valuable; but a quick way of extracting conceptual or social relationships could permit the operations to be modified on the spot to identify and gain control of key persons, block attacks, capture materiel, and so on. Tactical intelligence in current operations is shifting to a human intelligence (HUMINT)-oriented process (*1, 2*). A key element in HUMINT operations against terrorist organizations is the determination and then use of the underlying social and organizational map of the actual terrorist organization and its support structure.

Although the social networks in terrorist organizations may well span a large area, the information that can be gathered at the local level is also increasingly important. For example, local social patterns are becoming more important in suicide attacks than in larger terrorist organizations spanning greater geographic areas (*3*). Increasingly, suicide attackers are being recruited on a personal, family, or social basis rather than by an organized recruitment process similar to a business.

This shift will make the gathering of strictly local information on social networks increasingly important in stability operations. Support of many small unit operations with a local focus may require MT and rapid analysis of MT documents. The recruitment and vetting of informants and the identification of influential locals require development of insight into organizational, family, and tribal affiliations. The response to tactical needs also demands unprecedented understanding of our enemies' social engines. SNA can provide insight into this social milieu. Using SNA with MT may also improve document exploitation.

The tacit assumption in applying SNA or TNA to MT documents is that the underlying map of concepts—personalities, entities, relationships between them, and so on—is preserved to an adequate degree under translation and in a form that can then be extracted adequately from the document by SNA/TNA software. Both these assumptions are reasonable; but because of the importance of the results to intelligence analyses and current operations, some investigation is warranted.

The adequacy of the MT engines and methods in preserving useful concept map information can be expressed in terms of accuracy and understanding entities. That is, a concept map may be preserved reasonably well under translation; but if too much of the map is lost, not enough information may be preserved to be useful. Likewise, if enough extraneous concepts are added in translation, the important relationships may be too hard to find to be useful. These cases are illustrated in figure 1.

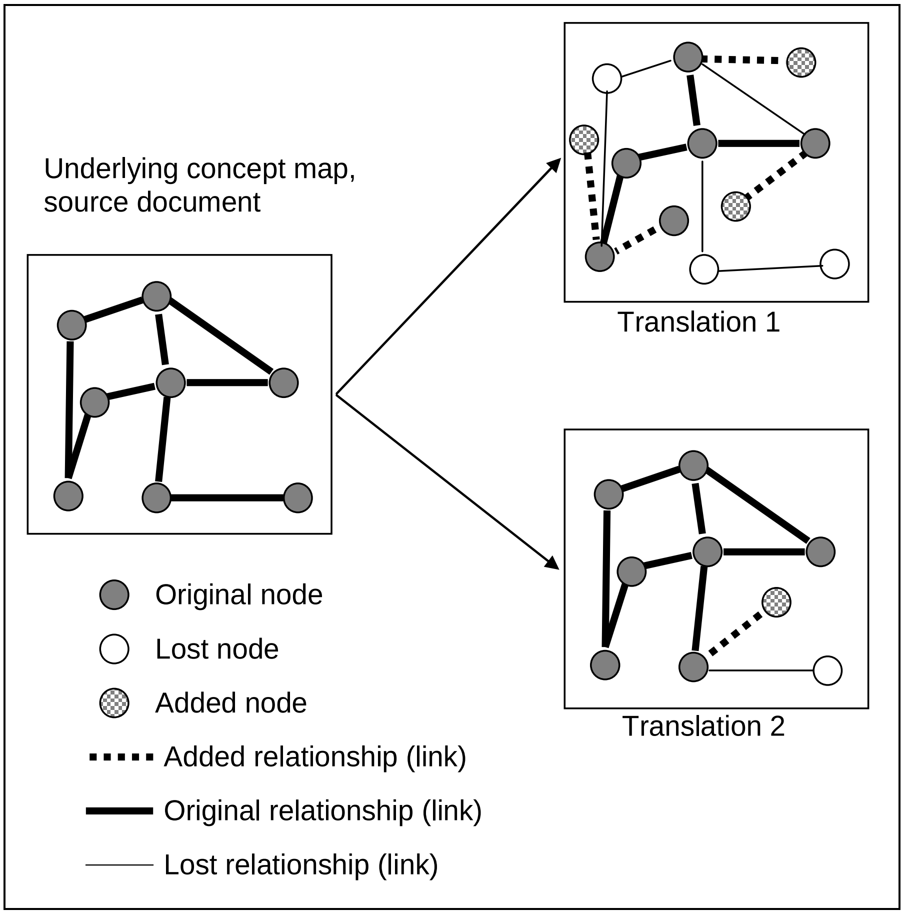


Figure 1. Concept map transformation under translation.

The original nodes and relationships that survive translation are shown for two notional translations as well as lost nodes and spurious added nodes. The question is, how much of these types of degradation is too much? Clearly, if a missing node is the name of a key terrorist leader, the loss is the link of a major terrorist figure to an entity such as a person, place, or event and is important; if the missing node is the name of his favorite TV show, less so.

In this study, investigations were conducted into (1) the structural aspects of the transformation of concept maps under translation and (2) the relational aspects of maps under translation. In the first case, several translations of the same documents were analyzed to determine qualitatively the proportions of signal to noise. In the second, the correlations between the concept maps resulting from translations were investigated. These investigations were exploratory in nature and will require confirmation using larger data sets.

2. SNA Toolset

The SNA software suite used to extract concepts from textual material and analyze their relationships was developed by the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University. The suite includes AutoMap, a freeware social/text network analysis tool, and several other visualization and analysis tools (*4*). The principal visualization tool used in this investigation was the Organizational Risk Analyzer (ORA).

2.1 Text Preprocessing With AutoMap

AutoMap is "a software tool to analyze text using the method of network text analysis. It performs a specific type of network text analysis called semantic network analysis. Semantic analysis extracts and analyzes links among words to model an author's 'mental map' as a network of links" (*5*). AutoMap links concepts together based on the concepts' proximity to each other. That is, two adjacent concepts are listed as linked; two concepts separated by several intervening words are not. Two words linked together are called a bigram. The bigrams can be analyzed for frequency within the document.

One measure of the degree to which translations differ is the difference in concepts and concept pairs introduced or deleted by the translation means. In this study, the Human 1 (HU01) translation of document 6 was chosen arbitrarily as the baseline, or ground truth; the Human 2 (HU02) translation could have been used as the baseline equally well.

The raw collection of pair-wise combinations of all possible concepts in the source document will obscure the meaning or presence of relationships between important concepts. Preprocessing removes confusing material, unifies multiword concepts into single strings representing a single concept, and performs other functions on the basic text file or group of files under analysis. The key preprocessing tools are as follows:

- Named entity recognition
- Collocation/n-gram identification
- Deletion
- Thesauri
- Stemming functions

2.1.1 Named-Entity Recognition

Named entity recognition does not impact the overall data; however, this function does allow the user to retrieve proper names,[*] numerical figures, and abbreviations from texts (*6*). More specifically, this function detects single capitalized words,[†] adjacent capitalized words,[‡] and strings of adjacent capitalized words[§] (*5*).

2.1.2 Collocation/N-gram Identification

The n-gram feature is much like named entity recognition because it does not directly impact the data. This utility identifies a word's collocates i.e., those words appearing next to or near the word (*6*).

2.1.3 Deletion

The delete list removes noncontent bearing words, such as articles and conjunctions, from the text. A problem is that, if used too extensively, the delete list can remove or obscure meaning. For example, if modifiers such as "liquid" are removed, associations with the concept "liquid explosive" are also removed. Delete lists tailored to specific physical areas or subjects of interest may well be necessary, perhaps applied to the same set of documents with separate and sequential analyses performed to address different physical areas or subjects of interest.

An example of the effect of a delete list would be as follows:

Original input text:	The eldest of these, and Bilbo's favourite, was young Frodo Baggins.
Delete list entries:	and, eldest, of, the, these, was
Text after deletion:	Bilbo's favourite young Frodo Baggins

[*] The term "proper names" refers to names of people, places, and/or organizations.

[†] For example, Maryland.

[‡] Adjacent capitalized words are phrases like "Towson University."

[§] "University of Maryland Baltimore College" is an example of a string of capitalized words.

2.1.4 Thesauri

Two different thesauri found within AutoMap were employed. The first was the Generalization Thesaurus, which is a paired set that associates textual "concepts" (words or other symbol groupings, such as a number, are referred to as "concepts" in Automap) in a text with other concepts. This way, concepts that differ in spelling can be associated with a higher level abstraction. For instance, the string Inkatha will be treated as different from the string Zulu party Inkatha. The latter string is treated as three separate concepts unless the parts of the string are unified through the Generalization Thesaurus. Thus, both strings can be associated with the general concept or idea of Inkatha. Likewise, variant spellings can be unified into a general concept. This is especially important if the source document is written in a non-Roman alphabet such as Arabic. This way, all occurrences of the variant strings will be shown and treated as the same generalized concept.

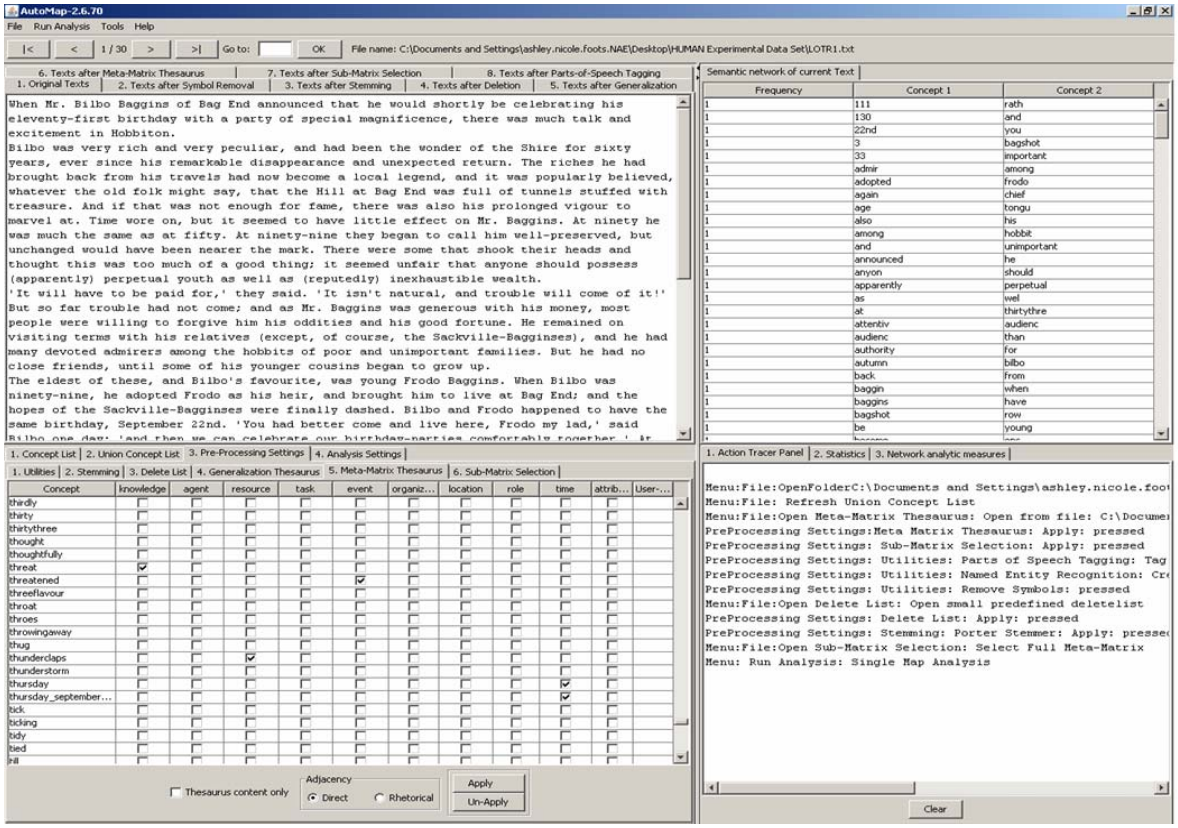
The second and, perhaps, more important thesaurus was the Meta-Matrix Thesaurus. This thesaurus associates text-level concepts with meta-matrix categories including agent, attribute, event, knowledge, location, organization, resource, role, task, and time. For example, the concept "Inkatha" would be tagged under the category "organization."[*] Using the Meta-Matrix Thesaurus in tagging concepts by category according to function in the document (e.g., "threatened" is an "event") is illustrated in figure 2.

2.1.5 Stemming

Stemming can, arguably, also act to bring translations to a common basis (*5*). Stemming is the process of reducing variants of a word to the common "stem." That is, "digs," "digging," "dug," and "dig" are reduced to "dig." This way, a great amount of information is lost, although the relationships are preserved. The modification of meaning may be undesirable or even intolerable. One example of stemming's modification of meaning is the conversion of "tanks" to "tank," "soldiers" to "soldier," etc. In many contexts, the loss of specific meaning (plural vs. singular, possessive case vs. nominative case, etc.) through the use of stemming can be a dangerous oversimplification. Whether to accept the simplification is a judgment call by the analyst.

In view of the possibility of removing or obscuring important concepts and relationships, iteration is the best method when beginning to develop the capability to analyze a given corpus. For instance, removal of concepts can be done starting from a limited list of words to be removed. Then, check the processed text for words that do not add to meaning as they are also candidates for removal. The candidates for removal may be put in the delete list and the list removed from the text before the results are scrutinized again. In field use, there may not be time for an iterative approach, so prior preparation and establishment of preprocessing files, such

[*] The Meta-Matrix Thesaurus function was also used to enhance the visual display of concept relationships by tagging concepts under different categories according to factors such as inclusion in a specific document. This allowed the display to depict the concept in contrasting colors based on provenance, as in section 5. In this case, the categories associated with a given concept did not refer to the meaning or function of the concept, but to factors such as what document the concepts were found in.



AutoMap-2.6.70
File Run Analysis Tools Help
|< < 1 / 30 > >| Go to: OK
6. Texts after Meta-Matrix Thesaurus | 7. Texts after Sub-Matrix Selection | 8. Texts after Parts-of-Speech Tagging
1. Original Texts | 2. Texts after Symbol Removal | 3. Texts after Stemming | 4. Texts after Deletion | 5. Texts after Generalization
Semantic network of current Text
1. Concept List | 2. Union Concept List | 3. Pre-Processing Settings | 4. Analysis Settings
1. Utilities | 2. Stemming | 3. Delete List | 4. Generalization Thesaurus | 5. Meta-Matrix Thesaurus | 6. Sub-Matrix Selection
1. Action Tracer Panel | 2. Statistics | 3. Network analytic measures
Apply
Un-Apply
Clear

File name: C:\Documents and Settings\ashley.nicole.foots.NAE\Desktop\HUMAN Experimental Data Set\LOTR1.txt

```
When Mr. Bilbo Baggins of Bag End announced that he would shortly be celebrating his
eleventy-first birthday with a party of special magnificence, there was much talk and
excitement in Hobbiton.
Bilbo was very rich and very peculiar, and had been the wonder of the Shire for sixty
years, ever since his remarkable disappearance and unexpected return. The riches he had
brought back from his travels had now become a local legend, and it was popularly believed,
whatever the old folk might say, that the Hill at Bag End was full of tunnels stuffed with
treasure. And if that was not enough for fame, there was also his prolonged vigour to
marvel at. Time wore on, but it seemed to have little effect on Mr. Baggins. At ninety he
was much the same as at fifty. At ninety-nine they began to call him well-preserved, but
unchanged would have been nearer the mark. There were some that shook their heads and
thought this was too much of a good thing; it seemed unfair that anyone should possess
(apparently) perpetual youth as well as (reputedly) inexhaustible wealth.
'It will have to be paid for,' they said. 'It isn't natural, and trouble will come of it!'
But so far trouble had not come; and as Mr. Baggins was generous with his money, most
people were willing to forgive him his oddities and his good fortune. He remained on
visiting terms with his relatives (except, of course, the Sackville-Bagginses), and he had
many devoted admirers among the hobbits of poor and unimportant families. But he had no
close friends, until some of his younger cousins began to grow up.
The eldest of these, and Bilbo's favourite, was young Frodo Baggins. When Bilbo was
ninety-nine, he adopted Frodo as his heir, and brought him to live at Bag End; and the
hopes of the Sackville-Bagginses were finally dashed. Bilbo and Frodo happened to have the
same birthday, September 22nd. 'You had better come and live here, Frodo my lad,' said
Bilbo one day: 'and then we can celebrate our birthday-parties comfortably together.' At
```

Frequency	Concept 1	Concept 2
1	111	rath
1	130	and
1	22nd	you
1	3	bagshot
1	33	important
1	admir	among
1	adopted	frodo
1	again	chief
1	age	tongu
1	also	his
1	among	hobbit
1	and	unimportant
1	announced	he
1	anyon	should
1	apparently	perpetual
1	as	wel
1	at	thirtythre
1	attentiv	audienc
1	audienc	than
1	authority	for
1	autumn	bilbo
1	back	from
1	baggin	when
1	baggins	have
1	bagshot	row
1	be	young

Concept	knowledge	agent	resource	task	event	organiz...	location	role	time	attrib...	User-...
thirdly	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
thirty	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
thirtythree	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
thought	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
thoughtfully	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
threat	☑	☐	☐	☐	☐	☐	☐	☐	☐	☐	
threatened	☐	☐	☐	☐	☑	☐	☐	☐	☐	☐	
threeflavour	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
throat	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
throes	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
throwingaway	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
thug	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
thunderclaps	☐	☐	☑	☐	☐	☐	☐	☐	☐	☐	
thunderstorm	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
thursday	☐	☐	☐	☐	☐	☐	☐	☐	☑	☐	
thursday_september...	☐	☐	☐	☐	☐	☐	☐	☐	☑	☐	
tick	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
ticking	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
tidy	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
tied	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
till	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

☐ Thesaurus content only

Adjacency: ● Direct ○ Rhetorical

```
Menu:File:OpenFolderC:\Documents and Settings\ashley.nicole.foo
Menu:File: Refresh Union Concept List
Menu:File:Open Meta-Matrix Thesaurus: Open from file: C:\Docume
PreProcessing Settings:Meta Matrix Thesaurus: Apply: pressed
PreProcessing Settings: Sub-Matrix Selection: Apply: pressed
PreProcessing Settings: Utilities: Parts of Speech Tagging: Tag
PreProcessing Settings: Utilities: Named Entity Recognition: Cr
PreProcessing Settings: Utilities: Remove Symbols: pressed
Menu:File:Open Delete List: Open small predefined deletelist
PreProcessing Settings: Delete List: Apply: pressed
PreProcessing Settings: Stemming: Porter Stemmer: Apply: presse
Menu:File:Open Sub-Matrix Selection: Select Full Meta-Matrix
Menu: Run Analysis: Single Map Analysis
```

Figure 2. Use of the Meta-Matrix Thesaurus in the Lord of the Rings excerpt, document set B.

as a relevant delete list, may be the difference between useful, actionable intelligence and a fallacious indication of no useful data.

The analyst may wish to have a set of situation-specific preprocessing input files, including delete lists and thesauri available for different missions. For instance, a unit may ask for all intelligence about a given address; the request regarding the address and people, places, and things of interest related to that address may be entered into the initial version of a Generalization Thesaurus. Likewise, words that might be included in a delete list used for one purpose or locale might be of crucial importance for another and so should be retained in the text.

2.2 ORA

ORA provides a graphical representation of the network or textual input (e.g., bigrams). For a detailed description of the display options for ORA, see the User's Guide and software description (*7*, *8*). Briefly, ORA organizes the display of the concepts and relationships in either operator-selected or automatic default patterns. One pattern, a dendritic tree, is especially valuable for examining differences between two or more representations of what is nominally the same material.

3. Language Data Sets

To explore the functionality of AutoMap, several trial documents, which were grouped into sets A and B, were chosen. Document set A was used for the qualitative analyses reported in sections 4 and 5; document set B was used for the quantitative analyses reported in section 6. The emphasis in set A was the multiplicity of MT engines; the emphasis in set B was the multiplicity of documents translated by one MT engine. Human translations were used in each set as ground truth.

3.1 Set A

The U.S. Army Research Laboratory Multilingual Computing Research Branch provided three sets of translations of short source excerpts in Spanish, treated here as document sets 6, 7, and 8. Each source excerpt is accompanied by two human linguist translations and five translations by different MT engines. These are labeled within each document set HU01 and HU02 for the human translations and MT01–05 for the machine translations. Short excerpts were used to allow direct examination of the documents at each stage of processing and analysis. Direct examination at each stage allowed the researchers to gain some confidence in the software and detect unexpected events that would easily be lost with larger files. Set A was extracted from the 1994 Defense Advanced Research Projects Agency corpus used to evaluate Spanish MT systems. The source and translations for one document in the family, document 6, are given in appendix A.

3.2 Set B

A set of longer translations was obtained to analyze the precision, recall, and accuracy between maps produced by different translation methods. Spanish texts were chosen for this analysis because MT engines have been vigorously trained with this language, which should lead to more reliable translations. Thirty Spanish texts of varying document lengths and three genres representative of different styles were collected: six short stories, six love letters, and 18 pages from J. R. R. Tolkien's Lord of the Rings (*9*). Each Spanish document had a respective human translation (or, in the case of the Tolkien excerpt, the original English document) that served as its parallel text, or ground truth, in English. The sources and translation for one excerpt are given in appendix B.

4. Descriptive Statistical Measures for Qualitative Assessment

AutoMap produces a number of descriptive statistical measures that, unfortunately, are not very informative in terms of differentiating between information content of files. The measures include:

- The number of concepts analyzed, both unique and total.
- The number of concepts in statements, both unique and total.
- The number of statements, both unique and total.
- The density, based on statements, both unique and total.

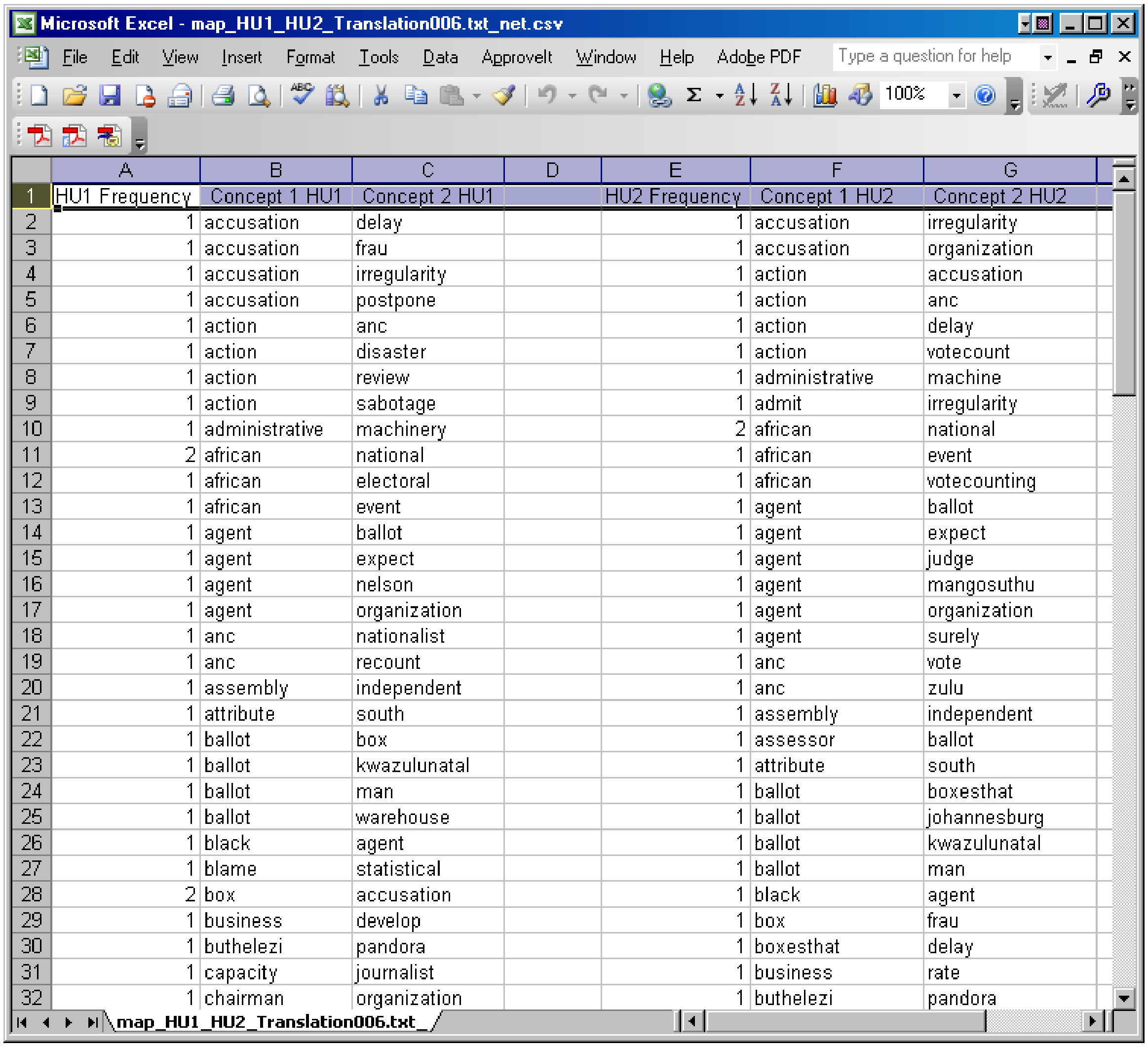
Other statistical measures may be more helpful in estimating the differences between documents. For instance, two documents will share a common set of concepts that were translated the same way, and each will also have its own set of unique concepts. These unique concepts can be different translations of the same word or nontranslated words. In each case, the enumeration of unique concepts may be a useful measure of the quantity of information lost (for example, words not translated) and, perhaps, of the noise introduced by a given translation engine; however, the actual relationships between concepts are an index of meaning. Those relationships are, in part, reflected in the list of bigrams, but complex relationships are difficult to discern from the textual list of bigrams. Interpretation of relationships through inspection of the diagrammatic representations of the bigrams is greatly facilitated by visualization techniques, such as ORA.

ORA can be used to generate a concept map and data associated with that map. Some of the bigrams in a text file (not included in the documents analyzed) are shown in the upper right pane in figure 2.* Note the frequency of occurrence listed for each bigram. This document set requires further preprocessing to clean up the output—unimportant numbers and less important words need to be removed to avoid obscuring the relations that are important. Important numbers that modify an important concept need to be linked by a thesaurus entry with that concept. Standalone important numbers (such as 9/11) need to be modified so a wholesale deletion of numbers does not affect them and unimportant numbers deleted one by one through inclusion in a delete list.

The list of bigrams may be used to estimate whether two documents differ in a significant fashion. Presently, this must be by judgment. That is, if one document is taken as ground truth, the presence or absence of concepts judged by the analyst as important may be tracked. The pairing of concepts may also be analyzed in the same way. Two important concepts may be translated differently in two documents, or the concepts may be linked or related differently in the two translations. The relationship in both translated documents may be different from the source document as well as being different with respect to each other. If one translation is chosen as the standard or ground truth, the presence or absence in the translated documents of concepts present in the ground truth document is a useful metric. The presence or absence in the translated documents of relationships between concepts present in the ground truth documents is also useful.

*Note that the list consists of linked “concepts,” where a “concept” can be a number or even a jumble of symbols. The bigram is the linkage of the two concepts.

A weight may be assigned to the existence of a given concept pair; at present, this must be a judgment call. A list of concept pairs from two translations, HU01 and HU02, is shown in figure 3. (A list of all concepts for all translations of document excerpt 6 is provided in appendix C.)

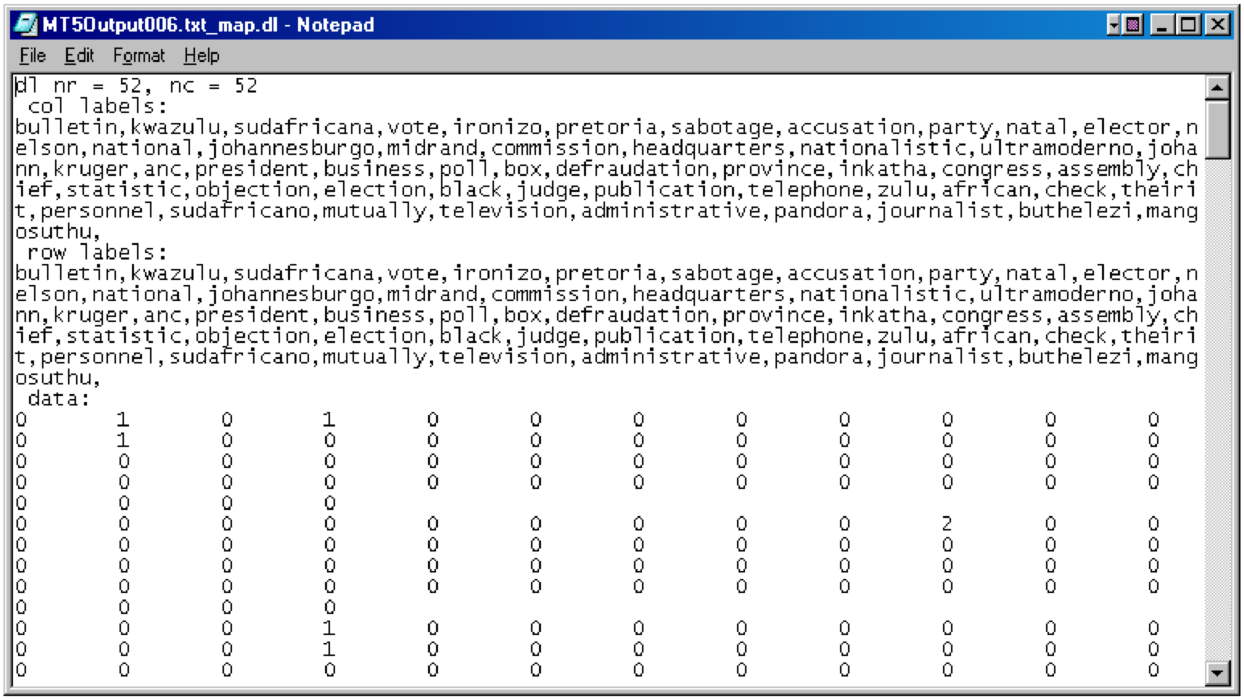


	A	B	C	D	E	F	G
1	HU1 Frequency	Concept 1 HU1	Concept 2 HU1		HU2 Frequency	Concept 1 HU2	Concept 2 HU2
2	1	accusation	delay		1	accusation	irregularity
3	1	accusation	frau		1	accusation	organization
4	1	accusation	irregularity		1	action	accusation
5	1	accusation	postpone		1	action	anc
6	1	action	anc		1	action	delay
7	1	action	disaster		1	action	votecount
8	1	action	review		1	administrative	machine
9	1	action	sabotage		1	admit	irregularity
10	1	administrative	machinery		2	african	national
11	2	african	national		1	african	event
12	1	african	electoral		1	african	votecounting
13	1	african	event		1	agent	ballot
14	1	agent	ballot		1	agent	expect
15	1	agent	expect		1	agent	judge
16	1	agent	nelson		1	agent	mangosuthu
17	1	agent	organization		1	agent	organization
18	1	anc	nationalist		1	agent	surely
19	1	anc	recount		1	anc	vote
20	1	assembly	independent		1	anc	zulu
21	1	attribute	south		1	assembly	independent
22	1	ballot	box		1	assessor	ballot
23	1	ballot	kwazulunatal		1	attribute	south
24	1	ballot	man		1	ballot	boxesthat
25	1	ballot	warehouse		1	ballot	johannesburg
26	1	black	agent		1	ballot	kwazulunatal
27	1	blame	statistical		1	ballot	man
28	2	box	accusation		1	black	agent
29	1	business	develop		1	box	frau
30	1	buthelezi	pandora		1	boxesthat	delay
31	1	capacity	journalist		1	business	rate
32	1	chairman	organization		1	buthelezi	pandora

Figure 3. Comparison of lists of bigrams, HU01 and HU02, document 6.

Concept maps may be represented numerically or graphically. Matrices provide convenient means of displaying numerical representations of the relations between concepts. An example of a matrix representation of the concept map for MT05, document 6, is shown in figure 4. The frequency with which one concept is paired with others is given by the matrix elements.

Figure 4 is a screen shot of part of the matrix of relationships between concepts. Because the columns and rows are not labeled individually and the matrix rows are folded multiple times, the matrix is difficult to interpret. The matrix has both columns and rows corresponding to the concepts in the document, enumerated as “col labels” and “row labels.” The matrix represented



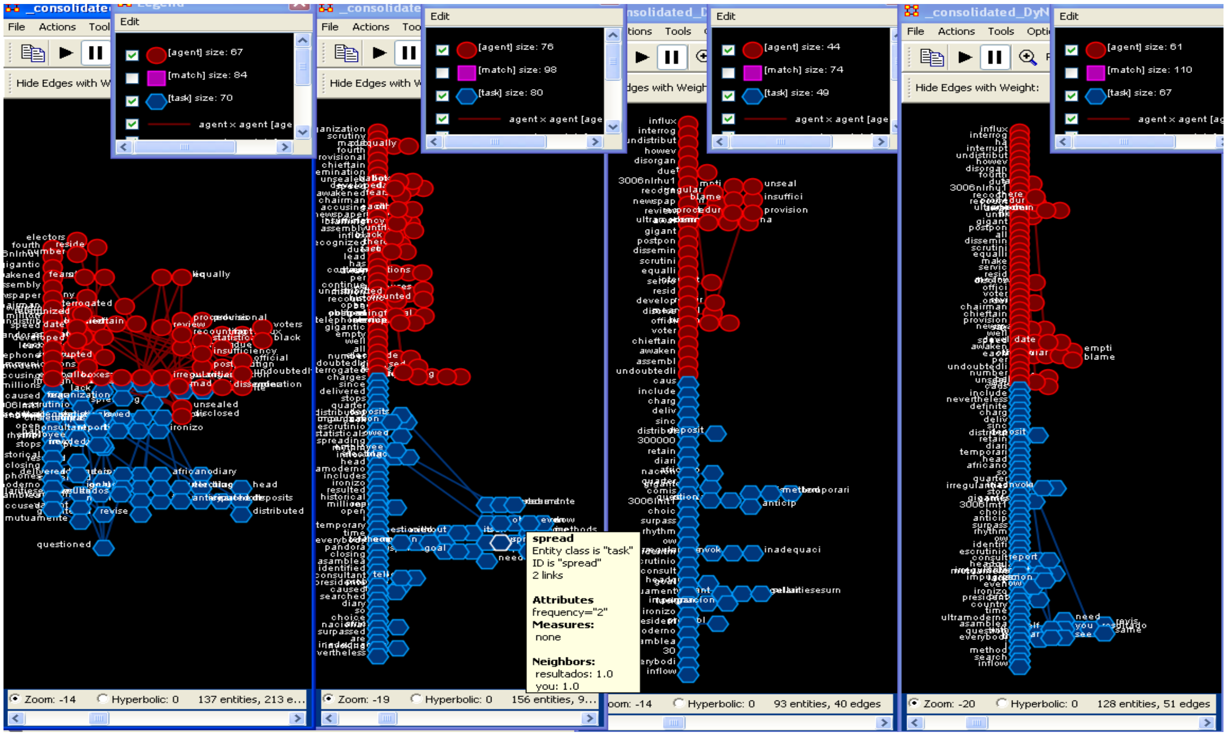
```
dl nr = 52, nc = 52
 col labels:
bulletin,kwazulu,sudafricana,vote,ironizo,pretoria,sabotage,accusation,party,natal,elector,n
elson,national,johannesburgo,midrand,commission,headquarters,nationalistic,ultramoderno,joha
nn,kruger,anc,president,business,poll,box,defraudation,province,inkatha,congress,assembly,ch
ief,statistic,objection,election,black,judge,publication,telephone,zulu,african,check,theiri
t,personnel,sudafricano,mutually,television,administrative,pandora,journalist,buthelezi,mang
osuthu,
 row labels:
bulletin,kwazulu,sudafricana,vote,ironizo,pretoria,sabotage,accusation,party,natal,elector,n
elson,national,johannesburgo,midrand,commission,headquarters,nationalistic,ultramoderno,joha
nn,kruger,anc,president,business,poll,box,defraudation,province,inkatha,congress,assembly,ch
ief,statistic,objection,election,black,judge,publication,telephone,zulu,african,check,theiri
t,personnel,sudafricano,mutually,television,administrative,pandora,journalist,buthelezi,mang
osuthu,
 data:
0       1       0       1       0       0       0       0       0       0       0       0
0       1       0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0       0       0
0       0       0       0
0       0       0       0       0       0       0       0       0       2       0       0
0       0       0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0       0       0
0       0       0       0
0       0       0       1       0       0       0       0       0       0       0       0
0       0       0       1       0       0       0       0       0       0       0       0
0       0       0       0       0       0       0       0       0       0       0       0
```

Figure 4. Concept map in matrix form.

in the figure has 52 rows and 52 columns, and each row is folded, for the window size chosen, five times. Thus, in the five text lines representing the first row of the matrix, 1s appear that correspond to the 2nd, 4th, and 14th columns, which, in turn, represent single occurrences of the association of the concept bulletin and the concepts kwazulu, vote, and johannesburgo. The next row has a double occurrence of the association of kwazulu and natal. Note that the relationships are directional; that is, row label bulletin and column label kwazulu show a single occurrence of bulletin linked to kwazulu, but row kwazulu and column bulletin do not show a corresponding occurrence of kwazulu linked to bulletin. (Graphical representations of concept maps may be seen in figures 5–9.)

Once the files have been processed and analyzed, the resulting lists of concepts may be compared. This is done using the text set comparer tool. The text set comparer tool allows an estimate of the degree to which key concepts, per se, are translated, and the degree the meaning is obscured by the generation of concepts not present in the source document. The comparison highlights terms common to two documents as well as those found in one but not the other. Thus, if important concepts are common to both documents, the datum will be noted; if new concepts are found in one document but not in the other, one or both translations may be faulty. For instance, if one document is a ground truth document and the other a machine translation, a concept found in the translation but not the ground truth document was probably generated erroneously. These spurious concepts may obscure important relationships and may also add false relationships as well.

Using the dictionary comparison function of the text set comparer tool for all the files permits estimation of the noise added and meaning subtracted during translation, compared to the a priori ground truth, HU01. The measures of noise added and meaning lost may be considered as candidate measures subject to further study given additional translations (human and machine).



[agent] size: 67
[match] size: 84
[task] size: 70

Zoom: -14 ○ Hyperbolic: 0 137 entities, 213 e...

[agent] size: 76
[match] size: 98
[task] size: 80

Zoom: -19 ○ Hyperbolic: 0 156 entities, 9...

spread
Entity class is "task"
ID is "spread"
2 links

Attributes
frequency="2"
Measures:
none

Neighbors:
resultados: 1.0
you: 1.0

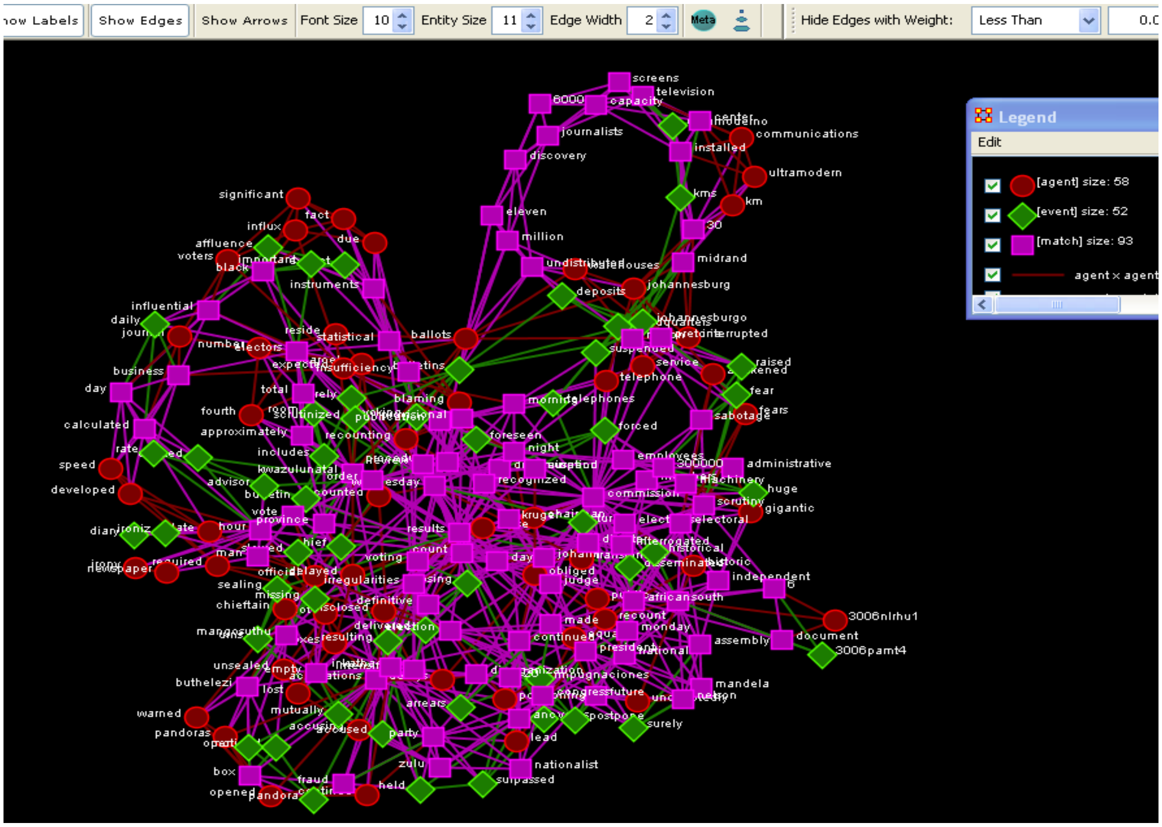
[agent] size: 44
[match] size: 74
[task] size: 49

oom: -14 ○ Hyperbolic: 0 93 entities, 40 edges

[agent] size: 61
[match] size: 110
[task] size: 67

Zoom: -20 ○ Hyperbolic: 0 128 entities, 51 edges

Figure 5. Four concept maps of the differences between HU01 and MT01, document 6 excerpt, with different preprocessing paths. Concepts in HU01, but not MT01, are shown as ovals; concepts in MT01, but not HU01, are shown as hexagons.



how Labels | Show Edges | Show Arrows | Font Size 10 | Entity Size 11 | Edge Width 2 | Hide Edges with Weight: Less Than 0.0

Legend

Edit

☑ [agent] size: 58

☑ [event] size: 52

☑ [match] size: 93

☑ agent x agent

Figure 6. A map of concepts in HU01, document 6 , merged with MT01, document 6. Concepts unique to HU01 are shown as ovals, concepts unique to MT01 are shown as diamonds, and concepts common to both are shown as squares.

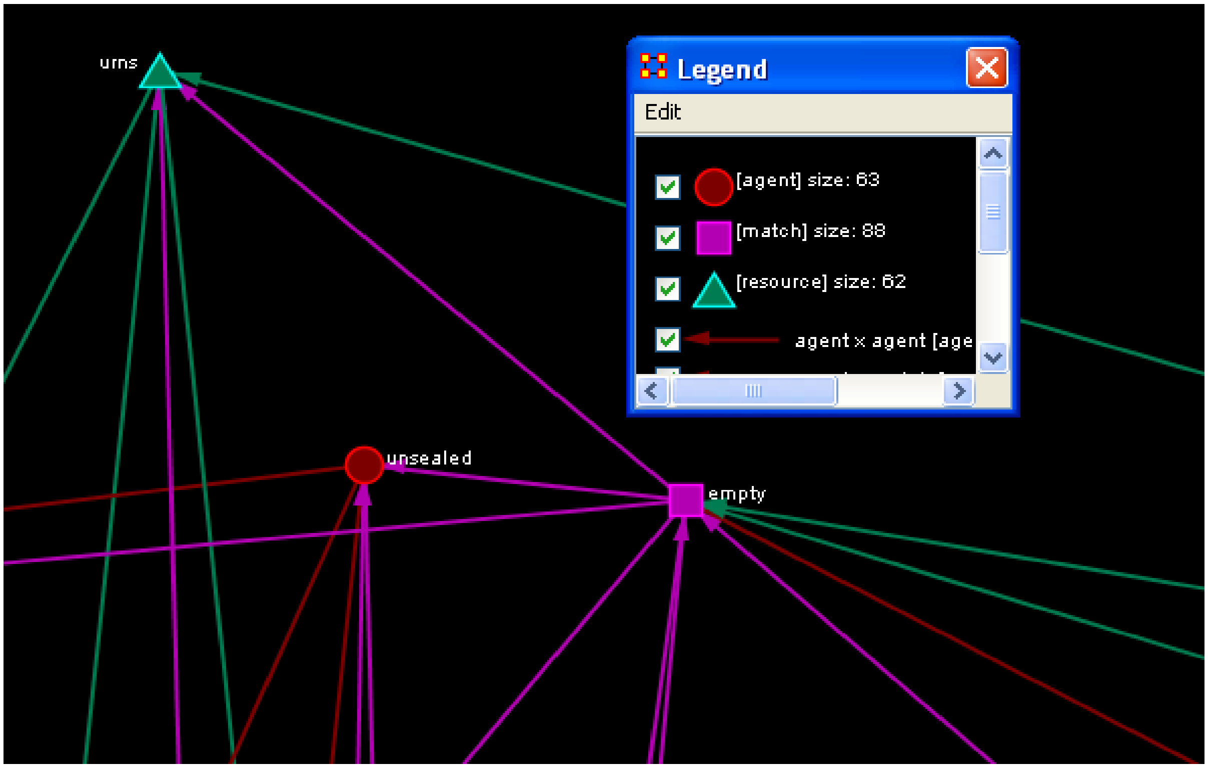
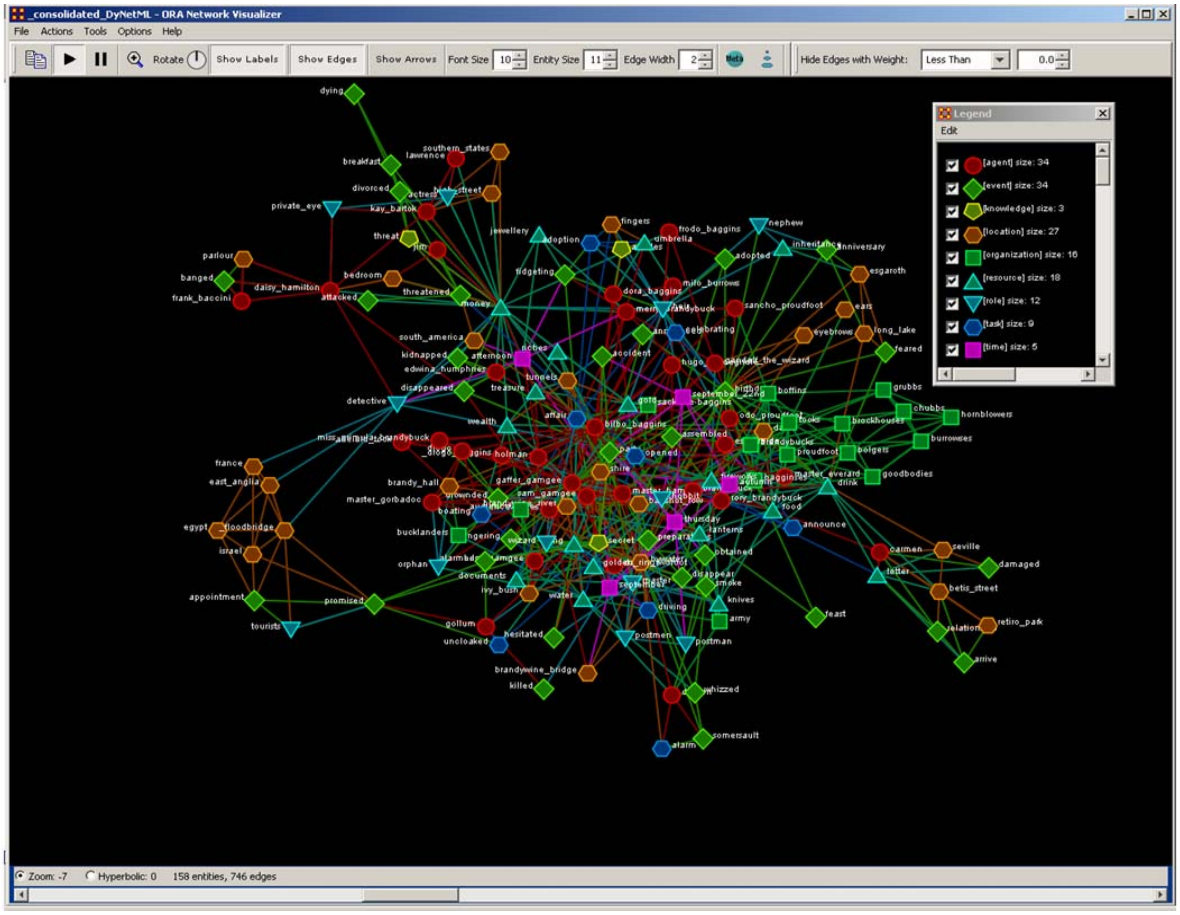


Figure 7. Intersection points in the documents, between HU01 and MT05, document 6. There are three different examples of the intersection points.



_consolidated_DyNetML - ORA Network Visualizer

Font Size 10 Entity Size 11 Edge Width 2 Hide Edges with Weight: Less Than 0.0

☑ ● [agent] size: 34

☑ ◆ [event] size: 34

☑ ⬟ [knowledge] size: 3

☑ ⬢ [location] size: 27

☑ ■ [organization] size: 16

☑ ▲ [resource] size: 18

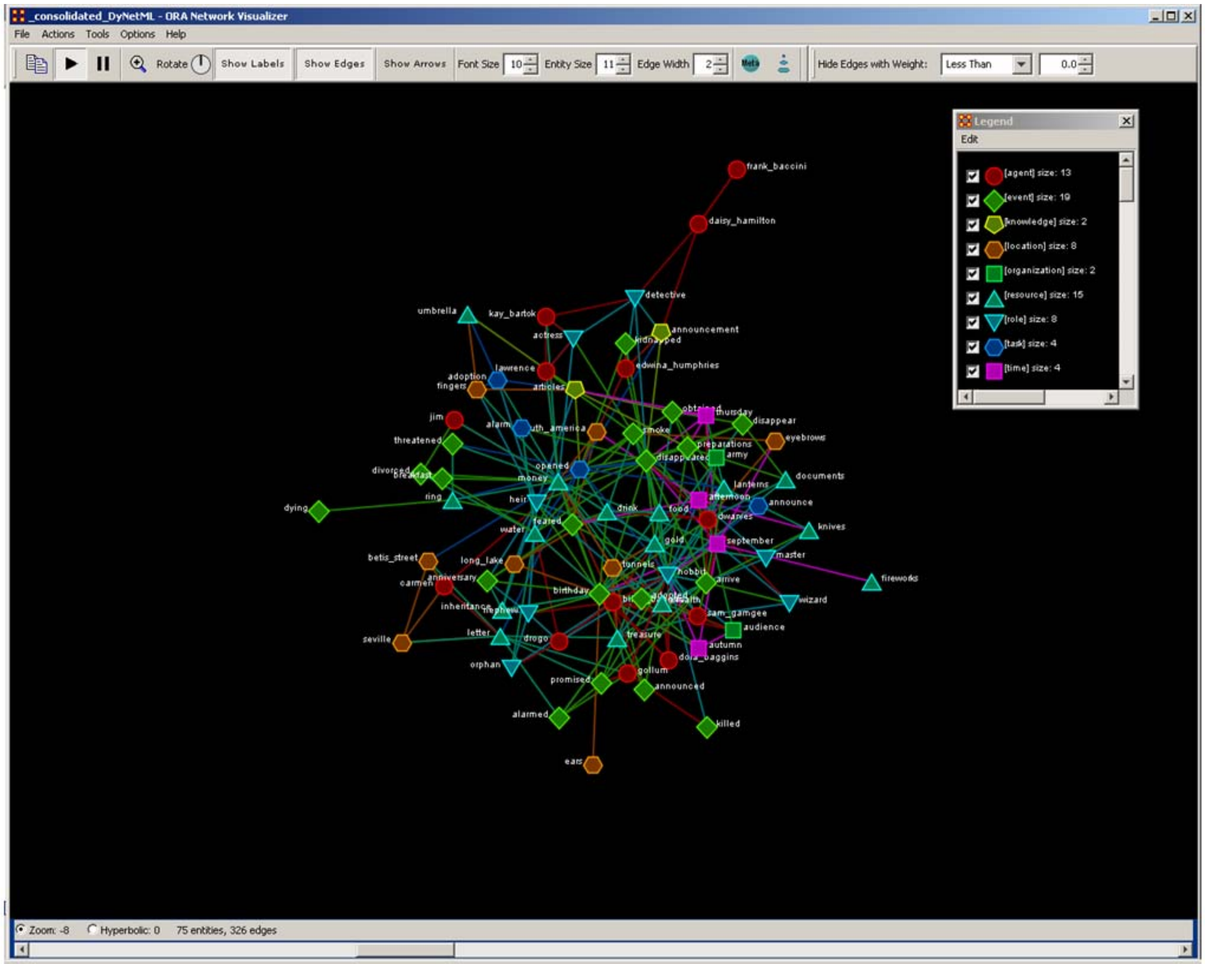
☑ ▼ [role] size: 12

☑ ⬢ [task] size: 9

☑ ■ [time] size: 5

● Zoom: -7 ○ Hyperbolic: 0 158 entities, 746 edges

Figure 8. ORA map of document set B human translations, all files included.



_consolidated_DyNetML - ORA Network Visualizer

Font Size 10 Entity Size 11 Edge Width 2

Hide Edges with Weight: Less Than 0.0

Legend

☑ ● [agent] size: 13
☑ ◆ [event] size: 19
☑ ⬟ [knowledge] size: 2
☑ ⬢ [location] size: 8
☑ ■ [organization] size: 2
☑ ▲ [resource] size: 15
☑ ▼ [role] size: 8
☑ ⬢ [task] size: 4
☑ ■ [time] size: 4

○ Zoom: -8 ○ Hyperbolic: 0 75 entities, 326 edges

Figure 9. ORA map of document set B machine translations, corresponding to those of figure 8.

A simple and direct partial measure of loss of meaning may be gained by consideration of the number of non-English-dictionary words in the translations compared to a reference. Mistranslations and nontranslations may show up in the list of non-English-dictionary words, along with place and personal names not found in the reference dictionary file. A presumed good translation will have the person and place names faithfully rendered; deviation from that may be considered a very rough measure of degradation compared to the ground truth.

Statistical measures describing the documents are taken from the AutoMap output files. The results are shown in table 1.

A full list of the concepts and nondictionary terms in the several translations of the document 6 excerpt is presented in appendix C along with various counts of concepts generated, lost, etc. A summary of the concept list properties of most interest is presented in table 2.

Table 1. Some basic descriptive statistical measures derived from AutoMap analysis of translations HU01, HU02, and MT01–05.

Descriptive Statistical Measure	HU01	HU02	MT01	MT02	MT03	MT04	MT05
Word count (MS Word 2003 tool) of unprocessed document, less header, including title	348	334	374	374	357	381	370
Paragraph count of unprocessed document, less header, less title	4	9	7	8	7	8	6
Difference in total word count of unprocessed document, with respect to HU01	—	–14	26	26	9	33	22
Difference in total paragraphs in unprocessed document, with respect to HU01	—	5	3	4	3	4	2
Total unique concepts, after preprocessing	55	71	85	69	86	68	62
Total concepts after preprocessing	78	89	106	89	102	88	79
Difference in unique concepts after preprocessing, with respect to HU01	—	16	30	14	31	13	7
Difference in total concepts after preprocessing, with respect to HU01	—	11	28	11	24	10	1
Nondictionary concepts after preprocessing, in common with HU01	10	10	8	8	6	8	7
Nondictionary concepts after preprocessing, in addition to those in HU01	—	1	16	8	10	5	9
Number of unique concepts in statements in preprocessed document	77	88	105	87	101	86	78
Total number of concepts in statements in preprocessed document	77	88	105	88	101	87	78
Number of unique statements in preprocessed document	77	88	105	87	101	86	78
Total number of statements in preprocessed document	77	88	105	88	101	87	78
Density of unique concepts in preprocessed document	1.4	1.24	1.24	1.26	1.17	1.26	1.26
Density of total concepts in preprocessed document	1.4	1.24	1.24	1.28	1.17	1.28	1.26
Entropy of preprocessed document (rounded)	5.59	6.04	6.27	5.95	6.34	5.93	5.81

Table 2. Summary of concept list properties, HU01-MT05, document 6 excerpt.

total unique concepts HU01 after deletion, generalization: 55	total unique concepts HU02 after deletion, generalization: 71	total unique concepts MT01 after deletion, generalization: 85	total unique concepts MT02 after deletion, generalization: 69	total unique concepts MT03 after deletion, generalization: 86	total unique concepts MT04 after deletion, generalization: 68	total unique concepts MT05 after deletion, generalization: 62
total concepts after deletion, generalization, HU01: 78	total concepts HU02 after deletion, generalization: 89	total concepts MT01 after deletion, generalization: 106	total concepts MT02 after deletion, generalization: 89	total concepts MT03 after deletion, generalization: 102	total concepts MT04 after deletion, generalization: 88	total concepts MT05 after deletion, generalization: 79
---	no. of unique concepts HU02 wrt HU01 after deletion, generalization: 16	no. of unique concepts MT01 wrt HU01 after deletion, generalization: 30	no. of unique concepts MT02 wrt HU01 after deletion, generalization:14	no. of unique concepts MT03 wrt HU01 after deletion, generalization: 31	no. of unique concepts MT04 wrt HU01 after deletion, generalization: 13	no. of unique concepts MT05 wrt HU01 after deletion, generalization: 7
---	Difference in total concepts HU02 wrt HU01 after deletion, generalization: 11	Difference in total concepts MT01 wrt HU01 after deletion, generalization: 28	Difference in total concepts MT02 wrt HU01 after deletion, generalization: 11	Difference in total concepts MT03 wrt HU01 after deletion, generalization: 24	Difference in total concepts MT04 wrt HU01 after deletion, generalization: 10	Difference in total concepts MT05 wrt HU01 after deletion, generalization: 1
10 concepts in HU01 listed as non-dictionary	non-dictionary concepts in HU02 after deletion, generalization in common with HU01: 10	non-dictionary concepts in MT01 after deletion, generalization in common with HU01: 8	non-dictionary concepts in MT02 after deletion, generalization in common with HU01:8	non-dictionary concepts in MT03 after deletion, generalization in common with HU01: 6	non-dictionary concepts in MT04 after deletion, generalization in common with HU01: 8	non-dictionary concepts in MT05 after deletion, generalization in common with HU01: 7
---	non-dictionary concepts in addition to those in HU01: 1	non-dictionary concepts in addition to those in HU01: 16	non-dictionary concepts in addition to those in HU01: 8	non-dictionary concepts in addition to those in HU01: 10	non-dictionary concepts in addition to those in HU01: 5	non-dictionary concepts in addition to those in HU01: 9
HU01 word count, less header, incl. title: 348	HU02 word count, less header, incl. title: 334	MT01 word count, less header, incl. title: 374	MT02 word count, less header, incl. title: 374	MT03 word count, less header, incl. title: 357	MT04 word count, less header, incl. title: 381	MT05 word count, less header, incl. title: 370
HU01 paragraph count, less header, less title: 4	HU02 paragraph count, less header, less title: 9	MT01 paragraph count, less header, less. title: 7	MT02 paragraph count, less header, less. title: 8	MT03 paragraph count, less header, less title: 7	MT04 paragraph count, less header, less title: 8	MT05 paragraph count, less header, less title: 6
---	Difference, total words HU02 wrt HU01: -14	Difference in total words MT01 wrt HU01: +26	Difference in total words MT02 wrt HU01: +26	Difference in total words MT03 wrt HU01: +9	Difference in total words MT04 wrt HU01: +33	Difference in total words MT05 wrt HU01: +22
---	Difference, total paras HU02 wrt HU01: +5	Difference in total paras MT01 wrt HU01: +3	Difference in total paras MT02 wrt HU01: +4	Difference in total paras MT03 wrt HU01: +3	Difference in total paras MT04 wrt HU01: +4	Difference in total paras MT05 wrt HU01: +2

5. Visual Map Analysis Using ORA*

ORA may be used to display the way concepts relate to one another. The ORA displays concepts as nodes linked by lines representing relationships. The nature of a single relationship is lost, and several relationships between two concepts are aggregated. That is, if verbs are

*This analysis was performed by Sean Murray.

discarded for some reason, man-bites-dog and man-pets-dog are both shown as a bigram (man, dog), with two occurrences. A great deal of meaning is lost. To explore the divergence of concept maps between different translations, several different paths were explored. Document 6 excerpt was used to explore the concept map differences, and the HU01 translation was arbitrarily chosen as ground truth. The initial set of concept maps developed used a small delete list of 29 concepts. The next set of concept maps developed used the small delete list and Porter English stemming (*5*). A set of concept maps was generated that initially used a large delete list of 2526 concepts; however, the list was created for different documents that were not all used in this exploration. This large delete list was therefore simplified. The last set of concept maps was produced using a delete list of only 570 concepts, a medium-sized delete list.

A Generalization Thesaurus with 4383 concepts, developed using different documents, was used. Use of a multiple-purpose Generalization Thesaurus does not, in general, create problems but should be examined carefully by the operator.

A Meta-Matrix Thesaurus was used to alias concept roles by the document it came from.[*] That is, in processing each translation, each word was given an alias role rather than its real, functional role. The alias allowed a visual depiction of concepts labeled by the document in which they were found. For example, in the HU01 translation, each word unique to that document was labeled as “agent”; in HU02, each unique word was labeled as “task”; and words unique to MT05 as “resource.” Words common to pairs of documents were labeled as “match.” Each meta-matrix was combined pair-wise in Microsoft Excel with the HU01 translation, so HU01 and MT01 were combined, HU01 and MT02 were combined, etc. The words from the separate meta-matrices were then sorted alphabetically so that matching words were together, allowing the matching words to be labeled as “match” and the two separate translation labels deleted.

The translated document combinations were opened as pairs in AutoMap, using the new combined Meta-Matrix Thesaurus. Different editing processes in AutoMap, shown in figure 5, produced different outputs in ORA. For better observations, ORA enabled the graphical shifting of nodes (concepts) that obscured each other.

Inspection of the concept maps created from different document translations indicated a fair amount of discrepancy between translations. In these cases, about 33%–34% of the concepts in one document were different from those in another document. This discrepancy between translations is shown by the high proportions of concepts unique to each translation in pair-wise comparisons. These proportions of unique words change depending on preprocessing paths but remain fairly high.

[*] That is, a set of concepts from a given document might all be categorized as “action,” regardless of the function of any given concept in its document, hence the use of the term “alias.”

Figure 5 shows examples of the discrepancy between pairs of translations using four different preprocessing paths. The icons shown in the graphs represent the concepts not common between the two translations. In figure 5, hexagonal icons represent the concepts in MT01 not seen in HU01; square icons represent the HU01 concepts not seen in MT01. Considerably more concept differences are shown between the human and machine translations compared to the differences between the two human translations. The numbers of concepts common to both translations for the different preprocessing paths are dispalyed in the inset boxes but not shown graphically. As can be seen by inspection, there is a large difference in the noncommon words, depending on the preprocessing strategy. This means that any comparison between files must involve the same preprocessing for both files.

The differences are summarized in table 3. These data can only be considered illustrative rather than representative of exact trends concerning the effects of preprocessing. For these engines and document excerpts, there is a relatively large number of concepts unique to the different translation means.

Table 3. Comparison HU01 and MT01, document 6 excerpt, four different preprocessing paths.

Information	**Path 1** **Medium Delete List**	**Path 2** **Small Delete List**	**Path 3** **Large Delete List, English Stemming, Generalization Thesaurus**	**Path 4** **Small Delete List, Stemming, Generalization Thesaurus**
No. concepts unique to MT01 (not found in HU01 ground truth)	70 concepts; 45% of 154 concepts	80 concepts; 45% of 178 concepts	49 concepts; 40% of 123 concepts	67 concepts; 38% of 177 concepts
Total concepts in MT01 under preprocessing	154	178	123	177
No. concepts unique to HU01 (not found in MT01)	67 concepts; 44% of 151 concepts	76 concepts; 44% of 174 concepts	44 concepts; 37% of 118 concepts	61 concepts; 36% of 171 concepts
Total concepts in HU01 under preprocessing	151	174	118	171
No. concepts common to both HU01 and MT01	84 concepts held in common, for path 4	98 concepts held in common, for path 3	74 concepts held in common, for path 2	110 concepts held in common, for path 1

In the case of this particular excerpt of this particular document, with these two translations, there are several points of interest. For the four different processing paths, 38%–45% of the concepts in MT01 are not present in the ground truth translation. They may be considered as, very possibly, spurious. Another possible point of interest is the 35%–44% of the concepts in the ground truth document not found in MT01. These may be considered as, very possibly, information lost.

Some of the divergences in concepts among the translations are due to nontranslation, synonyms, or mistranslations. For instance, some concepts in the MT documents were simply not translated and were left in Spanish. Other divergences are due to use of synonyms, such as the case where HU01 translated a concept as “man,” while one of the MT engines translated the same word as “old chap.” This can be corrected by placing “old chap” in the Generalization Thesaurus and having it changed to “man.” Another issue is inaccurate concept translations. For instance, in document 8, HU01 translated a term as “southeast,” while MT04 had “southwest.” This could seriously impact a unit in the field, deploying incorrectly to intercept a target.

Figure 7 is an example of a divergence concept map, with symbols arranged by the default map generator. The square icons indicate matching concepts (concepts found in both documents). The round and diamond icons indicate concepts unique to HU01 or MT01, respectively. This is before temporary removal of any entities. A medium-sized delete list was applied.

By clearly showing the differences or divergences between the two documents, the common concepts become easily visible and simple to verify. Color is an excellent cue in the program display but is not shown in the figures. In the actual program display, the lines connecting the icons representing the concepts are the color of the source concept. The relationships are directed relationships, with the direction being from the first item (source) in a bigram of concepts to the second. That is, the bigram is shown with the color of the connector line representing the direction of the relationship. In that case, for instance, a red icon representing “man” would have a red link to, perhaps, a differently colored icon representing “dog.”

Another convention uses arrows to indicate directionality; the arrows may also be colored to emphasize directionality. The convention using arrows is illustrated in figure 7. Triangles represent concepts found in MT05 only; ovals represent concepts found only in HU01. Concepts found in both documents are represented by squares. The direction of the relationship is indicated by an arrow; the first concept in the bigram is the origin of the arrow. Thus, a concept common to both documents, such as “empty” (square), shows an arrow directed to the triangle representing the concept “urn,” which is found in MT05 but not in HU01. Again, in a color display, the line representing the relationship shares the color of the source icon.

Visual map analysis is potentially of enormous importance in extraction of information, especially relationships, from textual material. Difficulties arise from the complexity of the visual display and the need for considerable experience in both use of the programs and in the social and tactical contexts of the operation. Perseverance will be needed by any analyst using the visual maps in a free-play setting; the important relationships may not reveal themselves right away. Pruning unimportant material will be essential to discerning the important.

The pruning of unneeded material is likely to be time consuming, making the maps difficult to use in near real-time support roles without considerable prior preparation. In particular, delete lists and thesauri tailored to the situation and the specific questions anticipated will be important. Such prior preparation, however, may involve a risk of excluding relationships not thought to be important ahead of time. The impact of such a risk could be evaluated by trials based on real tactical situations.

Several observations may be made regarding the use of the SNA/TNA toolset under analysis, Automap and its attendant suite of applications, to extract information from MT documents.

- Extraction of information in the form of bigrams and trigrams is labor intensive and highly iterative. Noise must be reduced and information conserved by successive application of thesauri and delete lists.
- If different translation engines are in use, the characteristics of the engines must be thoroughly understood, both in terms of what they translate and what they do not. Application of the engines to several common documents is in order, followed by study of both common and unique concepts from the same document translated in different ways.
- Use of the visualization tools allows examination of the sequential linkage of concepts—information that may be lost when examining mere lists of bigrams and trigrams. This is especially important for links that involve intermediate steps.

6. Experimental Protocol* for Quantitative Assessment

In addition to the qualitative exploration of the translated excerpts described in sections 4–5, a quantitative experiment was performed using longer texts and a single reference translation engine. Document set B was used for this analysis. The experiment involved the following two phases:

1. Phase one included text collection and preprocessing using AutoMap.
2. Phase two of the experiment consisted of concept map analyses.

6.1 Translation Engine

Document set B is a limited corpus comprises 30 Spanish documents of varying length—six short stories, six love letters, and 18 pages from J. R. R. Tolkien's Lord of the Rings (*9*).

Each Spanish document had a respective human translation (or, in the case of Lord of the Rings, the English original) that served as its parallel text, or ground truth, in English. The MT engine chosen was System Analysis Translator (SYSTRAN), because this MT system has a long history of development and is considered a stable, respected technology (*10*). Each Spanish text was run through SYSTRAN, which generated 30 MT documents. These were coupled with 30 human

*This experiment was performed by Michelle McVey and Ashley Foots (U.S. Army Research Laboratory, Human Research and Engineering Directorate) under the supervision of and with advice from Bornstein.

translations to generate a corpus of 60 translated documents. Examples of the human and the high-quality machine translations are provided in appendix B.

6.2 Text Preprocessing

AutoMap's text preprocessing techniques were applied to the translated documents. These techniques are semiautomated and required multiple iterations of the steps discussed in section 2.1.

6.3 AutoMap Analysis

In phase two of the experiment, AutoMap was used to determine if relationships between concepts could be found. Multiple meta-matrices for each of the 30 human translations and 30 machine translations were analyzed. Descriptive results included the number of concepts analyzed, number of concepts in statements, number of isolated concepts, and number of statements. The comparison of the two sets of translations is displayed in table 4. The machine translated documents generated many more concepts than the human translated documents but retained fewer relationships. The data from the multiple meta-matrices analyses were further analyzed using ORA.

Table 4. Comparison concept structure of human and machine translations.

Human Translation	**Machine Translation**
Number of Concepts Analyzed	
196	199
245	267
225	314
252	322
Number of Concepts in Statement	
724	772
729	795
176	198
1494	1914
Number of Isolated Concepts	
0	0
0	0
0	0
0	0
Number of Statements	
274	772
729	795
732	876
747	957

6.4 ORA Visualization Analysis

ORA was used to visualize the design structure of the networks of concepts found in the documents. Design structure refers to relations among entities such as agents, knowledge, resources, and tasks (*8*). The resulting visualizations of the concept structures vary greatly between the human translations (figure 8) and the machine translations (figure 9). These figures are included solely to show the diagrammatic differences.

The visualized networks differ in the numbers and linkages of entities (nodes) and edges (ties). The differences in numbers of the different categories of entities found within the networks are enumerated in table 5. These descriptive frequencies are indicators that MT documents do not appear to produce concept maps or visualizations comparable to those of human translated documents.

Table 5. Numbers of specific categories of entities and total number of entities and edges for human and machine translated documents.

Category	Human Translated	Machine Translated
Agent	34	13
Event	34	19
Knowledge	3	2
Location	27	8
Organization	16	2
Resource	18	15
Role	12	8
Task	9	4
Time	5	4
Entities	158	75
Edges	746	326

A striking feature of the comparison of the maps for document set B, human and machine translations, is the difference in complexity between the two corpora—the map of the MT corpus is far simpler. This may relate to variation between human translations as opposed to more uniformity between translations generated by a single MT engine. These differences in complexity can be seen in figures 8 and 9.

6.5 Precision and Recall for MT Documents

The visual variations seen in the concept maps were confirmed with the calculation of precision, recall, and the F-measure. In information retrieval, precision is defined as a system's ability to retrieve top-ranked entities that are mostly relevant. Recall is defined as the ability of the system to find all the relevant entities in the corpus (*11*). The F-measure fuses precision and recall. A low value for recall coupled with a large value for precision suggests that although some relevant entities were found, many were missed. A high value for recall coupled with a low value for precision implies that most relevant entities were found, but many nonrelevant entities were also retrieved. The ideal value for precision and recall is 1.00; however, this is rarely the case because although recall can be increased by retrieving more entities, this can lead to a reduction in precision (*12*). The trade-off between precision and recall is with low recall comes high precision and with high recall comes low precision.

An example calculation of precision and recall for the MT documents under the “Role” category uses data in table 6. The number of “Role” entities was 12 for the human translated documents and 8 for the machine translated documents. The “Role” entities are listed in table 6. The human translations were used as reference documents, or ground truth. Precision was calculated as the number of relevant retrieved entities divided by the total of number of entities retrieved. Relevant entities were defined as those entities retrieved from the ground truth human translations. The ORA output from the MT documents generated eight entities, all of which were retrieved from the human translations as well. This means that of the eight entities retrieved, all were relevant, resulting in a precision of 1.000. Recall was calculated as the number of relevant entities retrieved divided by the total number of relevant entities. As shown in table 7, the number of relevant entities retrieved was eight. This value was divided by the total number of relevant entities, 12. The resulting recall score was 0.667.

Precision and recall were computed for the remaining meta-matrix categories. The results are shown in table 7.

Table 6. Relevant entities for the “Role” meta-matrix category found by ORA for human translated and machine translated document sets.

Relevant Role Entity	Retrieved in Human Translation	Retrieved in Machine Translation
Nephew	Y	Y
Hobbit	Y	Y
Orphan	Y	Y
Detective	Y	Y
Heir	Y	Y
Master	Y	Y
Wizard	Y	Y
Actress	Y	Y
Private eye	Y	N
Postman	Y	N
Postmen	Y	N
Tourists	Y	N

Notes: Y = yes, N = no.

Table 7. Precision and recall for specific categories of entities for machine translated documents.

Category	Precision	Recall
Agent	1.000	0.382
Event	0.500	0.895
Knowledge	0.500	0.333
Location	1.000	0.296
Organization	1.000	0.125
Resource	1.000	0.833
Role	1.000	0.667
Task	1.000	0.444

Overall, precision was high (i.e., 1.000 for six categories), while recall was low (i.e., < 0.500 for six categories). The high precision and low recall values indicate that while some relevant entities were found, many were overlooked in the MT document set.

The F-measure captures precision and recall and is calculated as F-measure = (2 * precision * recall) / (precision + recall) (*11*), which is the harmonic mean of precision and recall. Both precision and recall must be high for the harmonic mean to be high (*12*). Using the sample average precision, 0.875, and sample average recall, 0.497, the F-measure was calculated as 0.634. The fairly low F-measure combined with the low recall values could be indicative of degraded MT output serving as input to ORA. Degraded documents may no longer contain pertinent information, thus hindering ORA's performance.

The low values for recall and the F-measure confirm that ORA's performance is degraded when analyzing machine translated documents.

7. Conclusion

Most descriptive statistical measures may serve to differentiate between document sets but may not illuminate differences in the concept mappings of translated documents. Direct examination of both concept lists and comparison of the bigrams will be needed. Since the importance of concepts is situation dependent, analyst judgment may be needed to assess the value added or subtracted during translation.

Initial examination of the concept maps indicates that the basic meaning as represented by the concept maps is truncated and obscured, in some cases, greatly. The issues of how much meaning is lost and the degree to which noise is introduced require a more substantive means of measurement.

Applying SNA to MT documents will require human intervention. Currently, there appears to be no reason to preclude application of the SNA software to MT documents data linked to an analysis center in real or near real time. Use of an analysis center with decentralized document gathering is probably required due to the need for an expert software operator who is also well versed in the culture of the area and the current intelligence requirements.

Preliminary results from evaluating SNA output of MT and human translated documents indicated that AutoMap and ORA did not produce comparable concept maps and visualizations from documents translated by several different algorithms. This result was based on a limited Spanish-language corpus, and the results should not be assumed true for all language character sets and may not apply to larger documents. Further, the translation algorithms were selected for convenience and may not represent the best of current practice. Nevertheless, this issue should be investigated for languages of current importance and the current MT engines for those languages.

8. Summary

A method has been developed for assessing the degree to which meaning, as represented by concept maps, is preserved under translation by various means (e.g., machine and human translation). Guidelines for employing SNA software on MT documents have been developed and are outlined in this report.

Should this research be resumed, there are several action items:

- The document corpus needed for the experiment must be secured.
- The experimental design for testing must be refined.
- An evaluation of meaning beyond that represented by Automap’s proximity-based concept pairing, such as latent semantic analysis, should also be performed to provide an alternative view of a document’s meaning.

9. References

1. Baker, R. O. The Decisive Weapon: A Brigade Combat Team Commander's Perspective on Information Operations. *Military Review* May–June **2006**.
2. Baker, R. O. Human-Centric Operations: Developing Actionable Intelligence in the Urban Counterinsurgency Environment. *Military Review* March–April **2007**.
3. Pedahzur, A.; Perliger, A. The Changing Nature of Suicide Attacks: A Social Network Perspective. *Social Forces* **2006**, *84* (4).
4. Center for Computational Analysis of Social and Organizational Systems. http://www.casos .cs.cmu.edu/ (accessed 20 June 2007).
5. Carley, K. M.; Diesner, J.; De Reno, M. AutoMap User's Guide; CMU-ISRI-06-114; Carnegie Mellon University: Pittsburgh, PA, October 2006. http://www.casos.cs.cmu .edu/publications/ papers/CMU-ISRI-06-114.pdf (accessed 27 September 2007).
6. Magnini, B.; Negri, M.; Prevete, R.; Tanev, H. A WordNet-Based Approach to Named Entities Recognition. In *Proceedings of SemaNet'02: Building and Using Semantic Networks*, Taipei, Taiwan, August 2002.
7. Carley, K. M.; DeReno, M. *ORA 2006: User's Guide*; CASOS technical report CMU-ISRI-06-113; Carnegie Mellon University: Pittsburgh, PA, August 2006. http://www.casos .cs.cmu.edu /publications/papers/CMU-ISRI-06-113.pdf (accessed 26 September 2007).
8. Carley, K. M.; Reminga, J. *ORA: Organization Risk Analyzer*; CASOS technical report CMU-ISRI-04-106; Carnegie Mellon University: Pittsburgh, PA, January 2004. http://www.casos.cs.cmu .edu/publications/papers/ carley_2004_oraorganizationrisk.pdf (accessed 26 September 2007).
9. Bilbrough, M. Parallel Texts: Stories and Poems. http://www.englishspanishlink.com /stories%20and%20poems.htm (accessed 16 July 2007).
10. Flanagan, M.; McClure, S. SYSTRAN and the Reinvention of MT. http://www.systransoft .com/index/About-Systran/News-And-Events/Articles (accessed 25 July 2007).
11. Klavans, J.; Resnik, P. *The Balancing Act Combining Symbolic and Statistical Approaches to Language (Language, Speech, and Communication)*; MIT Press: Cambridge, MA, 1996.
12. Manning, C. D.; Schutze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, 2000.

INTENTIONALLY LEFT BLANK.

Appendix A. Document 6 Source and Translations, From Document Set A

This appendix appears in its original form, without editorial change.

Appendix A. Document 6 source and translations, from Document Set A

Source, Document 6

Source TEXT006

ESCRUTINIO SUDAFRICANO: DESASTRE PARA LA COMISIÓN ELECTORAL

El recuento de las históricas elecciones sudafricanas se transformó en un desastre, pues este miércoles, cinco días después del cierre de la votación, no había sido entregado ningún resultado definitivo, mientras se intensificaban las acusaciones de irregularidades.

Los atrasos, la desorganización y las impugnaciones obligaron a postergar hasta el próximo lunes la elección del futuro Presidente, seguramente Nelson Mandela, por la nueva Asamblea nacional sudafricana.

La Comisión electoral independiente, una gigantesca maquinaria administrativa con unos 300.000 empleados, se vio obligada anoche a suspender la difusión de los resultados para revisar sus procedimientos de recuento. No obstante, estaba prevista para la noche del miércoles la publicación de resultados provisorios.

Esta mañana, todos los teléfonos de la Comisión fueron suspendidos, tanto en la sede de Johannesburgo como en Midrand, a unos 30 km, donde había sido instalado un centro ultramoderno, con pantallas de televisión y capacidad para más de 6.000 periodistas.

El descubrimiento de once millones de boletines no distribuidos, en los depósitos de la región de Johannesburgo y de Pretoria, suscitaron el temor de un sabotaje del escrutinio, y cinco miembros de la Comisión fueron interrogados sin que se hayan difundido los resultados.

Igualmente grave, el Congreso Nacional Africano (ANC) y el partido nacionalista zulu Inkatha se acusaron mutuamente de irregularidades - urnas vacías, perdidas o sin sellar- que atrasaron los resultados en la provincia de KwaZulu-Natal, que incluye aproximadamente un cuarto del total de los electores.

Este miércoles había sido escrutado apenas un tercio de los boletines de voto en KwaZulu-Natal y el jefe del Inkatha, Mangosuthu Buthelezi, advirtió que "se abrirá una caja de Pandora si las acusaciones de fraude contra mi partido son retenidas". El Inkatha superaba al ANC con un 20% de los votos mientras continuaba el recuento.

El presidente de la Comisión, el juez Johann Kruger, reconoció ciertas irregularidades, invocando la insuficiencia de instrumentos estadísticos, debido a que no se esperaba una afluencia tan importante de electores negros.

El influyente diario Business Day calculó que al ritmo empleado hasta ahora se necesitaba una hora por asesor para contar un boletín de voto. "Un hombre, una hora, un voto", ironizó el diario.

Human Translation 1 (HU01)

3006NLR[*]

South African Electoral Scrutiny: Disaster for the Commission

The recount of the historic South African elections was transformed into a disaster as, this Wednesday, five days after the close of the voting, no definitive results have been disclosed.

Meanwhile, accusations of irregularities have intensified. Delays, disorganization and accusations have obliged postponing the election of the future President, undoubtedly Nelson Mandela, by the new South African National Assembly until next Monday. The independent Electoral Commission, a gigantic administrative machinery with some 300,000 employees, was obliged last night to suspend the dissemination of results in order to review its recounting procedures.

The publication of provisional results, however, was expected for Wednesday night. This morning all Commission telephone service was interrupted, in Johannesburg as well as in Midrand, some 30 km. away, where an ultramodern communications center has been installed with television screens and a capacity for more than 6,000 journalists. The discovery of eleven million undistributed ballots in warehouses in the Johannesburg and Pretoria region awakened fears of sabotage of the electoral scrutiny.

Five members of the Commission were interrogated but the results were not made public. Equally serious, the African National Congress (ANC) and the nationalist Zulu party Inkatha were accusing each other of such irregularities as empty, lost or unsealed ballot boxes, accusations which delayed results in the province of KwaZulu-Natal, where approximately one fourth of the total number of electors reside. Scarcely a third of the ballots in KwaZulu-Natal had been counted this Wednesday and Inkatha chieftain Mangosuthu Buthelezi warned that "a Pandora's box will be opened if accusations of fraud continue against my party." The Inkatha party continued to lead ANC by 20% as the recount continued. Judge Johann Kruger, Chairman of the Commission, recognized that there had been certain irregularities, blaming the insufficiency of statistical instruments due to the fact that such a significant influx of black voters had not been expected. The influential journal Business Day calculated that the speed developed to date required one hour per election official to count one ballot.

"One man, one hour, one vote," said the newspaper with irony.

[*] Note: this header was in the documents as received. The header, in this case 3006NLR, identifies the translation method, that is, the specific machine translation engine or specific human.

Human Translation 2 (HU02)

3006WG

SOUTH AFRICAN VOTE-COUNTING: A DISASTER FOR THE ELECTORAL COMMISSION

The recount of the historic South African elections became a disaster this Wednesday when, five days after voting ended, no definitive result had been given, while accusations of irregularities increased.

The delays, disorganization, and challenges forced the postponement until next Monday of the election of the future President, surely Nelson Mandela, by the new South African National Assembly.

The independent Electoral Commission, a gigantic administrative machine with about 300,000 employees, was obliged to suspend dissemination of the results last night to revise its recount procedures. Nevertheless, the publication of provisional results was anticipated for Wednesday evening.

This morning, all the Commission's telephones were shut off, both in the Johannesburg headquarters and in Midrand, about 30 km away, where an ultramodern center had been installed, with television screens and capacity for more than 6,000 journalists.

The discovery of eleven million undistributed ballots, in the Johannesburg and Pretoria regions' warehouses, raised the fear of a vote-count sabotage, and five members of the Commission were questioned without the results having been disseminated.

Equally serious, the African National Congress (ANC) and the Zulu nationalist party Inkatha mutually accused each other of irregularities--empty, lost, or unsealed ballot boxes--that were delaying results in the KwaZulu-Natal province, which includes approximately a fourth of the total voters.

This Wednesday scarcely a third of the ballots in KwaZulu-Natal had been counted, and the Inkatha leader, Mangosuthu Buthelezi, warned that "a Pandora's box will be opened if the fraud accusations against my party are retained." Inkatha was leading the ANC with 20% of the votes while the recount continued.

The Commission's President, Judge Johann Kruger, admitted certain irregularities, citing the insufficiency of statistical instruments, due to the fact that such a sizable turnout of black voters was not expected.

The influential daily Business Day calculated that at the rate used up to now an hour per assessor to count a ballot was needed. "One man, one hour, one vote," the daily said ironically.

Machine Translation 1 (MT01)

3006L

Escrutinio South African: disaster for the electoral commission

The count of the historical South African elections itself transformed in a disaster, since this Wednesday, five days after the closing of the voting, not had been delivered no resulted definitive, while itself intensified the accusations of irregularitieses.

The delays, the lack of organization and the impugnaciones obligated to delay even the close Monday the choice of the future Presidente, probably Nelson Mandela, by the new national South African Asamblea.

The electoral independent Comision, a gigantesca administrative machinery with some 300,000 employees, itself saw obligated last night to suspend the spreading of the resultados to revise its methods of count. Nevertheless the publication of temporary resultados, was anticipated for the night of Wednesday.

This morning, everybody them telephones of the Comision was suspended, goal in the headquarters of Johannesburg as in Midrand, to some 30 km, which had been installed an ultramoderno center, with screens of television and capacity stops more of 6,000 journalists. The discovery of eleven millions of reports not distributed, in the deposits of the region of Johannesburg and of Pretoria, caused the fear of a sabotage of the escrutinio, and five members of the Comision was questioned without that itself you have spread the resultados.

The same serious, the Congress Nacional Africano (ANC) and the party nationalist zulu Inkatha itself accused mutuamente of vacant irregularitieses-urns, lost or without sellar- that delayed the resultados in the province of KwaZulu-Natal, that includes approximately a quarter of the total of the electing.

This Wednesday had been searched scarcely one third of the reports of vote in KwaZulu-Natal and the head of the Inkatha, Mangosuthu Buthelezi, warned that "itself will open a box of Pandora if the charges of fraud against my party are retained". The Inkatha surpassed to the ANC with 20% of the votes while continued the count.

The president of the Comision, the judge Johann Kruger, identified certain irregularitieses, invoking the inadequacies of instruments statisticals, owed to that not itself expected a so significant inflow of electing blacks. The influential diary Business Day calculated that to the rhythm employee even now itself needed an hour by consultant to tell a report of vote. "A man, a time, one I vote", ironizo the journal.

Machine Translation 2 (MT02)

3006SY

SOUTH AFRICAN SCRUTINY: DISASTER FOR THE ELECTORAL COMMISSION

The count of the historical South African elections was transformed into a disaster, because this Wednesday, five days after the closing of the voting, had been given no definitive result, while the accusations of irregularities intensified.

The atrasos , disorganization and the oppositions forced to delay until the next Monday the election of future Presidente, surely nelson mandela , by the new South African national Assembly.

The independent electoral Commission, a gigantic administrative machinery with about 300,000 employees, was forced last night to suspend the diffusion of the results to review its procedures of count. However, it was anticipated for the night of the Wednesday the publication of provisorios results.

This morning, all the telephones of the Commission were suspended, as much in the seat of Johannesburg like in midrand , to about 30 km, where an ultramodern center had been installed, with screens of television and capacity for more than 6,000 journalists.

The discovery of eleven million distributed bulletins, in the deposits of the region of Johannesburg and pretoria , did not provoke the fear of a sabotage of the scrutiny, and five members of the Commission were interrogated without the results have spread.

Also it burdens, the African National Congress (ANC) and nationalistic party zulu inkatha mutually accused of irregularities - empty ballot boxes, lost or without sellar- that the results in the province of KwaZulu- birthday retarded, that approximately includes a quarter of the total of the voters.

This Wednesday had been scrutinized as soon as a third of bulletins of vote in KwaZulu-birthday and the head of inkatha , mangosuthu buthelezi , noticed that "a box of pandora will be opened if the accusations of fraud against my party are retained". inkatha surpassed to the ANC with a 20% of the votes while it continued the count.

The president of the Commission, judge johann kruger , recognized certain irregularities, invoking the insufficiency of statistical instruments, because a so important affluence of black voters was not expected.

The influential daily business Day calculated that to the rate used until now one hour by adviser was needed to count a vote bulletin. "A man, one hour, a vote", ironizo' the daily.

Machine Translation 3 (MT03)

3006SY

SOUTH AFRICAN SCRUTINY: DISASTER FOR THE ELECTORAL COMMISSION

The count of the historical South African elections was transformed into a disaster, because this Wednesday, five days after the closing of the voting, had been given no definitive result, while the accusations of irregularities intensified.

The atrasos , disorganization and the oppositions forced to delay until the next Monday the election of future Presidente, surely nelson mandela , by the new South African national Assembly.

The independent electoral Commission, a gigantic administrative machinery with about 300,000 employees, was forced last night to suspend the diffusion of the results to review its procedures of count. However, it was anticipated for the night of the Wednesday the publication of provisorios results.

This morning, all the telephones of the Commission were suspended, as much in the seat of Johannesburg like in midrand , to about 30 km, where an ultramodern center had been installed, with screens of television and capacity for more than 6,000 journalists.

The discovery of eleven million distributed bulletins, in the deposits of the region of Johannesburg and pretoria , did not provoke the fear of a sabotage of the scrutiny, and five members of the Commission were interrogated without the results have spread.

Also it burdens, the African National Congress (ANC) and nationalistic party zulu inkatha mutually accused of irregularities - empty ballot boxes, lost or without sellar- that the results in the province of KwaZulu- birthday retarded, that approximately includes a quarter of the total of the voters.

This Wednesday had been scrutinized as soon as a third of bulletins of vote in KwaZulu- birthday and the head of inkatha , mangosuthu buthelezi , noticed that "a box of pandora will be opened if the accusations of fraud against my party are retained". inkatha surpassed to the ANC with a 20% of the votes while it continued the count.

The president of the Commission, judge johann kruger , recognized certain irregularities, invoking the insufficiency of statistical instruments, because a so important affluence of black voters was not expected.

The influential daily business Day calculated that to the rate used until now one hour by adviser was needed to count a vote bulletin. "A man, one hour, a vote", ironizo' the daily.

Machine Translation 4 (MT04)

3006PA

South African scrutiny: disaster for the electoral commission

The count of the historical South African elections was turned into a disaster, then this Wednesday, five days after the closing of the voting, there had not been delivered any resulting definitive, while there were intensified the accusations of irregularities.

The arrears, the disorganization and the impugnaciones made it necessary to postpone until next Monday the election of the future President, surely Nelson Mandela, by the new Assembly national South African.

The independent electoral Commission, a huge administrative machinery with some 300,000 employees, was forced last night to suspend the dissemination of the results in order to review their procedures of count. However, there was foreseen for the night of Wednesday the publication of provisional results.

This morning, all the telephones of the Commission were suspended, both at the headquarters of Johannesburgo and in Midrand, to some 30 kms, where there had been installed an ultramoderno center, with screens of television and capacity for more than 6,000 journalists.

The discovery of eleven million undistributed bulletins, in the deposits of the region of Johannesburgo and of Pretoria, raised the fear of a sabotage of the scrutiny, and five members of the Commission were interrogated without there have been disseminated the results.

Also serious, the African National Congress (ANC) and the nationalist party zulu Inkatha were accused mutually of irregularities - empty urns, missing or without sealing - that they slowed down the results in the province of KwaZulu-Natal, who includes approximately a room of the total of the electors.

This Wednesday there had been barely scrutinized a third of the bulletins of vote in KwaZulu-Natal and the chief of the Inkatha, Mangosuthu Buthelezi, noticed that "will open a box of Pandora if the accusations of fraud against my party are held." The Inkatha surpassed the ANC with 20% of the votes while there continued the count.

The president of the Commission, the judge Johann Kruger, recognized certain irregularities, invoking the insufficiency of statistical instruments, since one did not expect an as important affluence of black electors.

The influential daily Business Day calculated that at the rate utilized until now there was a need a hour by advisor in order to count a bulletin of vote. "A man, a hour, a vote", ironizó the diary.

Machine Translation 5 (MT05)

3006GP

SCRUTINY SUDAFRICANO: DISASTER FOR THE ELECTORAL COMMISSION

The inventory of the historical elections sudafricanas became a disaster, since this Wednesday, five days after of the close of the voting, had not been delivered any definitive result, while was intensified the irregularity accusations.

The lags, the disorganization and the objections compelled to defer until the next Monday the election of the future President, certainly Nelson Order it, by the new national Assembly sudafricana.

The independent electoral Commission, a gigantic administrative machinery with some 300.000 personnel, was seen obligated last night to discontinue the diffusion of the results to check their/its inventory procedures. Nevertheless, was anticipated for the night of the Wednesday the temporary results publication.

This morning, all the telephones of the Commission were discontinued, in the headquarters of Johannesburgo as well as in Midrand, to some 30 km, where had been installed a center ultramoderno, with television screens and capacity for more than 6.000 journalists. The eleven discovery million of not distributed bulletins, in the deposits of the region of Johannesburgo and of Pretoria, raised the dread of a sabotage of the scrutiny, and five members of the Commission were interrogated without may have been spread the results.

Equally serious, the African National Congress (ANC) and the nationalistic party zulu Inkatha were accused mutually of irregularities- empty urns, lost or without sealing- that they retarded the results in the province of KwaZulu - Natal, that includes approximately a quarter of the total of the electors.

This Wednesday had been polled hardly a third of the vote bulletins in KwaZulu - Natal and the chief of the Inkatha, Mangosuthu Buthelezi, warned that "will be opened a Pandora's box if the defraudation accusations against my party are held". The Inkatha was surpassing to the ANC with a 20% of the votes while was continuing the inventory.

The president of the Commission, the judge Johann Kruger, recognized certain irregularities, invoking the statistic instruments insufficiency, due to the fact that were not waited an abundance so important of black electors. The influential daily Business Day calculated that to the pace employed up until now were needed a hour by advising to count a vote bulletin. "A man, a hour, a vote", ironizó the newspaper.

INTENTIONALLY LEFT BLANK.

Appendix B. Example Excerpts From a Human Translation and Its Corresponding Machine Translation, Document Set B

This appendix appears in its original form, without editorial change.

Appendix B. Example excerpts from a human translation and its corresponding machine translation, Document Set B.

Human Translated Document

When Mr. Bilbo Baggins of Bag End announced that he would shortly be celebrating his eleventy-first birthday with a party of special magnificence, there was much talk and excitement in Hobbiton.
Bilbo was very rich and very peculiar, and had been the wonder of the Shire for sixty years, ever since his remarkable disappearance and unexpected return. The riches he had brought back from his travels had now become a local legend, and it was popularly believed, whatever the old folk might say, that the Hill at Bag End was full of tunnels stuffed with treasure. And if that was not enough for fame, there was also his prolonged vigour to marvel at. Time wore on, but it seemed to have little effect on Mr. Baggins. At ninety he was much the same as at fifty. At ninety-nine they began to call him well-preserved, but unchanged would have been nearer the mark. There were some that shook their heads and thought this was too much of a good thing; it seemed unfair that anyone should possess (apparently) perpetual youth as well as (reputedly) inexhaustible wealth.

Machine Translated Document

When Mr. Bilbo Handbag of Closed Handbag announced that in a moment he would celebrate its birthday hundredth eleventh with a celebration of special magnificencia, was many commentaries and excitation in Hobbiton. Bilbo was very rich and very peculiar and had been the astonishment of the Region during sixty years, from its memorable disappearance and unexpected return. The wealth that had brought of those trips had turned local legend and was common belief, against everything what they could say the old ones, that in the hill of Closed Handbag there were many stuffed tunnels of treasures. As if this was not sufficient to give fame him, the prolonged vigor of Mr. Bolso'n was the wonder of the Region. The time happened, but it seemed to affect it very little. To the ninety years it had the same aspect that to the fifty. To ninety and the nine began to consider it "conserved well", but "without changes" it had more been near the truth. There were many moved the head thinking that they were too many good things; it seemed unjust that somebody had (in appearance) an eternal youth and simultaneously (it assumed) inexhaustible goods.

Appendix C. Master Concept List, Translations HU01, HU02, and MT01–MT05, Document 6 Excerpt

The tables in this appendix appear in their original form, without editorial change.

Appendix C. Master Concept List, Translations HU01-HU02 and MT01-MT05, Document 6 Excerpt.

Table C-1. Master concepts lists, all translations, document 6 excerpt.

HU01 concept list after generalization and deletion	#	HU2 concept list after generalization and deletion	#	MT01 concept list after generalization and deletion	#	MT02 concept list after generalization and deletion	#	MT03 concept list after generalization and deletion	#	MT04 concept list after generalization and deletion	#	MT05 concept list after generalization and deletion	#
accusations	4	accusations	2	accusations	1	accusations	2	ability	1	accusations	2	accusations	2
administrative	1	administrative	1	administrative	1	administrative	1	adviser	1	administrative	1	administrative	1
african_national _congress	2	admitted	1	**african_national _ congress**	3	adviser	1	**african_national_ congress**	3	advisor	1	**african_national _ congress**	2
assembly	1	**african_national _congress**	2	*africano*	1	**african_national _congress**	2	area	1	**african_national_ congress**	2	assembly	1
ballot	2	assembly	1	*asamblea*	1	assembly	1	award	1	arrears	1	black	1
ballots	2	assessor	1	blacks	1	*atrasos*	1	ballot	1	assembly	1	box	1
black_voters	1	ballot	2	box	1	ballot	1	*ballotbox*	1	black	1	bulletin	1
blaming	1	ballots	2	business	1	birthday	2	billets	1	box	1	bulletins	2
box	1	**black_voters**	1	caused	1	**black_voters**	1	box	1	bulletin	1	business	1
boxes	1	box	1	charges	1	box	1	bulletin	1	bulletins	2	chief	1
business	1	business	1	choice	1	boxes	1	bulletins	1	business	1	commission	5
chairman	1	challenges	1	closing	1	bulletin	1	business	1	chief	1	compelled	1
chieftain	1	citing	1	*comision*	4	bulletins	2	chairman	2	closing	1	continuing	1
commission	5	commission	3	commission	1	burdens	1	challenge	1	commission	5	*defraudation*	1
communications	1	commissions	2	consultant	1	business	1	chap	1	deposits	1	deposits	1
counted	1	counted	1	deposits	1	closing	1	charges	2	disseminated	1	discontinued	1
election	2	delaying	1	electing	2	commission	5	clerk	1	election	1	election	1
elections	1	disseminated	1	elections	1	deposits	1	commission	2	elections	1	elections	1
electors	1	election	1	employees	1	election	1	committee	1	electors	2	electors	2
employees	1	elections	1	*escrutinio*	2	elections	1	consequence	1	employees	1	headquarters	1
expected	2	employees	1	expected	1	employees	1	consequences	1	forced	1	includes	1
fraud	1	ended	1	fraud	1	expected	1	converted	1	fraud	1	**inkatha**	3

inkatha	3	evening	1	*gigantesca*	1	forced	2	counted	1	headquarters	1	instruments	1
instruments	1	expected	1	goal	1	fraud	1	counting	1	*impugnaciones*	1	*ironizo*	1
irregularities	3	forced	1	head	1	head	1	elections	1	includes	1	irregularities	2
johann_kruger	1	fraud	1	headquarters	1	includes	1	electors	2	inkatha	3	johann_kruger	1
johannesburg	2	headquarters	1	identified	1	inkatha	3	eleventh	1	instruments	1	*johannesburgo*	2
journalists	1	includes	1	*impugnaciones*	1	instruments	1	*este*	1	*ironiz*	1	journalists	1
judge	1	increased	1	inadequacies	1	*ironizo*	1	excelled	1	irregularities	3	judge	1
kwazulu_natal	2	inkatha	3	includes	1	irregularities	3	executive	1	johann_kruger	1	kwazulu_natal	2
mangosuthu_buthelezi	1	instruments	1	inflow	1	johann_kruger	1	expected	1	*johannesburgo*	2	mangosuthu_buthelezi	1
members	1	ironically	1	inkatha	3	johannesburg	2	fellow	1	journalists	1	members	1
midrand	1	irregularities	2	instruments	1	journalists	1	fiance	1	judge	1	midrand	1
morning	1	johann_kruger	1	*ironizo*	1	judge	1	fraud	1	*kms*	1	morning	1
national	1	johannesburg	2	*irregularitieses*	2	*kwazulu*	2	*guyrope*	1	kwazulu_natal	2	national	1
nationalist	1	journalists	1	*irregularitiesesurns*	1	mangosuthu_buthelezi	1	headquarters	1	mangosuthu_buthelezi	1	*nationalistic*	1
nelson	1	judge	1	johann_kruger	1	members	1	includes	1	members	1	nelson	1
nelson_mandela	1	kwazulu_natal	2	johannesburg	2	midrand	1	inkatha	3	midrand	1	objections	1
obliged	2	leader	1	journalists	1	morning	1	johann_kruger	1	missing	1	pandoras	1
official	1	leading	1	judge	1	national	1	*johannesburgo*	2	morning	1	party	2
pandoras	1	machine	1	kwazulu_natal	2	*nationalistic*	1	journalists	1	national	1	personnel	1
party	2	mangosuthu_buthelezi	1	lack	1	nelson	1	judge	1	nationalist	1	polled	1
postponing	1	members	1	mangosuthu_buthelezi	1	nelson_mandela	1	kwazulu_natal	2	nelson	1	president	2
president	1	midrand	1	members	1	noticed	1	lack	1	nelson_mandela	1	pretoria	1
pretoria	1	morning	1	methods	1	*oppositions*	1	leader	1	noticed	1	province	1
province	1	national	1	midrand	1	pandora	1	mangosuthu_buthelezi	1	pandora	1	retarded	1
recount	2	nationalist	1	millions	1	party	2	members	1	party	2	sabotage	1
sabotage	1	nelson	1	morning	1	president	1	**midrand**	1	president	2	sealing	1
south_african	3	**nelson_mandela**	1	*mutuamente*		*presidente*		millions	1	pretoria	1	statistic	1
telephone	1	obliged	1	*nacional*	1	pretoria	1	national	1	province	1	*sudafricana*	1
television	1	**pandoras**	1	national	1	province	1	nationalist	1	resulting	1	*sudafricanas*	1
transformed	1	party	2	nationalist	1	*provisorios*	1	needing	1	sabotage	1	*sudafricano*	1
vote	1	*postponement*	1	nelson	1	retained	1	nelson	1	scrutinized	1	surpassing	1
voting	1	president	2	**nelson_mandela**	1	retarded	1	**nelson_mandela**	1	sealing	1	telephones	1

warehouses	1	pretoria	1	organization	1	sabotage	1	*obligor*	1	slowed	1	television	1
		province	1	owed	1	scrutinized	1	**pandoras**	1	**south_african**	3	*theirits*	1
		questioned	1	pandora	1	seat		perpetration	2	surpassed	1	*ultramoderno*	1
		recount	3	party	1	*sellar*	1	posting	1	suspended	1	urns	1
		regions	1	president	2	**south_african**	1	pretoria	1	telephones	1	vote	3
		retained	1	*presidente*	1	surpassed	1	province	1	television	1	votes	1
		revise	1	pretoria	1	suspended	3	reckoned	1	turned	1	voting	1
		sabotage	1	province	1	telephones	1	recommend	1	*ultramoderno*	1	zulu	1
		south_african	3	questioned	1	television	1	recount	2	urns	1		
		telephones	1	reports	1	transformed	1	reviewed	1	utilized	1		
		television	1	*resultados*	2	vote	1	ridiculed	1	vote	3		
		vote	1	resulted	4	voters	1	sabotage	1	votes	1		
		voters	1	retained	1	votes	3	selection	1	voting	1		
		votes	1	revise	1	voting	1	sequels	1	zulu	1		
		voting	1	sabotage	1	zulu	1	*setbacks*	1				
		warehouses	1	searched	1		1	shutdown	1				
		zulu	1	*sellar*	1		1	**south_african**	3				
				south_african	1			stirred	1				
				spreading	3			suffrage	1				
				statisticals	1			survey	1				
				stops	1			telephones	1				
				surpassed	1			television	1				
				suspended	1			thousand	1				
				telephones	1			told	1				
				television	1			tomorrow	1				
				transformed	1			tools	1				
				ultramoderno	1			transmitted	1				
				vote	1			vote	2				
				votes	3			votes	2				
				voting	1			weds	1				
				zulu	1			yards	1				
					1			zulu	1				

Notes: Words in bold type are nondictionary words found in HU01. Words in italic type are untranslated. Bold type words are strings generated by the Generalization Thesaurus.

Table C-2. Statistical summary of the master concepts lists from all translations, document 6 excerpt.

total number of concepts unique to HU01: 55	total number of concepts unique to HU02: 71	total number of concepts unique to MT01: 85	total number of concepts unique to MT02: 69	total number of concepts unique to MT03: 86	total number of concepts unique to MT04: 68	total number of concepts unique to MT05: 62
HU01 total concepts after deletion and generalization: 78	HU02 total concepts after deletion and generalization: 89	MT01 total concepts after deletion and generalization:106	MT02 total concepts after deletion and generalization: 89	MT03 total concepts after deletion and generalization:102	MT04 total concepts after deletion and generalization: 88	MT05 total concepts after deletion and generalization: 79
	no. of unique concepts HU02 wrt HU01 after deletion, generalization: 16	no. of unique concepts MT01 wrt HU01 after deletion, generalization: 30	no. of unique concepts MT02 wrt HU01 after deletion, generalization:14	no. of unique concepts MT03 wrt HU01 after deletion, generalization: 31	no. of unique concepts MT04 wrt HU01 after deletion, generalization: 13	no. of unique concepts MT05 wrt HU01 after deletion, generalization: 7
	no. of total concepts HU02 wrt HU01 after deletion, generalization: 11	no. of total concepts MT01 wrt HU01 after deletion, generalization: 28	no. of total concepts MT02 wrt HU01 after deletion, generalization: 11	no. of total concepts MT03 wrt HU01 after deletion, generalization: 24	no. of total concepts MT04 wrt HU01 after deletion, generalization: 10	no. of total concepts MT05 wrt HU01 after deletion, generalization: 1
10 concepts listed as non-english-dictionary	non-english-dictionary concepts in HU02 after deletion and generalization in common with HU01: 10	non-english-dictionary concepts in MT01 after deletion and generalization in common with HU01: 8	non-english-dictionary concepts in MT02 after deletion and generalization in common with HU01:8	non-english-dictionary concepts in MT03 after deletion and generalization in common with HU01: 6	non-english-dictionary concepts in MT04 after deletion and generalization in common with HU01: 8	non-english-dictionary concepts in MT05 after deletion and generalization in common with HU01: 7
HU01 MSO 2003 word count, less header, incl. title: 348	non-english-dictionary concepts in addition to those in HU01: 1	non-english-dictionary concepts in addition to those in HU01: 16	non-english-dictionary concepts in addition to those in HU01: 8	non-english-dictionary concepts in addition to those in HU01: 10	non-english-dictionary concepts in addition to those in HU01: 5	non-english-dictionary concepts in addition to those in HU01: 9

HU01 MSO 2003 paragraph count, less header, less title: 4	HU02 word count, less header, incl. title: 334	MT01 word count, less header, incl. title: 374	MT02 word count, less header, incl. title: 374	MT03 word count, less header, incl. title: 357	MT04 word count, less header, incl. title: 381	MT05 word count, less header, incl. title: 370
	HU02 paragraph count, less header, less. title: 9	MT01 paragraph count, less header, less. title: 7	MT02 paragraph count, less header, less. title: 8	MT03 paragraph count, less header, less title: 7	MT04 paragraph count, less header, less title: 8	MT05 paragraph count, less header, less title: 6
	no. of total words HU02 wrt HU01: -14	no. of total words MT01 wrt HU01: +26	no. of total words MT02 wrt HU01: +26	no. of total words MT03 wrt HU01: +9	no. of total words MT04 wrt HU01: +33	no. of total words MT05 wrt HU01: +22
	no. of total paras HU02 wrt HU01: +5	no. of total paras MT01 wrt HU01: +3	no. of total paras MT02 wrt HU01: +4	no. of total paras MT03 wrt HU01: +3	no. of total paras MT04 wrt HU01: +4	no. of total paras MT05 wrt HU01: +2

Source: SNA-MT DRAFT MR v19.doc, 4 August 2009.

Bibliography

Conover, W. J. *Practical Nonparametric Statistics: Second Edition*; Wiley & Sons, Inc.: New York, NY.

Tanenbaum, W.; Brand, J. *Using AutoMap for Social and Textual Network Analysis*; technical note 321; U.S. Army Research Laboratory: Aberdeen Proving Ground, MD, July 2008.

NO. OF
COPIES ORGANIZATION

1 (PDF only) DEFENSE TECHNICAL
INFORMATION CTR
DTIC OCA
8725 JOHN J KINGMAN RD
STE 0944
FORT BELVOIR VA 22060-6218

1 DIRECTOR
US ARMY RESEARCH LAB
IMNE ALC HRR
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
RDRL CIM L
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
RDRL CIM P
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
RDRL D
2800 POWDER MILL RD
ADELPHI MD 20783-1197

ABERDEEN PROVING GROUND

1 DIR USARL
RDRL CIM G (BLDG 4600)

NO. OF COPIES	ORGANIZATION
4	DIRECTOR US ARMY RESEARCH LAB RDRL CII B BROOME RDRL CII T M HOLLAND M VANNI C VOSS 2800 POWDER MILL RD ADELPHI MD 20783-1197
	ABERDEEN PROVING GROUND
7	DIR USARL RDRL CII C A BORNSTEIN (5 CPS) J BRAND D WELSH

INTENTIONALLY LEFT BLANK.