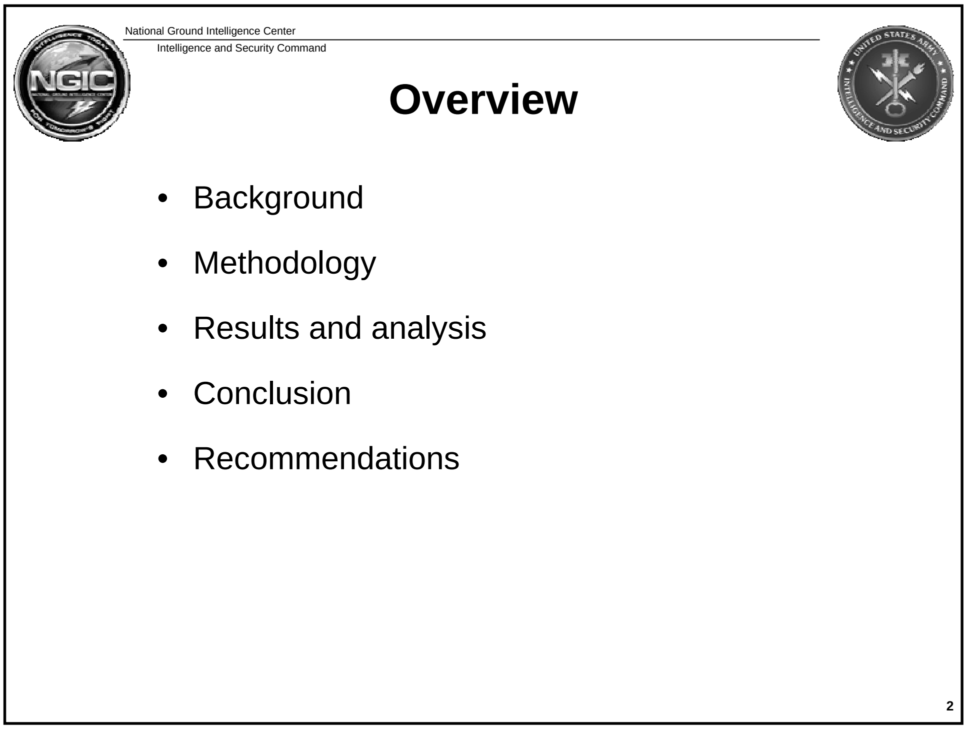
NATIONAL GROUND INTELLIGENCE CENTER
INTELLIGENCE TODAY FOR TOMORROW'S FIGHT

MT Errors in CH-to-EN MT Systems: user feedback

Prepared for the AMTA Conference
21-25 October 2008

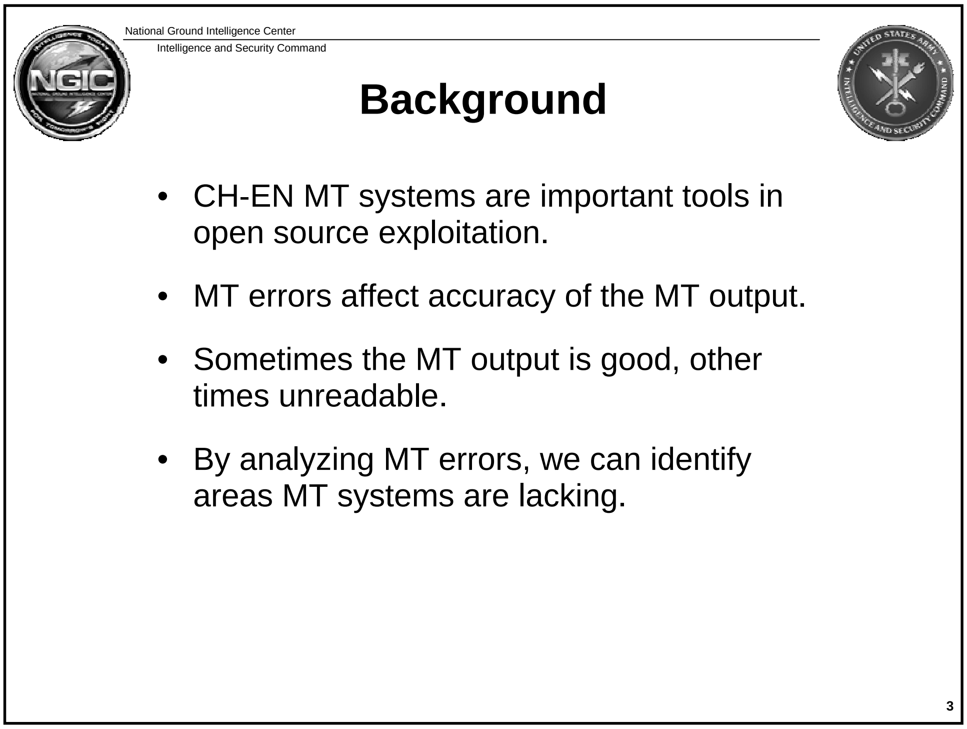
Shin Chang-Meadows

The Overall classification of this Briefing is
UNCLASSIFIED



Overview

- Background
- Methodology
- Results and analysis
- Conclusion
- Recommendations

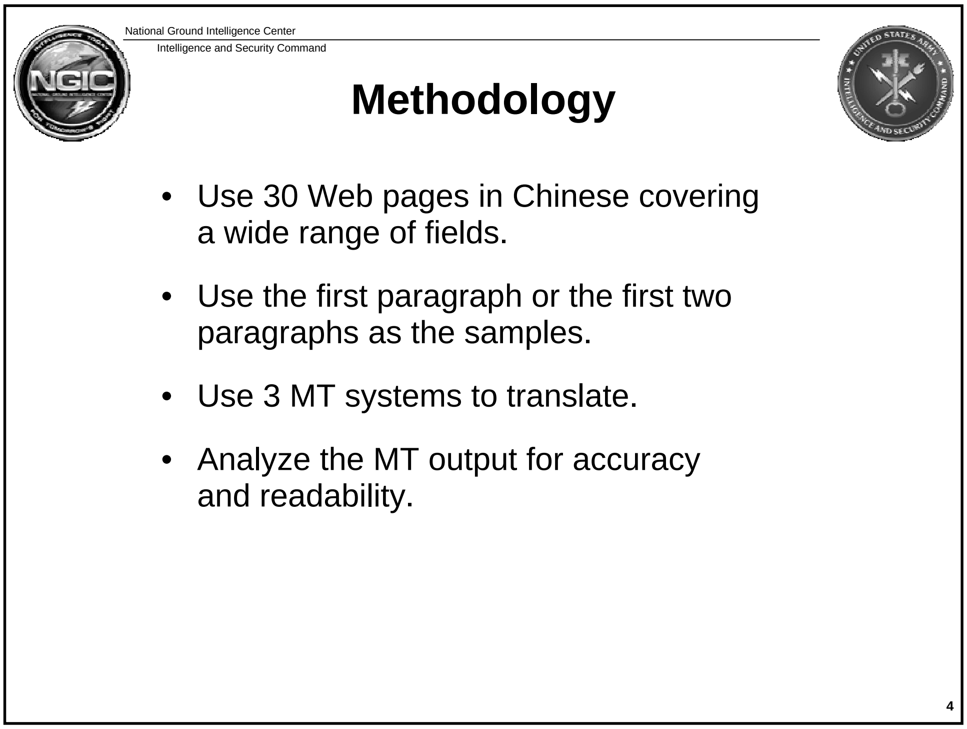


National Ground Intelligence Center

Intelligence and Security Command

Background

- CH-EN MT systems are important tools in open source exploitation.
- MT errors affect accuracy of the MT output.
- Sometimes the MT output is good, other times unreadable.
- By analyzing MT errors, we can identify areas MT systems are lacking.

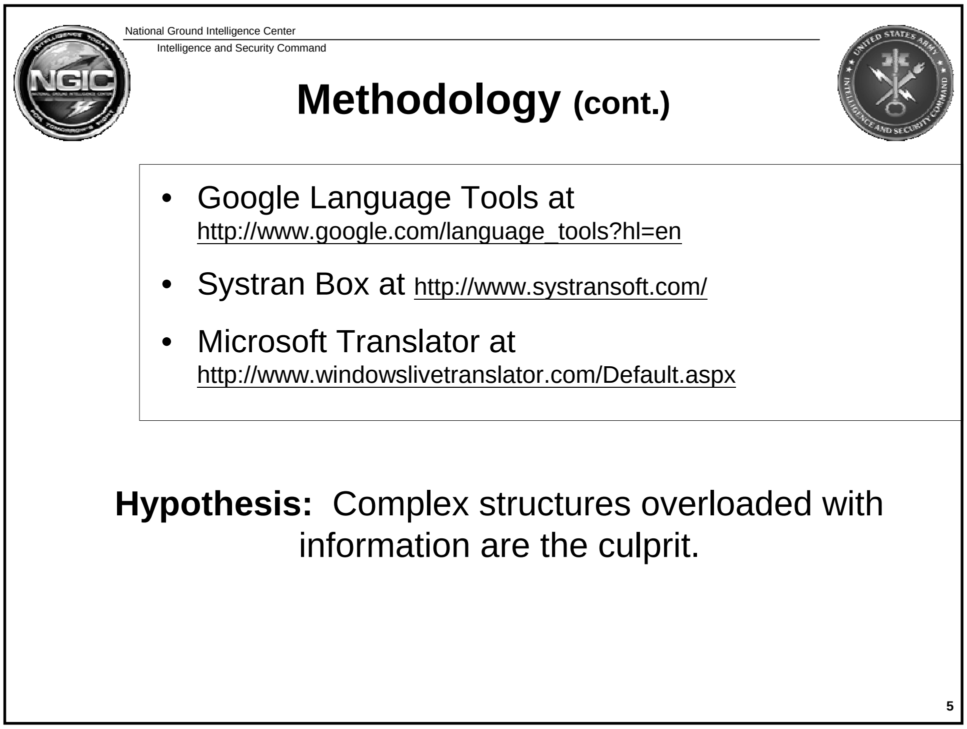


National Ground Intelligence Center

Intelligence and Security Command

Methodology

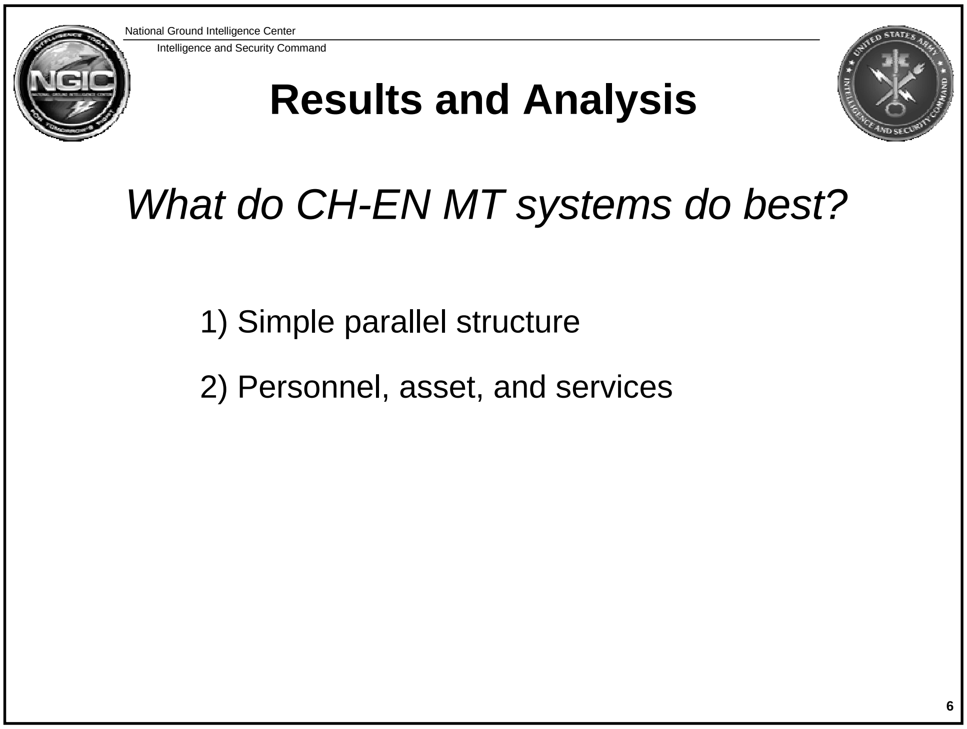
- Use 30 Web pages in Chinese covering a wide range of fields.
- Use the first paragraph or the first two paragraphs as the samples.
- Use 3 MT systems to translate.
- Analyze the MT output for accuracy and readability.



Methodology (cont.)

- Google Language Tools at http://www.google.com/language_tools?hl=en
- Systran Box at http://www.systransoft.com/
- Microsoft Translator at http://www.windowslivetranslator.com/Default.aspx

Hypothesis: Complex structures overloaded with information are the culprit.

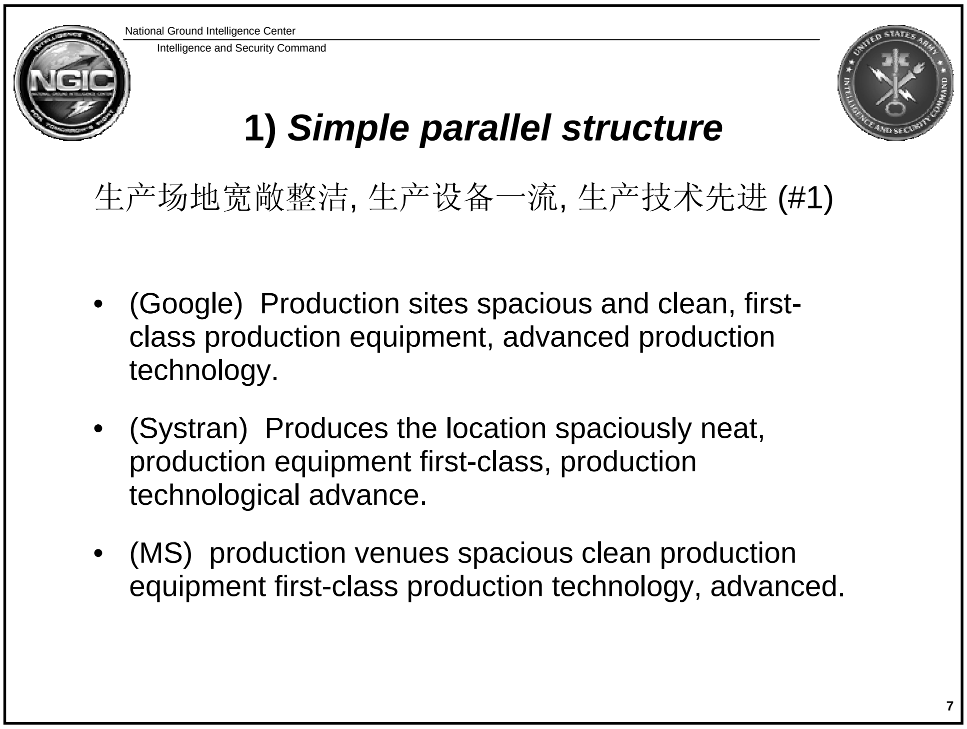


Results and Analysis

What do CH-EN MT systems do best?

1) Simple parallel structure

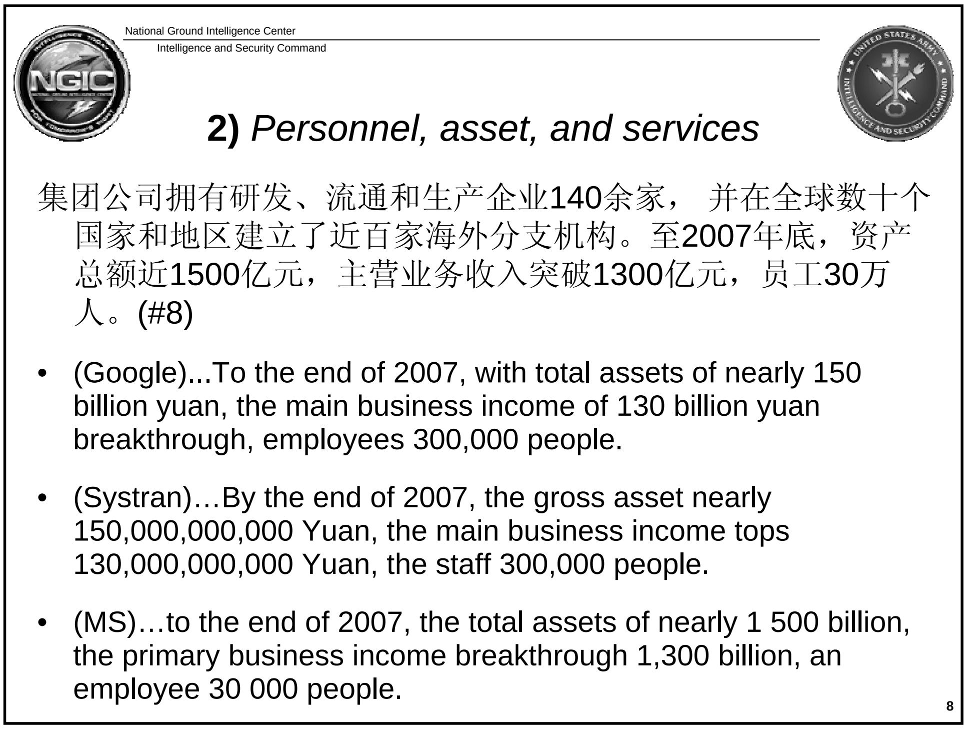
2) Personnel, asset, and services



1) *Simple parallel structure*

生产场地宽敞整洁, 生产设备一流, 生产技术先进 (#1)

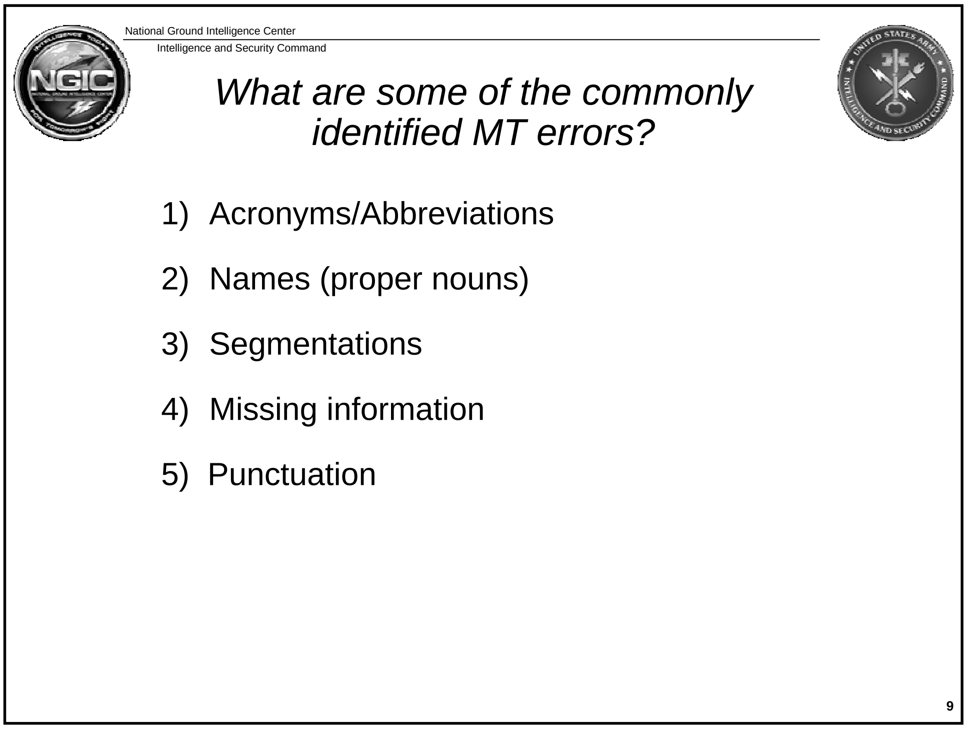
- (Google) Production sites spacious and clean, first-class production equipment, advanced production technology.
- (Systran) Produces the location spaciously neat, production equipment first-class, production technological advance.
- (MS) production venues spacious clean production equipment first-class production technology, advanced.



2) *Personnel, asset, and services*

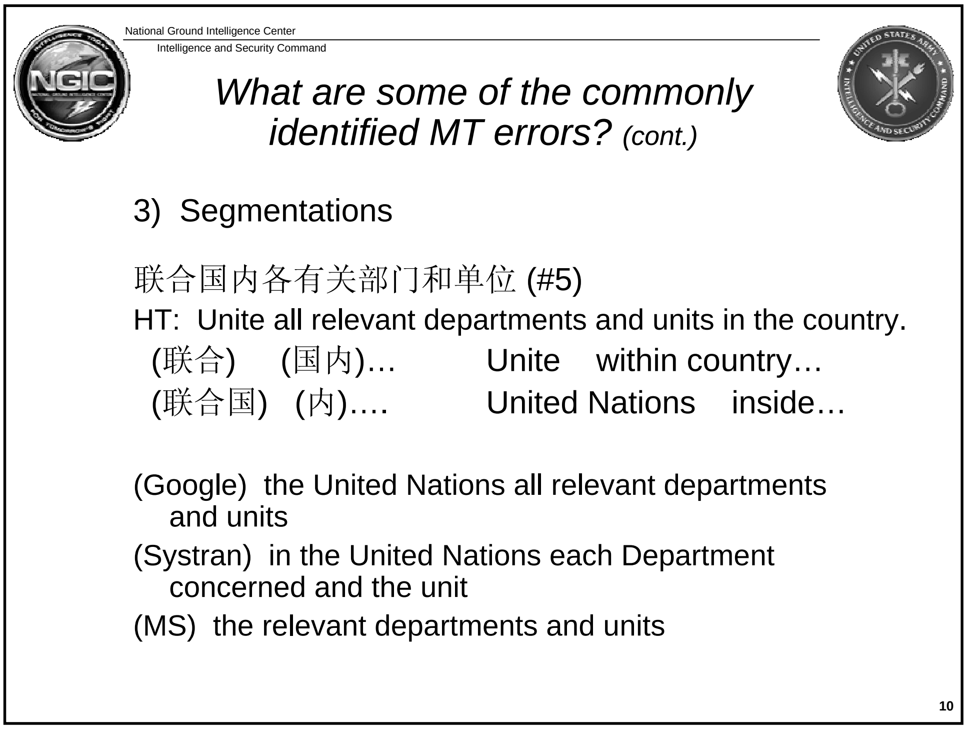
集团公司拥有研发、流通和生产企业140余家， 并在全球数十个国家和地区建立了近百家海外分支机构。至2007年底，资产总额近1500亿元，主营业务收入突破1300亿元，员工30万人。(#8)

- (Google)...To the end of 2007, with total assets of nearly 150 billion yuan, the main business income of 130 billion yuan breakthrough, employees 300,000 people.
- (Systran)…By the end of 2007, the gross asset nearly 150,000,000,000 Yuan, the main business income tops 130,000,000,000 Yuan, the staff 300,000 people.
- (MS)…to the end of 2007, the total assets of nearly 1 500 billion, the primary business income breakthrough 1,300 billion, an employee 30 000 people.



What are some of the commonly identified MT errors?

1) Acronyms/Abbreviations
2) Names (proper nouns)
3) Segmentations
4) Missing information
5) Punctuation



What are some of the commonly identified MT errors? (cont.)

3) Segmentations

联合国内各有关部门和单位 (#5)

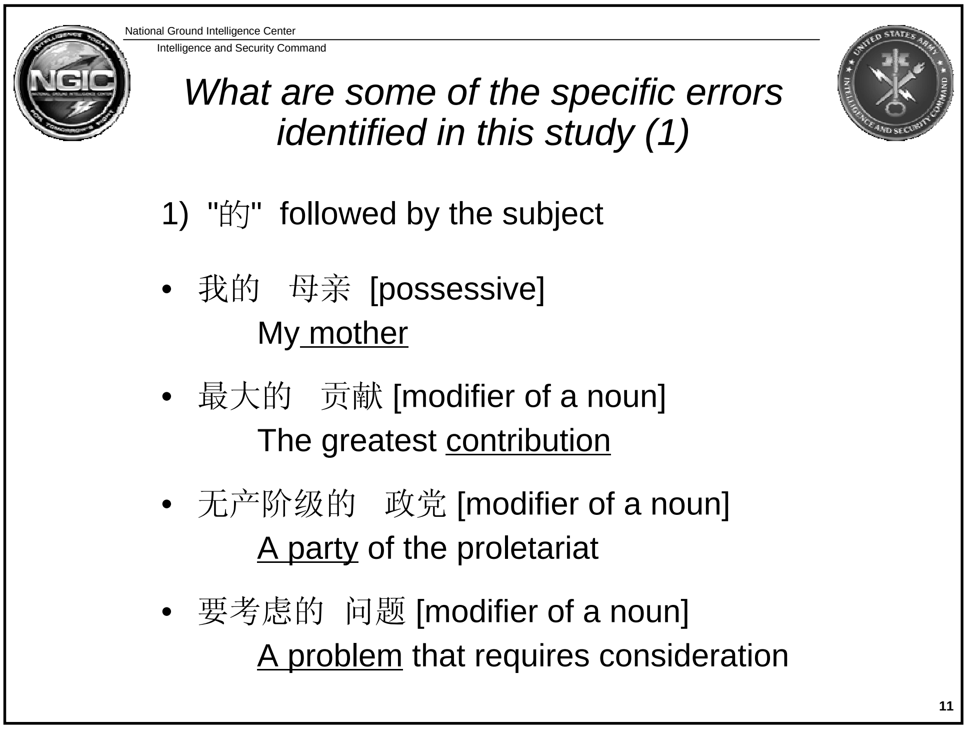
HT: Unite all relevant departments and units in the country.

(联合)	(国内)…	Unite	within country…
(联合国)	(内)….	United Nations	inside…

(Google) the United Nations all relevant departments and units

(Systran) in the United Nations each Department concerned and the unit

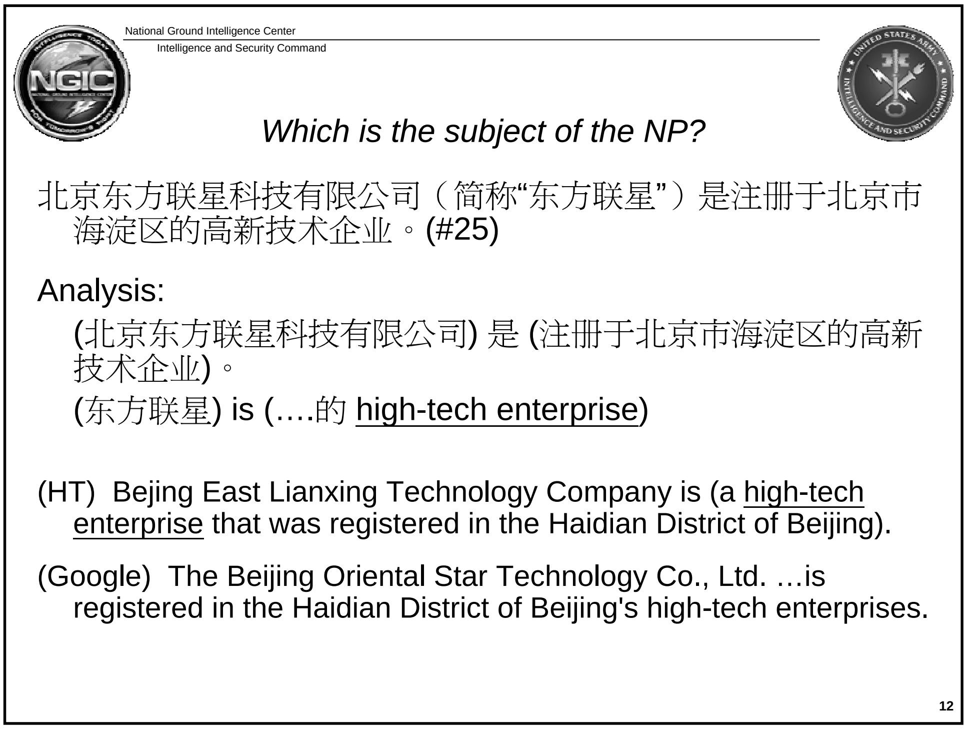
(MS) the relevant departments and units



What are some of the specific errors identified in this study (1)

1) "的" followed by the subject

- 我的　母亲 [possessive]
  My mother
- 最大的　贡献 [modifier of a noun]
  The greatest contribution
- 无产阶级的　政党 [modifier of a noun]
  A party of the proletariat
- 要考虑的 问题 [modifier of a noun]
  A problem that requires consideration



Which is the subject of the NP?

北京东方联星科技有限公司（简称“东方联星”）是注册于北京市海淀区的高新技术企业。(#25)

Analysis:

(北京东方联星科技有限公司) 是 (注册于北京市海淀区的高新技术企业)。

(东方联星) is (….的 high-tech enterprise)

(HT) Bejing East Lianxing Technology Company is (a high-tech enterprise that was registered in the Haidian District of Beijing).

(Google) The Beijing Oriental Star Technology Co., Ltd. …is registered in the Haidian District of Beijing's high-tech enterprises.

Which is the subject of the NP?

东大系规范化的股份制企业，专业从事水煤浆锅炉及水煤浆制备机组的研 究、开发、生产与销售。(#26)

Analysis: 东大 is (…的 enterprise), specializes in (…的 research and development) .

(HT) Dongda is (a standard joint-stock enterprise), specializes in (the research, development, production, and sales of coal-water slurry boilers and coal-water slurry preparation units.)

(Google) East of standardized joint-stock enterprises, specialized in the coal slurry and coal slurry boiler unit of the research, development, production and sales.

What are some of the specific errors identified in this study (2)

1) "的" followed by the subject

2) Identifying relationships among related noun phrases

What is the relationship among related NPs?

华建集团是中国科学院直接投资成立的高科技企业，(#24)

Analysis: (华建集团) is **(**(中国科学院) …的 (高科技企业)**)**

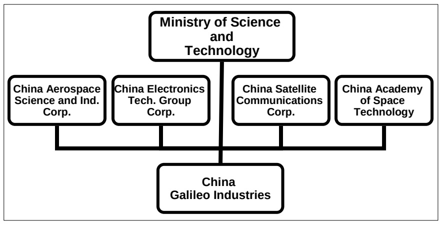
(HT) The Huajian Group is (<u>a high-tech enterprise</u> invested and established directly by the China Academy of Sciences).

(Google) Hua Jian Group is a direct investment in the establishment of the Chinese Academy of Sciences of the high-tech enterprises,

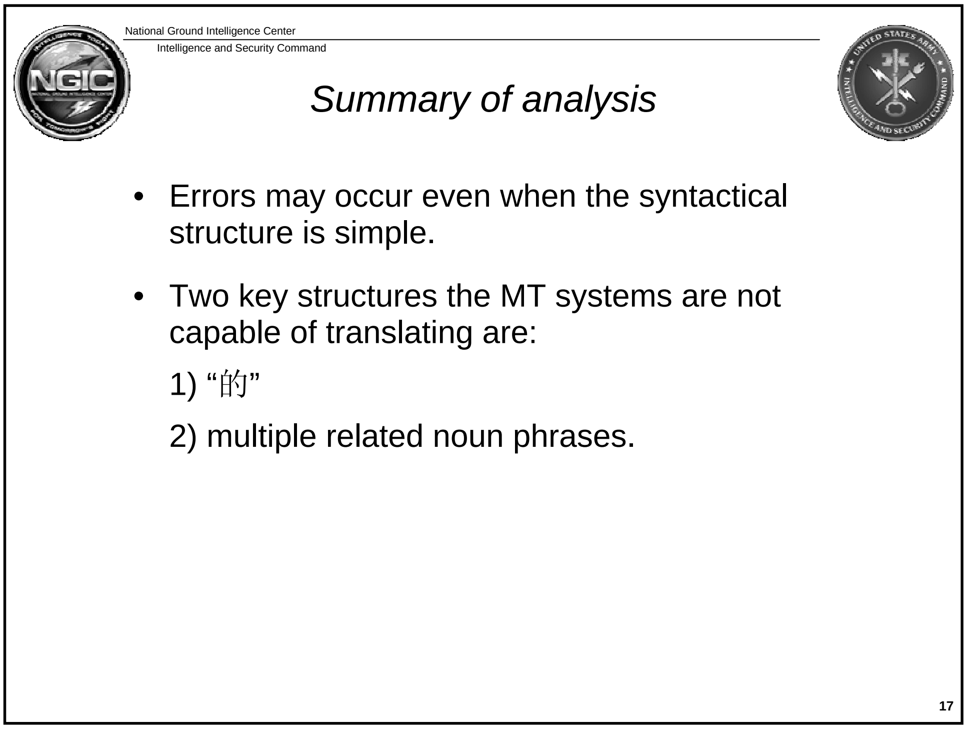
What is the relationship among related NPs?

伽利略卫星导航有限公司是在国家科技部的支持和领导下，由中国航天科工集团公司、中国电子科技集团公司、中国卫星通信集团公司和中国空间技术研究院四大股东共同组建的一家从事卫星导航系统技术开发、应用和运营服务的有限责任公司。(#5)

(HT) Sponsored by (the Ministry of Science and Technology). (China Galileo Industries) is (a jointly formed limited liability company) by the (China Aerospace Science & Industry Corp.), (China Electronics Technology Group Corp.), (China Satellite Communications Corp) and (China Academy of Space Technology)…..


Ministry of Science and Technology
China Aerospace Science and Ind. Corp.
China Electronics Tech. Group Corp.
China Satellite Communications Corp.
China Academy of Space Technology
China Galileo Industries

(Systran) The Galileo satellite navigation Limited company is under the National Technical department's support and the leadership, by the China Aerospace Science and Industry Corp. Company, China Electronic Technology Group Corporation, China Satellite Communications Corporation and the China Academy of Space Technology four major stockholders sets up one to be engaged in the satellite navigational system technology development, the application and the operation service Limited liability company together.

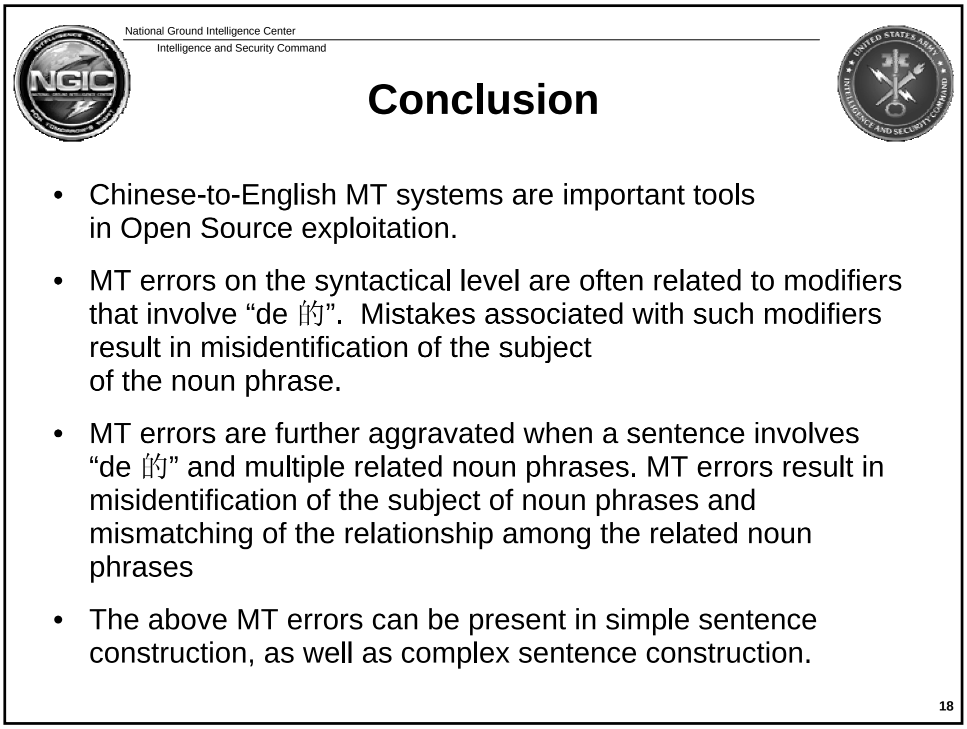


Summary of analysis

- Errors may occur even when the syntactical structure is simple.
- Two key structures the MT systems are not capable of translating are:

  1) “的”

  2) multiple related noun phrases.



Conclusion

- Chinese-to-English MT systems are important tools in Open Source exploitation.
- MT errors on the syntactical level are often related to modifiers that involve “de 的”. Mistakes associated with such modifiers result in misidentification of the subject of the noun phrase.
- MT errors are further aggravated when a sentence involves “de 的” and multiple related noun phrases. MT errors result in misidentification of the subject of noun phrases and mismatching of the relationship among the related noun phrases
- The above MT errors can be present in simple sentence construction, as well as complex sentence construction.

Recommendations

To MT users:

1. Misidentification of the subject and mismatching of the relationship among the noun phrases are two problems that seriously affect the accuracy of the MT output.
2. When in doubt, use more than one MT system to improve accuracy
3. Provide user feedback to MT developers.

To MT developers:

1. Address syntactical structures such as those identified in this study that current MT systems are unable to handle.
2. Seek user feedback.

人非圣贤, 孰能无过?

(HT) Ordinary people are not saints and sages. Who can make no mistakes?

(HT) Humans are not saints. Therefore, humans make mistakes.

(Google) Non-saints, without a Shuneng»

(Systran) The person non-saints and sages, who can not have?

(MS) The people of non - saint question cannot be too?