KOREAN NATIONAL CORPUS
AKA THE SEJONG CORPUS

- Constructed under the 21st Century Sejong project
- Size: 500 million eojuls; Goal: ~200 million eojuls
- Collection of corpora of modern Korean, International Korean, old Korean and oral folklore literature
- Accessible online in Korean

http://www.sejong.or.kr/user/main.do

LO JIA YI ERIN U1330845D

CREATION

- 2000: Subdivided the division for efficiency
- Primary Corpus Construction
  - Deals with the corpus of the modern South Korean language with various annotations
  - Raw, grammatically tagged, parsed and semantically tagged corpus
- Special Corpus Construction
  - Includes various types of corpus classified according to time, region and language
  - Transcribed colloquial expressions, parallel corpus, International Korean, historical data

ANNOTATION

- Applied TEI (Text Encoding Initiative) P3 to the corpus in 1998
- All corpus documents consist of the TEI header and main text
    - Header: Bibliography, text category, history of computerization, record of correction etc.
- SGML for text encoding
- Plans to convert current data into TEI P5 for standardization
    - TEI P5 is based on XML

USAGE

- The Sejong treebank
  - Parsed sentences composed of 150,000 words
- Sejong Morph Tagged Corpus
  - Morphologically analyzed written and spoken Korean
  - Yonsei University, Korea University
  - ~10 million words (written); ~30,000 words (spoken)
  - (1) geoleosseo ("walked") :
    geod_VV + eoss_EP + eo_EF
    walk PAST DECLARATIVE

References

- Junho, J.P., Jo, Y. and Shin, H. (2010). *The KOLON System: Tools for Ontological Natural Language Processing in Korean.* Paper presented at the Pacific Asia Conference on Language, Information and Computation. Retrieved February 2, 2015, from http://aclweb.org/anthology/Y10-1048.
- Kang, B. and Kim H. (2004). *Sejong Korean Corpora in the Making*. Paper presented at the International Conference on Language Resources and Evaluation. Retrieved February 2, 2015, from http://www.lrec-conf.org/proceedings/lrec2004/pdf/66.pdf.
- Kim, H. (n.d.). Korean National Corpus in the 21st Century Sejong Project. Retrieved February 2, 2015, from http://wenku.baidu.com/view/7bbc2f6825c52cc58bd6bec6.

ISRLN

Title	Korean National Corpus; Sejong Corpus
Full official name	Korean National Corpus
Resource Type	Primary text
Source/URL	http://www.sejong.or.kr/user/main.do
Format/MIME Type	text/SGML
Size/Duration	500,000,000 eojuls
Access Medium	Online
Description	The KNC is a collection of various corpora including that of modern Korean, International Korean, old Korean and oral folklore literature. It is constructed under the 21st century King Sejong project which was initiated by the government.
Version	?
Media Type	Text
Language	Korean
Resource Creator	The National Institute of Korean Language
Distribution	The National Institute of Korean Language
Rights Holder	The National Institute of Korean Language