# (U) Building Human-Language Technology

FROM: [redacted] and
Human Language Technology (S23)
Run Date: 09/07/2006

(U) **cor¿pus** \'korpəs..\ noun. *plural* : **corpo¿ra** \'korp(ə)rə\ ... A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language.
*-- Dictionary of Linguistics and Phonetics, 3rd edition, 1991.*

(S//SI) *Analysts:* Imagine a future where the tools you use are tailored to handle the complex SIGINT data that you process every day... a future where tools are developed to deal with the unique challenges you face ... a future where a commercial product from an outside vendor can be carefully evaluated using real operational data so that smart decisions can be made about spending Agency funds to provide you with technology that really works....

(U//FOUO) The Corpora Activity, an effort within the Human Language Technology Program Management Office (HLT PMO), is helping to make that future a reality. But, how are corpora related to technology? **Language corpora, annotated sets of linguistic data,** are not merely related, they are actually crucial to the research, development, and evaluation of any HLT tool. They **are the foundation on which the tools are built.**

(U//FOUO) The careful preparation of data sets that reflect Agency language challenges are especially important in the creation of tools that make their way to analysts' desks, because the tools need to be trained to deal with the material that inundates those desks. Acquiring, preparing, and disseminating corpora -- for voice, text and image, for any stage of HLT development -- is the primary task of the Corpora Activity.

(U) Outside of NSA, in the commercial and academic worlds, linguistic corpora are created using open-source, unclassified data. Qualified workers annotate, or mark up, the data to reflect some aspect of their content, like an interesting speaker or a particular foreign language. Such open-source data are instrumental in conducting foundational research to develop the mathematical algorithms that underlie HLT tools. These data sets are carefully controlled and marked for variables that the technology will be taught to recognize, such as gender, speaker, language, and other factor, and thus they are ideal for many research applications.

(U//FOUO) For the development and evaluation of HLT tools that Agency language analysts use, in addition to unclassified data sets, HLT researchers and developers need classified corpora. **The creation of these agency data sets poses unique challenges that are not found in the open-source world.** In the outside world, a company may pay to acquire whatever data they need, crafted to their exact specifications; within the agency, we must create corpora with the data available to us -- SIGINT intercept.

**SERIES:**
**(U) HLT**

1. Human-Language Technology in Your Future
2. For Media Mining, the Future Is Now!
3. For Media Mining, the Future Is Now! (conclusion)
4. 'Knowledge Discovery': Finding the Best Material
5. Human-Language Technology -- Everywhere
6. Dealing With a 'Tsunami' of Intercept
7. Building Human-Language Technology
8. Strangers in a Strange Land?

(U//FOUO) Keeping and distributing open-source data poses no problems in the outside world so long as licensing restrictions are obeyed, yet corpora based on SIGINT must be stored and shared in a way that carefully obeys policy and security restrictions. When annotating open-source data, a commercial company needs only to worry about the qualifications of the workers who annotate them; since our data are classified, any workers who annotate them, of course, must be highly qualified linguists, but also must have the appropriate clearances. These issues and others make the development of classified corpora particularly challenging.

(U//FOUO) But when there is a challenge, there is also an element of excitement. The work of annotating SIGINT language material may be difficult, but it can also be an interesting and intense diversity activity that may teach language analysts more about their own languages. It can be rewarding to analysts who want to help their mission by supporting the development of tools that will be tailored to their specific requirements. Diversity tours and details are available within the Research Directorate (R6), Analytic Automation Technologies (S202B1) and the HLT PMO (S23) itself.

(U) Acquiring and preparing linguistic corpora are essential basic steps for conducting the research and development of tomorrow's HLT products. Laying the groundwork with linguistic data that accurately reflects Agency issues will help bring the analysts' future closer to today