

PROPOSER INFORMATION PAMPHLET

(PIP)

FOR THE

**Research on English and Foreign Language EXploitation (REFLEX)
Broad Agency Announcement (BAA)**

BAA 04-01-FH

15 March 2004

BAA 04-01-FH
Research on English and Foreign Language EXploitation (REFLEX)
BROAD AGENCY ANNOUNCEMENT (BAA)

PROPOSER INFORMATION PAMPHLET (PIP)

I. GENERAL

The information provided in this pamphlet, in addition to that provided in the Federal Business Opportunities (FedBizOps) Announcement, BAA 04-01-FH, constitutes a Broad Agency Announcement (BAA) as contemplated in FAR 6.102 (d) (2) (i).

All administrative correspondence and questions concerning this BAA must be directed, in writing, to the administrative addresses, as follows:

Contracting Officers Representative: Roy L. Peters

Contracting Officer: Gloria M. Golden

Internet Web Site: <http://www.nbc.gov/reflex.html>

The Department of the Interior, National Business Center, Acquisition and Property Management Division, Southwest Branch, Fort Huachuca intends to use electronic mail for most technical and administrative correspondence regarding this BAA. Technical and contractual questions should include the originator's full name and return e-mail address in the text. Questions and answers will be posted to the solicitation home page, URL <http://www.nbc.gov/reflex.html>.

Written requests for information concerning this BAA may be sent by, as follows:

By facsimile:

(520)533-8954, addressed to ATTN: BAA 04-01-FH INFORMATION (Roy Peters)

By Email:

Roy_L_Peters@nbc.gov

By surface mail (USPS):

Department of the Interior
National Business Center
Acquisition and Property Management Division, Southwest Branch

PO Box 12924
ATTN: BAA 04-01-FH (BAA INFORMATION, Roy Peters)
Fort Huachuca, Arizona 85670-2924

By overnight delivery service:

Department of the Interior
National Business Center
Acquisition and Property Management Division, Southwest Branch
Building 22208, Augur Avenue
ATTN: BAA 04-01-FH (BAA INFORMATION, Roy Peters)
Fort Huachuca, Arizona 85613-6000

If email is not available, please direct questions to one of the above addresses. These requests must include the name, address, and phone number of a point of contact at the asking organization.

II. OBJECTIVES

The Intelligence Technology Innovation Center (ITIC) and other Defense, Intelligence Community, and Homeland Security agencies of the U.S. Government have an interest in research on algorithms, techniques, technologies, and methodologies for partially or fully automating aspects of the process of acquiring information from text documents and other natural language-based sources. Government information analysts need dramatic improvements over today's capabilities in order to be able to accomplish their analytic tasks in the future.

The National Business Center is soliciting research proposals to advance the process of converting information from documents and other language-based sources into forms that can best be used in the analytic process. These "documents" can be in English or one or more foreign languages; they could be written text, or spoken or image data that has been converted into text or text-like form via ASR or OCR; they could be in a range of genres; and they could be on a variety of topics or domains. There are two broad use cases where this research program's technology will be used.

- **Assistance to information analysts:** Automated tools will assist human analysts in judging, quickly and accurately, the utility of individual documents in English or one or more foreign languages, and of clusters of documents that are related topically or along certain other dimensions; and in identifying specific facts or other items of interest from single documents or sets/clusters of documents. A long-term goal is to allow the human analyst to be able to perform accurate information analysis from an automatically-produced English-language translation or from an automatically-produced, condensed, English-language textual or non-textual rendition (summary) of a document or document cluster, instead of from the foreign language original(s).
- **Input to automated analytic tools:** Automated technology will generate fully structured, language-independent representations of information as input to analytic tools; this representation would be based on unstructured information conveyed in a single document or a document cluster in English or one or more foreign languages. The analytic tools include various knowledge-base population or analysis tools, social network analysis tools, and the like.

II.A. SCOPE

Period of Performance: Phase 1 shall be a base period not to exceed 24 months after contract award and will be funded initially in FY2004, and incrementally in fiscal years thereafter. Phase 2 shall be a period not to exceed 12 months and be exercised as an option at the end of Phase 1 with another funding increment in a subsequent fiscal year. It is anticipated any subsequent rounds of this BAA will also have a similar Phase 1 and 2 approach, i.e., an initial base effort followed by an option period.

Individual Awards: Multiple individual awards are anticipated. Phase 1 and Phase 2 awards are expected to be in the range of \$100,000 to \$1,000,000 per year. The amount of the award will vary according to the type of effort undertaken. Individual offerors may submit multiple proposals.

The BAA will remain open for 3 years after the publication date. The first round deadline for submission of proposals is 30 April 2004.

II.B. RESEARCH AREAS

This Research Program is seeking to develop innovative algorithms, technologies, linguistic resources, and linguistic resource methodologies to address the program goals. It is envisioned that research in a number of areas may be needed to address program goals, including Machine Translation, Information Extraction, Semantic Analysis, and certain aspects of Text Summarization and Document Clustering. Additionally, a range of linguistic and knowledge resources is likely to be needed for English, Chinese, Arabic, Korean, and a broad range of languages less-commonly taught in the US; these resources include corpora, lexicons, ontologies, and basic Natural Language Processing (NLP) components. This solicitation suggests (but does not require) one possible progression of broad semantic types that capture increasingly richer amounts and types of information for addressing one or both of the use cases above.

The data to be used for testing and evaluation in this program will be provided by the Government or its representative to all program participants, in addition to limited training data of similar character to the testing and evaluation data. The specifics of the program evaluations for various tasks will be determined over the course of the program, but will include standard evaluations conducted by the National Institutes of Standards and Technology (NIST). All awardees are expected to participate in the evaluations relevant to their tasks, where available. For example, all contractors performing the tasks described in subsections 1.1, 2.1, or 3.1 will be expected to participate in Automated Context Extraction (ACE) information extraction evaluations (previous evaluations are described at <http://www.nist.gov/speech/tests/ace>), while full-text translation offerors will be expected to participate in NIST's MT Evaluation (<http://www.nist.gov/speech/tests/mt/index.htm>). These evaluations are in addition to any project-internal evaluation run by the offerors.

0 Program Parameters

Subsections 0.1 through 0.4 below address the language, media, domain, and genre parameters that apply to all program tasks in Section 1 Partial Semantic Content Processing: Entities, Section 2 Partial Semantic Content Processing: Relations, Section 3 Partial Semantic Content Processing: Events, Section 4 "Full" Document Content

Processing, Section 5 Enabling Resources, and Section 6 Evaluation Methodology Research (except where otherwise specified in the task description).

0.1 Languages

0.1.1 Major Program Languages

The languages to be investigated in the research program are as in Section 0.1.2 below. Proposals may address the tasks in Sections 1 through 4.1 for English only, or may address them for English and one or more of the other languages listed below. Offerors are expected to retain flexibility to address the Year 2 and Year 3 language, which is to be finalized 3 months prior to the beginning of the program year.

0.1.2 Language Schedule

L1: English (Year 1)

L2: Arabic (Year 1)

L3: Chinese (Year 1)

L4: Korean or other language to be determined (Year 2 and 3)

0.1.3 Less-commonly-taught languages (LCTL)

This program is also interested in techniques and enabling resources to allow for rapid development of the capabilities described in Sections 1 through 4, below, for a range of languages less commonly taught in the US, especially languages for which there are no or few computational linguistic resources. Sections 1.8, 2.6, and 4.2.2 address the development of specific NLP capabilities for LCTLs, while Section 5.1 addresses algorithm development, methodology, and the process of collecting and building enabling resources for LCTLs in support of those tasks.

0.2 Media

This research program is interested in addressing the information conveyed by language in any of several media and communication modes. Proposals addressing tasks in Sections 1 through 4 below that deal with data input in all the types described in Sections 0.2.1 and 0.2.2 are preferred.

The term 'document' is used throughout this solicitation to refer to any individual communication that originated in any of the allowed media, to include a single newswire, a single news story from broadcast news, a conversation, etc.

0.2.1 Text

Text documents in electronic form (such as ASCII or UNICODE). These documents may have some structure provided in XML.

0.2.2 Text derived from spoken or image data

Text that is generated from spoken data through Automated Speech Recognition (ASR) technology or from document image data through Optical Character Recognition (OCR)

technology. Research on improving ASR or OCR, however, is not in the scope of the program, and proposals in those areas will not be considered.

Of particular interest (but not required) is research on capabilities, as described in Sections 1 to 4 below (especially Section 1.2), that not only accept input from ASR or OCR as a (1-best) text string, but also take advantage of data structures used within ASR and OCR systems, such as word lattices, letter lattices, n-best lists, etc.

0.3 Domains and Genres

There will be no limitation on the subject areas or domains to be covered by the data. The genres to be covered include newspaper and newswire reporting, scientific and technical journal articles or abstracts, broadcast news (in ground truth and ASR transcript forms, such as the Topic Detection and Tracking (TDT) effort described at <http://www ldc.upenn.edu/Projects/TDT>), and (errorful) conversational speech transcripts produced by ASR systems (with very limited availability of ground-truth transcripts) of the type present in the FISHER, SWITCHBOARD, CALLHOME and CALLFRIEND corpora (<http://www ldc.upenn.edu/servlet/search.WebSearch?q=switchboard>, <http://www ldc.upenn.edu/servlet/search.WebSearch?q=callhome> and <http://www ldc.upenn.edu/servlet/search.WebSearch?q=callfriend>).

Proposals addressing all four of these general genre types will be preferred, however, the Machine Translation tasks and the LCTL tasks might not apply to the technical journal genre, depending on availability of corpora resources.

0.4 Applications

The technology to be developed under this program is expected to have a wide range of applications and uses. However, for the purpose of focusing the research and prioritizing the semantic types and areas of concentration, Social Network Analysis is identified as a specific primary customer application. While no work on Social Network Analysis technology per se is allowed in this solicitation, the offeror should expect that the outputs of the technologies described in Sections 1.1, 2.1, 3.1, and 4.1 will be used by existing and future Social Network Analysis tools.

Social Network Analysis (SNA) addresses the creation and analysis of a network-based representation of Entities of interest, along with specific relationships that are believed to exist between or among those Entities. Temporal, locative, and other attributes of relationships may also be of interest. The various relationships may be directly expressed in text, or may be inferred from Relations or Events (see below) or from other elements or characteristics of text.

1 Partial Semantic Content Processing: Entities

The first program task involves providing various capabilities for processing information about Entities. Entity types of interest include people, organizations, geopolitical entities (such as countries), locations, facilities, weapons, vehicles, certain substances, as well as definite temporal references. Other Entity types may be defined over the course of the program.

For the purpose of this program, a word or phrase in a document is not an Entity, but rather a *mention* that may aggregate with other mentions that refer to the same specific real world object (abstract or physical), and, taken together, the aggregated mentions model an Entity. Of primary importance to this program are these document-wide aggregated models of Entities (just referred to as Entities in this document), while the individual mentions of an Entity are still of secondary importance.

The information to be handled by the proposed systems includes (as relevant) the names of the Entities, semantic type and subtype of the Entities (see <http://www ldc.upenn.edu/Projects/ACE/docs/EDT-Guidelines-V2-5.pdf>), certain attributes of the Entities, and/or normalized forms of reference (such as standard date formats, for example as in <http://timex2.mitre.org>, or gazetteer indices for locations, to be specified or selected over the course of the program).

1.1 Entity Detection and Tracking (EDT)

For each Entity (of the semantic types of interest) referred to in a document, identify all mentions (whether names, noun phrases, pronouns, etc.) of the Entity, aggregate all these mentions into a single Entity object representation, determine the semantic type and subtype of the Entity, and normalize the attributes, names, etc. in ways suggested by the offeror and refined in consultation with the Government and other research groups working on this task.

1.2 EDT for Derived Text

Develop the capabilities described in Section 1.1, but assuming input in the form of ASR or OCR data structures such as word lattices, letter lattices, n-best lists, etc., as described in Section 0.2.2.

1.3 Name Handling

For certain Entities with names (people, places, and organizations), normalize the name representation to a standard form (to be proposed by the offeror, and to be finalized in consultation with the Government after award). For example, each portion of people's names shall be individually annotated and type-marked, using annotations such as GIVEN-NAME, FAMILY-NAME, PATRONYMIC, etc.

1.4 Entity Translation and Transliteration

For Entities with names that are found in non-English documents, translate or transliterate the names into English in a standard transliteration scheme, to be agreed

upon in consultation with the Government. For transliterated non-English names of Entities that are found in English documents, normalize the transliteration into a standard transliteration scheme to be agreed upon in consultation with the Government. Translate any non-name descriptors of the Entity into English, or represent that information in a language-independent representation formalism.

1.5 Entity-based English-language Summarization of Single Documents

Produce an English-language text summary of a single document, centrally focusing on the Entities in that document. The summary does not necessarily need to be narrative text, but any text does need to be in English (including Entity names, in a translated or transliterated form, as appropriate), regardless of the original language of the document. All proposals for this task must address L1 as well as at least one of L2 through L4 (as defined in Section 0.1.2 above)

1.6 Entity-based Document Clustering

Cluster a heterogeneous set of documents based on Entities that have some feature in common and/or based on the intersection of multiple Entities that meet some shared criteria. This task could involve matching names and date/time expressions across languages, as well as matching different spellings, forms, or transliterations of a name within a language. This task does not involve clustering based on any other words, terms, or concepts, but could involve all Entity types, including locations and date/time expressions.

1.7 Entity-based Summarization of Document Clusters

Produce an English-language text summary of a multi-document cluster resulting from the task in Section 1.6, centrally focusing on the Entities in that document. The summary does not necessarily need to be narrative text, but any text does need to be in English (including Entity names), regardless of the original language of the document(s). All proposals for this task must address L1 as well as at least one of L2 through L4.

1.8 LCTL Capabilities

Develop algorithms, technologies, and/or methodologies to rapidly develop the sorts of capabilities described in Sections 1.1 through 1.7 for LCTLs (as described in Section 0.1.3), assuming limited corpora and human resources. The offeror needs to explicitly identify the extent of any resources the offeror believes will be required for the proposed approach that exceed the resources described in Section 5.1.

2 Partial Semantic Content Processing: Relations

The second program task involves providing various capabilities for processing information about Relations between Entities. For the purpose of this program, Relations are as defined in <http://www ldc.upenn.edu/Projects/ACE/docs/RDC-Guidelines-V3-6.pdf>; Entities are as defined above. The information to be handled by the proposed systems includes (as relevant) the semantic type and subtype of the Relation, certain attributes of the Relations, and/or normalized forms of reference (such as standardized time/date or location formats, to be defined or selected during the course of the program, by the Government and program participants).

Offerors may either couple proposals for this task with their own Entity work (the task in Section 1), treat Relation arguments as unanalyzed strings with a semantic type, or assume an external Entity-level processor.

For the purpose of this program, a word or phrase in a document is not a Relation, but rather a *mention* that may aggregate with other mentions that refer to the same specific real world relation (abstract or physical) and entities, and, taken together, the aggregated mentions model a Relation. Of primary importance to this program are these document-wide aggregated models of Relation (just referred to as Relations in this document), while the individual mentions of a Relation are still of secondary importance.

2.1 Relation Detection and Tracking

For each Relation (of the semantic types of interest) referred to in a document, identify all mentions (whether prepositions, verbs, juxtaposition in a N-N compound, etc.) of the Relation, aggregate all these mentions into a single Relation object representation, determine the semantic type and subtype of the Relation, and normalize the attributes, etc. in ways suggested by the offeror and refined in consultation with the Government and other research groups working on this task.

2.2 Relation Translation

For Relations with strings as attributes or arguments (“mentions”, such as specifics describing Relation subtypes) that are found in non-English documents, translate those strings into English, or render the information in a language-neutral representation formalism. At least one of L2 through L4 must be included in proposals addressing this task.

2.3 Relation-based English-language Summarization of Single Documents

Produce an English-language text summary of a single document, centrally focusing on the Relations and Entities in that document. The summary does not necessarily need to be narrative text, but any text does need to be in English (including Entity names or Relation attributes or arguments), regardless of the original language of the document. All proposals for this task must address L1 as well as at least one of L2 through L4.

2.4 Relation-based Document Clustering

Cluster a heterogeneous set of documents based on Relations and Entities in common. This task may build on the offeror's proposal for the task in Sections 1.1, 1.4, and 1.6, or may involve clustering on string representations of entities. This task does not involve clustering based on words, terms, or concepts, but may involve all Relation and Entity types.

2.5 Relation-based Summarization of Document Clusters

Produce an English-language text summary of a multi-document cluster resulting from the task in Section 2.4, centrally focusing on the Entities and Relations in that document. The summary does not necessarily need to be narrative text, but any text does need to be in English (including Entity names), regardless of the original language of the document(s). All proposals for this task must address L1 as well as at least one of L2 through L4.

2.6 LCTL Capabilities

Develop algorithms, technologies, and/or methodologies to rapidly develop the sorts of capabilities described in Sections 2.1 through 2.5 for LCTLs (as described in Section 0.1.3) assuming limited corpora and human resources. The offeror needs to explicitly identify the extent of any resources the offeror believes will be required for the proposed approach that exceed the resources described in Section 5.1.

3 Partial Semantic Content Processing: Events

The third program task involves providing various capabilities for processing information about Events. For the purpose of this program, Events are loosely defined to include multi-place relations (more than 2 arguments), or predicate/proposition heads, that describe a change in the world's state. Events in this sense are frequently (but not always) expressed in English as verbs or deverbal nominalizations. Thus, events are defined here to be fairly fine-grained. The exact definition of Events for the overall program will be determined over the course of the program.

In representing the arguments of an Event, the offerors may either couple proposals for this task with their own Entity task proposal (the task in Section 1), or only treat arguments as strings with a semantic type, or assume an external Entity-level processor.

The information to be handled by the proposed systems includes (as relevant) the semantic type and subtype of the Event (see <http://www ldc.upenn.edu/Projects/ACE/docs/EnglishEDCV2.0.pdf>), certain attributes of the Event preferably including temporal and locative information (as available, and normalized according to a standard to be selected by the Government and contractors during the course of the program, such as <http://timex2.mitre.org>), and argument information labeled with semantic role and filled by either Entity objects or strings.

For the purpose of this program, a word or phrase in a document is not an Event, but rather a *mention* that may aggregate with other mentions that refer to the same specific real world event (abstract or physical) and entities, and, taken together, the aggregated mentions model an Event. Of primary importance to this program are these document-wide aggregated models of Events (just referred to as Events in this document), while the individual mentions of an Event are still of secondary importance.

3.1 Event Detection and Tracking

Identify all mentions of Events of the semantic types of interest in documents, determine the semantic type of each Event, collect all the mentions of the Event within one document into a single Event object representation, identify the semantic roles, and fill each semantic role with either a string or a representation of an Entity (or other object).

For each Event (of the semantic types of interest) referred to in a document, identify all mentions (whether verbs, deverbal nominalizations, etc.) of the Event, aggregate all these mentions into a single Event object representation, determine the semantic type and subtype of the Event, and normalize the attributes, etc. in ways suggested by the offeror and refined in consultation with the Government and other research groups working on this task.

Note that the ACE program will not be evaluating Event Detection and Tracking in 2004, but will start in the Fall of 2005.

4 “Full” Document Content Processing

In addition to the partial semantic processing capabilities described in the sections above, proposals are invited to address more complete content processing for full documents, as specified below.

4.1 Content Representation and Novel Semantic Analysis Algorithms

The primary goal of this task is to produce a structured language-neutral information representation from unstructured language data, to be used by a range of possible analytic applications, eliminating the need for language-specific capabilities for those applications. The primary such analytic application for this program is Social Network Analysis, as discussed in Section 0.4; other applications that may benefit from this representation include knowledge base population with a range of associated analysis tools. A secondary goal of this task is to investigate whether the representation may be of use as an intermediate structure in various NLP applications or processes.

Define a data structure, logical form, or other representation language or formalism that captures the semantic content of documents; develop and validate an algorithm for producing this representation for previously unseen text. (Associated manual construction of corpora that are annotated with this representation, for the purposes of training, concept validation, and testing, is covered in Section 5.2.1). This representation may parallel the flow of information in the document (e.g., line up sentence-by-sentence), may be a cumulative unordered network of all Entities, Relations, Events, and other elements of content, or both.

Proposals for research on word-sense disambiguation (WSD) of running text, for at least 2 of L1 through L4 or a LCTL (to be agreed in consultation with the Government), are acceptable under this task.

While the eventual long-term goal is a universal language-independent representation that captures much of the information content in text, for a range of applications, proposals are expected to be realistic, yet bold, in their expectations of what will be achieved during the course of the proposed effort. Proposals should clearly indicate whether the proposed representation is limited or unlimited along each of the following dimensions:

- language independence (for example, all terms in the representation language are drawn from a language-neutral ontology or other inventory)
- semantic depth
- semantic breadth (what is covered – entities? propositions? relations? speaker attitudes? modality? polarity? any inferences? etc.)
- domain independence
- application range (what kinds of analytic or NLP applications can rely on the representation)

It is not expected that the proposed representations be unlimited along each of these dimensions, but the proposal must indicate the assumptions along each dimension.

4.2 Full-Text Machine Translation Algorithms

Develop novel algorithms or technologies for improved full-text (i.e., entire document) translation, into English. Particular emphasis is expected on accurate translation of the sorts of Entities described in Section 1. The evaluation will include Entity translation/transliteration accuracy, as well as standard full-text translation evaluation in NIST's MT evaluation (<http://www.nist.gov/speech/tests/mt/index.htm>).

Proposals that explore hybrid algorithms that incorporate elements from various existing approaches are also encouraged, as are proposals that incorporate or build on the types of processing described in Sections 1 through 4.1 above.

Each proposal under this task must specify the subtask in Section 4.2.1 and/or the subtask in Section 4.2.2.

4.2.1 Full-Text Translation for the Major Program Languages

Validate algorithms by developing MT systems for translating at least 2 of L2 through L4 into English.

Proposals that suggest translation improvement primarily by increasing resource sizes (corpora, lexicons, etc.) are not encouraged.

4.2.2 Full-Text Translation for LCTLs

Of additional interest are approaches to translation that primarily require only limited resources, of the size and type described in Section 5.1.3. Evaluation on this task will involve building an MT system from those resources in a rapid timeframe for a LCTL (to be specified by the Government).

5 Enabling Resources

In support of the research goals of this program and the specific tasks described in Sections 1 through 4, proposals are solicited to build resources and to research resource development methodologies or technologies as described below.

All of the resources to be produced under this task must have all Intellectual Property Rights issues resolved, allowing distribution of these resources to any research organization or US Government element upon execution of appropriate agreements or licenses, either freely (preferred) or for a very modest fee.

5.1 LCTL

In addition to the thorough development of capabilities described in Sections 1 through 4 for Languages L1 through L4, proposals addressing technology and methodology for rapid development of basic enabling resources for a wide range of possible LCTLs are solicited.

5.1.1 Algorithm and Methodology Research

Develop language-independent algorithms, techniques, and methodologies to support rapid development of the basic resources described in Section 5.1.3 for any arbitrary language with a written form. Corpus-based unsupervised and lightly-supervised methods are acceptable, as are lightweight elicitation methodologies from untrained native speakers or other generally available (in the US) informants. Combinations of various techniques and methodologies are encouraged. Prototypes and other results of this work shall be made available to any contractors in this program who are working on the tasks described in 5.1.2 (which could provide valuable beta-testing).

5.1.2 Production of LCTL Text Resources

This subtask involves producing a set of basic resources that could be used for a range

of foreign language processing applications, for a number of languages. The production rate should be at least 7 languages per year, with the languages to be decided by the Government each year, in consultation with the offeror.

The quality level that is expected under this task must be sufficient to support a rapid start on developing initial capabilities of the sorts described in Sections 1 through 4.

Unicode UTF8 is the required encoding (for those languages for which Unicode is defined) for delivered resources, with annotations in XML.

Each offeror addressing this task is expected to cover all the resources specified in Section 5.1.3; offerors are encouraged to team if necessary to address the complete list.

Note that collection of corpora is opportunistically based on found data, so the processing elements described below may need to address the domains and/or genres of text that are found, not only general news text. For the same reasons, the requirements in Sections 0.2 and 0 do not necessarily hold for this task.

5.1.3 LCTL Text Resources

- Monolingual text corpus of at least 250,000 words. This corpus should include news and other genres; however, religious scriptures and government legislation should not be used to meet the minimum size requirement, but may be included, where found, as supplementary corpora.
- Parallel bilingual (with English) text corpus of at least 250,000 words, different from the monolingual corpus. The bulk (175,000 words) of this text should be material that was originally written in the foreign language, then translated into English (this may need to be produced by the offeror, if it cannot be found). This corpus should include news and other genres; however, religious scriptures and government legislation should not be used to address the minimum size requirement, but may be included, where found, as supplementary corpora. The remaining 75,000 words shall be produced by translating a standard corpus of English text (selected by the Government, in consultation with the contractor) into each of the foreign languages.
- Bilingual lexicon, wordlist, or Machine-Readable Dictionary for the foreign language and English, of at least 10,000 headwords or lemmas
- Fonts and codeset converters as needed to convert from the most common codesets to UTF8
- Sentence segmenter
- Word segmenter/tokenizer
- Part of Speech (POS) tagset and tagger, and (for at least 3 languages per year) a small manually-annotated corpus of running text sufficient for evaluation of the tagger

- If relevant: Morphological analyzer and (for at least 2 languages per year) a small manually-annotated corpus of running text with each morpheme and lemma identified, where the lemma forms are as found in the lexicon/wordlist/dictionary.
- Named Entity tagger (for People, Organizations, Locations, and Times/Dates, but with the final tagset to be determined in consultation with the Government) and an annotated 100,000 word corpus (could be either the monolingual or the foreign-language part of the bilingual corpus above)
- Person name transliterator into English
- A narrative descriptive grammar for the language in a form to be proposed by the offeror, to include a description of the components of people's names.

5.2 Semantic Resource Construction

5.2.1 Semantically Annotated Corpora

Construct a gold-standard corpus of text from one or more of languages L1 through L4 (see paragraph 0.1.2, above), or an LCTL, that is semantically annotated and validated for accuracy and consistency, to the degree feasible. The nature of the semantic annotation needs to be clearly specified in the proposal, as well as its expected benefit and application (an interlingua for MT, a Proposition Bank for producing event-level structured output, etc.) The proposal will also clearly specify the assumptions made along each of the dimensions specified in Section 4.1.

Specific measures of annotator consistency or annotation reproducibility and reliability must be proposed. Both stand-off and imbedded XML-based annotation formats are acceptable.

Corpora annotated for Information Extraction tasks, as described in Sections 1.1, 2.1, and 3.1, fall under Section 5.2.4 below.

5.2.2 Wordnets

Construct a wordnet for L2 or an LCTL, or construct improved wordnets for L3 or L4. Provide mappings from each synset to an equivalent synset (if any) in Princeton University's English-language WordNet™ 2.0 (as described at <http://www.cogsci.princeton.edu/~wn/>). The resulting resources are expected to be freely available to the research community and US Government.

5.2.3 Ontologies

Construct a language-neutral, general-domain ontology for supporting tasks described in Sections 1 through 4 or 5.2.1. Construct a lexicon that maps word senses (or word forms) in any of L1 through L4 or a LCTL (to be agreed on with the Government) into this ontology. Provide some measures and validation of completeness, consistency, and/or reproducibility of the ontology. Proposals for true ontologies that have significant semantic relations in addition to taxonomic ones are preferred. The requirement is for

an ontology that will support language processing tasks across languages. However, the proposal should demonstrate that the offeror understands and addresses culture-specific ontological bias issues.

This task addresses the actual construction of ontologies, necessary tools, and associated lexical resources for linguistic or NLP applications, and may be accompanied by additional work on algorithm development for (semi-) automatic generation of true ontologies or lexicons linked to them.

5.2.4 Annotated Extraction Corpora

Construct annotated corpora for languages L1 through L4 to support development and evaluation for all of the tasks described in Sections 1.1, 2.1, and 3.1. The corpus sizes are expected to be in the range of 250,000 words per language per year (L4 starts in Year 2). The exact specification of the annotations will evolve over the course of the Program, but example guidelines are referenced in the corresponding sections above. Robust annotation procedures involving double annotation with adjudication for the entire corpus are required. The offeror shall include detailed discussion of annotation methodology and consistency evaluation for ensuring high annotation quality.

6 Evaluation Methodology Research

In support of the research goals of this program, proposals are solicited to develop and assess the merits of novel evaluation methodologies, for evaluation of the technologies described in Sections 1 through 5. This BAA is not soliciting proposals for actually running any of the Program-wide evaluations, however.

III. PROPOSAL PREPARATION INSTRUCTIONS:

This announcement is an expression of interest only and does not commit the Government to pay for proposal preparation costs. The cost of preparing proposals in response to this BAA is not considered an allowable direct charge to any resulting contract or to any other contract. However, it may be an allowable expense to normal bid and proposal indirect costs as specified in FAR 31.205-18. If a subcontract(s) with a Federally Funded Research and Development Center (FFRDC) is proposed, offerors are reminded of the limitations in their use (see FAR 35.017) and must provide documentation in the proposal that work is not otherwise available from the private sector. Each proposal shall reflect the potential for commercial application and the benefits expected to accrue from this commercialization. Technology transition efforts, partners, or plans should be explicitly discussed. All data an offeror deems pertinent to the proposal shall be submitted with the proposal.

Discussions with any of the points of contact shall not constitute a commitment by the Government to subsequently fund or award any proposed effort. Only Contracting Officers are legally authorized to commit the Government.

PROPOSAL SCOPE: Offerors may submit proposals covering a base period of performance (Phase 1) not to exceed 24 months and, as needed, an optional effort (Phase 2) covering an additional period of performance not to exceed 12 months. Proposals submitted with optional periods of performance will be evaluated on the basis of the base period and all options. An individual proposal must identify explicitly which Research Areas tasks are being addressed, by identifying section (X), subsection (X.Y), **and** subsubsection (X.Y.Z, where given) numbers between 1.1 and 6. However, **each proposal must show a coherent project direction in the range of tasks covered.** Multiple proposals, each covering a specific research direction (potentially tasks from multiple Research Areas sections and subsections, etc.) may be submitted by a single offeror. However, for proposals addressing multiple Research Areas sections, subsections, or subsubsections, separate technical, cost, and deliverable information must be provided, as specified below.

CLASSIFICATION: Unclassified proposals **ONLY** will be accepted and evaluated.

TRAVEL and PROGRAM MEETINGS: Offerors will be expected to participate in various technical exchanges plus coordination and planning activities with the Government and other participants. Offerors are expected to participate in an initial 3-day kickoff meeting, followed by 2- to 3-day program/PI meetings every six months. Additionally, offerors should allow for 2 additional 1-day meetings per year. For costing estimation purposes, offerors may assume that many of the meetings will be in the general Washington, D.C. area. Offerors may propose travel to relevant annual government hosted or other evaluation meetings.

PERSONNEL REPLACEMENT: Any technical personnel who, during the performance of the contract, are assigned by the contractor to replace the technical personnel identified by the contractor in the technical proposal (or during any negotiations) for work on the contract shall possess at least the same technical qualifications. They shall also be capable of assuring satisfactory performance of the work required by the resulting contract. The Government reserves the right to review résumés of any replacements or substitutes for key personnel named in the contractor's proposal.

COLLABORATION: In order to help the program make maximum progress, contractors will be expected to share detailed technical information about any techniques that they develop or use.

FORMAT

Volumes: Proposals shall consist of two volumes: Volume I – Technical and Management and Volume II - Cost. The page format shall be 12 point or larger type, single-spaced, one inch margins, single sided, 8.5 by 11 inch pages. The page limitations for the Technical and Management Volume include all information (i.e., figures, tables, graphics, charts, indices, photographs, foldouts, etc.) and are given for each section in the Volume. Unnecessarily elaborate brochures or presentations beyond that sufficient to present a complete and effective proposal are not desired. Offerors shall submit an original and a paper copy of each proposal and an electronic copy in Microsoft Word for Windows (Microsoft Excel for any spreadsheet submissions) format on 3.5 inch 1.4MB floppy diskette or CD-ROM by the closing date. Proposals exceeding the maximum total length WILL NOT be considered.

Electronic Proposal Format: Electronic proposals shall be made using Microsoft Word (Version 6.0 or earlier) and Excel for Windows applications (compatible with Windows 95 through 2000 or Windows XP). Acrobat Portable Data File Format (PDF) is also an acceptable file format, provided these files are created with Version 5.5 or earlier. Diskettes or CDs shall be clearly labeled, referencing BAA 04-01-FH, marked with the proposer's organization and proposal title (short title recommended). Hard copy and electronic media must be submitted together. If using Microsoft Word, embed any graphics used. Microsoft Word documents with graphics as separate files are **NOT** acceptable. Volumes I and II must each be contained within a single electronic file, i.e., a single file containing all of Volume I and a second single file containing all of Volume II. All electronic media must be verified virus-free by using an up-to-date, reputable virus detection utility, such as Norton or McAfee anti-virus software, and so noted on the diskette or disk label.

Number of Copies: 2 Copies of each proposals shall be submitted (one must contain original signatures) and an electronic copy of both Volumes I and II.

Information or data contained in a full proposal deemed proprietary by the offeror should

be clearly marked. The offeror must mark the proposal with a protective legend in accordance with FAR Part 15.6, Use and Disclosure of Data, (modified to permit release to outside evaluators retained by either ITIC or the Department of the Interior, National Business Center, Acquisition and Property Management Division, Southwest Branch, Fort Huachuca) if protection is desired for proprietary or confidential information.

Volume I – Technical and Management

Each section shall begin on a new page and shall be limited in length as {indicated}. Foldouts will be counted as a single page and must be no larger than 11 x 17 inches. The number of foldouts shall not exceed five in number, and used for tables, graphics, and similar material. Offerors are encouraged to submit concise, but descriptive, technical proposals.

Cover Sheet: {1 page} The Cover Sheet provided as Attachment 1 of this document shall be completed by the offeror and submitted with the proposal. Include the cover sheet at the beginning of the file containing Volume I. All information requested must be provided. The CAGE, DUNS/CEC, and TIN codes provided shall be those of the offeror and not of the principal place of performance, if the two are different.

Part I: Summary of Proposal. This section shall provide an overview of the proposed work, as well as introduce associated technical and management issues.

- (a) {2 pages} In a manner of the offeror's choosing, provide an executive summary, including a succinct description of the uniqueness and benefits of the proposed project, a brief discussion of the technical rationale, technical approach, and constructive plans for accomplishing the technical goals
- (b) {1 page} A list of the **Research Areas section, subsection, and subsection numbers** that correspond with the tasks being addressed in this proposal. Identify the **languages** being addressed.
- (c) {1 page} Summary of innovative claims for the proposed research
- (d) {2 pages} A clearly defined organization chart for the program team with a listing of key personnel.

Part II: Detailed Proposal Information. This part shall provide more detailed, in-depth discussion of the proposed research. Specific attention must be given to addressing both the risks and payoffs of the proposed research making it desirable to pursue. This Part shall provide:

- (a) Statement of Work (SOW), describing the effort's scope, the specific tasks to be performed, and their associated schedules and relationship to the program goals and associated Research Areas, described above. At a minimum, SOW shall

consist of the following sections:

{2 pages} Scope—a description of the overall objectives and goals and major milestones for the effort.

{5 pages per task, not to exceed 25 pages total} Task/technical requirements – a description of proposed tasks, representing the work to be performed, developed in an orderly progression and in enough detail to establish the feasibility of accomplishing the overall program goals. The overall effort should be grouped into tasks and identified in a work breakdown structure (WBS)-like numbering system. Proposed costs shall have a one-to-one correlation to this reporting structure, which shall be depicted in the cost volume

For each task, address at least the following:

- Task Title
- The Language(s) and Research Areas section/subsection/subsubsection number that the task is addressing (where possible, avoid more than one Research Areas subsection or subsubsection per task, although a Research Areas section, subsection, or subsubsection may be addressed by multiple proposed tasks).
- Technical Challenge – Diagnosis of the challenge and the associated risks.
- Technical Objective – A clear statement of what is to be produced and benefits if successful.
- Technical Approach – A description of the approach, the rationale for the approach, why the proposed technical approach is expected to achieve the stated goals within the proposed cost and time schedule, and proposed evaluation
- Comparison with other work – Highlight the uniqueness of the proposed work and differences between the proposed effort and current state-of-the-art (especially work under previous US Government-sponsored programs such as ACE, EELD, and TIDES). Identify the advantages and disadvantages of the proposed work with respect to potential alternative approaches.

(b) {2 pages} A graphic illustration that depicts major milestones and schedule of the proposed effort arrayed against the proposed tasks and time estimates.

(c) {4 pages} Show how past/current performance justifies an award in this technical area; specifically, include any recent results from standard NIST or other evaluations, on appropriate tasks. Include capabilities, related experience, facilities, techniques, or unique combinations of these, which are integral factors for achieving proposal objectives; and references who can verify present and past performance. Include contract number(s), points of contact, and telephone numbers. Proposer is responsible for accuracy and currency of references' information.

- (d) {1 page per person} List of key personnel, concise summary of their qualifications, and discussion of the offeror's previous accomplishments and work in this or closely related research areas. Indicate the level of effort to be expended by each person during each contract year and other (current and proposed) major sources of support for them and/or commitments of their efforts.
- (e) {1 page} If any portion of the research is based on the use of Government-owned resources of any type, the offeror shall specifically identify the property or other resource required, the date the property or resource is required, the duration of the requirement, the source from which the resource will be obtained, if known, and the impact on the research if the resource cannot be provided. If no Government-furnished property is required for conduct of the proposed research, this section shall consist of a statement to that effect.
- (f) {2 pages} Deliverables, which should include demonstrations, associated with the proposed research, and any plans and capabilities to accomplish technology transition and commercialization. If relevant, plans for dual-use capability or technology transfer plans, such as plans leading to commercialization of technology developed as a result of these projects should be addressed in this subsection.
- (g) {1 page} Resources offered – any software or linguistic data that the offeror is willing to share with other Program sites
- (h) {3 pages} A management approach describing the overall plan to manage this effort, including brief discussions of total organizations, use of personnel, relationships among project/function/subcontractors, Government research and facility interface, and planning, scheduling and control practices. Discuss any teaming relationships, to include the programmatic relationship of team members; the unique capabilities of team members; the task responsibilities of team members; the teaming strategy among the team members; the key personnel from each team member along with the amount of effort to be expended by each person during each year
- (i) {1 page} A summary of any proprietary claims to results, prototypes, or systems supporting and/or necessary for the use of the research, results and/or prototype must be included. If there are no proprietary claims this section shall consist of a statement to that effect. In addition, and where appropriate, Volume I shall contain information concerning the identification and assertion of use, release, or disclosure restrictions and technical data or computer software previously delivered to the Government, as well as proposed licensing rights that will accrue to the Government.

Part III: Additional Information. This section has no page limits. This section shall include:

- (a) More detailed biographies or résumés for the Principal Investigator(s) and any other critical personnel

- (b) A brief bibliography of relevant technical papers or research notes (published and unpublished) which document the technical ideas upon which the proposal is based. Copies of up to three papers may be attached in their entirety. This material will be used at the discretion of evaluators, to enhance their understanding of relevant related work. It should not be used in place of the above-required information. When providing published work, the font/formatting requirements established for this volume do not apply.

Volume II - Cost.

Part 1: Cover Sheet. The Proposal Pricing Sheet at Attachment 2 shall be completed and submitted with each offer.

Part 2: Cost Summary. This section shall include:

- (a) A one-page cost and fee summary per year.
- (b) Detailed cost breakdowns by year and by tasks, correlated to Volume I, Statement of Work Task/Technical Requirements (cost detail reporting shall have a one-to-one correlation to the structure of the SOW and the WBS). The costs are to be broken down into appropriate accounting categories to help reviewers understand the proposed effort, and shall minimally include:
- Labor hours by labor category
 - Critical personnel assigned to this task (with labor category)
 - Subcontractors and consultants
- (c) Materials by vendor quotes and purchase history
- (d) Travel
- (e) Other direct and indirect costs

Part 3: Supporting Cost and Pricing Information. This part shall include supporting cost and pricing information in sufficient detail to substantiate the summary cost estimates in Part 2 above. Costs for subcontracts having 20% or more of the total value of the work must be substantiated to the same level of detail as the costs of the offeror.

Provide descriptions of each labor category referenced in the Cost Summary

HANDLING OF PROPOSALS: All proposals shall be handled as source selection information; contents will be disclosed only for the purposes of evaluation and only to members of the source selection panel.

The Government may use consultants and/or contractors to assist in evaluating the

proposals. These personnel will have signed, and will be subject to, the terms and conditions of non-disclosure agreements. By submission of its proposal, an offeror agrees that its proposal information may be disclosed to the aforementioned personnel for the limited purposes stated above. However, only the Government will make final award determinations under this BAA.

PROPOSAL SUBMISSION: Proposals are due on or before 4:00 PM, Mountain Standard Time, 30 April 2004 to the Department of the Interior, National Business Center, Acquisition and Property Management Division, Southwest Branch, Post Office Box 12924, ATTN: BAA 04-01-FH (Roy Peters), Fort Huachuca, Arizona, 85670-2924. Proposals which are hand-carried or delivered by overnight express carrier (such as Federal Express or United Parcel Service) are also due on or before 4:00 PM, Mountain Standard Time, 30 April 2004 to the Department of the Interior, National Business Center, Acquisition and Property Management Division, Southwest Branch, Second Floor, Building 22208, Augur Street, ATTN: BAA 04-01-FH (Roy Peters), Fort Huachuca, Arizona, 85613-6000. Proposals must be submitted in accordance with the requirements and procedures identified in the BAA and this PIP. To be considered, full proposals (in original, one copy, and electronic media) must be received. **Proposals submitted by fax or electronic mail are not acceptable and WILL NOT BE CONSIDERED. Proposals and/or proposal modifications received after the proposal submission closing date and time will be handled IAW FAR 15.208. Proposals not adhering to the form and format required by this BAA WILL NOT BE CONSIDERED.** The Government anticipates completing the evaluation process within 60 days after receipt of each proposal. Contract award will follow this evaluation and is estimated to be completed within 60 days.

IV. PROPOSAL SELECTION CRITERIA

Proposals will be selected through a technical/scientific/business decision process with technical and scientific considerations being most important. Evaluations will be performed using the following criteria listed in descending order of relative importance. Each period of the effort (base and options) must demonstrate these contributions independently and collectively. Proposals unresponsive to the Research Areas addressed in the BAA will not be fully evaluated and will not be considered for award.

- Overall scientific and/or technical merit, including technical approach, degree of innovation, understanding of the technical issues, and evaluation plan. If a proposal lacks overall scientific and/or technical merit, it will not be further considered for award.
- The proposed effort's potential contributions to the stated program goals.
- The offeror's capabilities, related experience, facilities, techniques, or unique combinations of these, which are integral factors for achieving proposal objectives; qualifications, capabilities, and experience of key personnel, and the offeror's record of present and past performance.

- Cost reasonableness and realism

Awards under this BAA will be made to responsible offerors on the basis of the evaluation criteria above and a BEST VALUE approach to the Government. Awards will be subject to the availability of funds. Awards may take the form of a procurement contract, grant, or cooperative agreement, depending upon the nature of the work proposed, the required degree of interaction between parties, and other factors.

The Government reserves the right to select for award all, some, or none of the proposals received and to incrementally fund any award instrument. The Government also reserves the right to fund all or any part of a proposal evaluated as eligible for award. Awards are subject to the availability of Government funds.

BAA 04-01-FH
Research on English and Foreign Language EXploitation (REFLEX)
BROAD AGENCY ANNOUNCEMENT (BAA)

PROPOSAL COVER SHEET

ATTACHMENT 1

BAA 04-01-FH
Research on English and Foreign Language EXPloitation (REFLEX)
Broad Agency Announcement (BAA)

Organization/Company	
CAGE Code	
DUNS/CEC Number	
TIN Number	
Type of Business	
Proposal Title and Identification Number	
Team Members/Type of Business	

Research Areas (list section, subsection, and subsection numbers)	
Language(s)	
Principal Investigator Name	
Mail Address	
Phone Number	
Fax Number	
E-mail Address	
Administrative Contact Name	
Mail Address	
Phone Number	
Fax Number	
E-mail Address	

Proposal Duration	
Base Year 1	\$
Base Year 2	\$
Option Year	\$
Total	\$

BAA 04-01-FH
Research on English and Foreign Language EXploitation (REFLEX)
Broad Agency Announcement (BAA)

PROPOSAL PRICING SHEET

ATTACHMENT 2

BAA 04-01-FH PROPOSAL PRICING SHEET

1. Company/Agency Information:

(Company/Agency Name)

(First Line of Address)

(Street Address)

(City)

(State)

(Zip Code)

2. Company/Agency Point of Contact Information:

(POC Name)

(POC Title)

(POC Telephone and FAX Nos. (Include Area Code))

(POC e-mail)

3. Type Of Contract (Check One):

_____ FFP

_____ CPFF

_____ CPAF

_____ FPI

_____ CPIF

_____ Other (Specify)

4. Proposed Cost (A + B = C):

4.a. Cost

4.b. Profit/Fee

4.c. Total

5. Performance:

5.a. Place (1) _____
(2) _____

5.b. Period (1) _____
(2) _____

6. Line Item Costs (List and reference the identification, quantity and total price proposed for each contract line item. A line item cost breakdown supporting this recap is required unless otherwise specified by the Contracting Officer. Continue on reverse, and then on plain paper, if necessary. Use same headings.)

6.a. Line No.	6.b. Identification	6.c. Quantity	6.d. Price	6.e. Prop. Pg. No.
_____	_____	_____	_____	_____
_____	_____	_____	_____	_____

7. Provide the Following (If available):

_____			_____		
(Name of Contract Administration Office)			(Name of Audit Office)		
_____	_____	_____	_____	_____	_____
(City)	(State)	(Zip Code)	(City)	(State)	(Zip Code)
_____			_____		
(Telephone (Include Area Code))			(Telephone (Include Area Code))		

8. Will you require the use of any Government property in the performance of this work? Yes No

9. Do you require Government contract financing to perform this proposed contract? Yes No
Type of financing (Check One) Advanced Payments Progress Payments Guaranteed Loans

10. Have you been awarded any contracts or subcontracts for the same or similar items within the past 3 years?
 Yes No (If "Yes," identify items(s), customer(s) and contract number(s) on reverse of form.)

11. Is this proposal consistent with your established estimating and accounting practices and procedures and FAR Part 31, Cost Principles? Yes No (If "No," explain on reverse of form.)

12. Cost Accounting Standards Board (CASB) Data (Public Law 91-379 as amended and FAR Part 30)

12.a. Will this contract action be subject to CASB regulations? Yes No
(If "No," explain on reverse of form.)

12.b. Have you submitted a CASB disclosure statement (CASB DS-1 or 2)? Yes No
(If "yes," specify in proposal the office to which submitted and if determined to be accurate.)

12.c. Have you been notified that you are or may be in compliance with your disclosure statement or cost accounting standards? Yes No (If "Yes," explain in proposal.)

12.d. Is any aspect of this proposal inconsistent with your disclosed practices or applicable cost accounting standards? Yes No (If "Yes," explain in proposal.)

This proposal is submitted in response to BAA 04-01-FH and reflects our estimates and/or actual costs as of this date and conforms to the instructions in FAR15.804-6(b)(1), and Table 15-2. By submitting this proposal, the offeror, if selected for negotiation, grants the contracting officer and authorized representatives(s) the right to examine, at any time before award, those records which include books, documents, accounting procedures and practices, and other data regardless of type and regardless of whether such items are in written form, in the form of computer data, or whether such supporting information is specifically referenced or included in the proposal as the basis for pricing, that will permit an adequate evaluation of the proposed price.

_____	_____	_____
13. Name (Typed)	14. Title	15. Company/Agency Name

_____	_____
16. Signature	Date