

Broad Operational Language Translation (BOLT)

Joseph Olive
Program Manager



Approved for Public Release, Distribution Unlimited

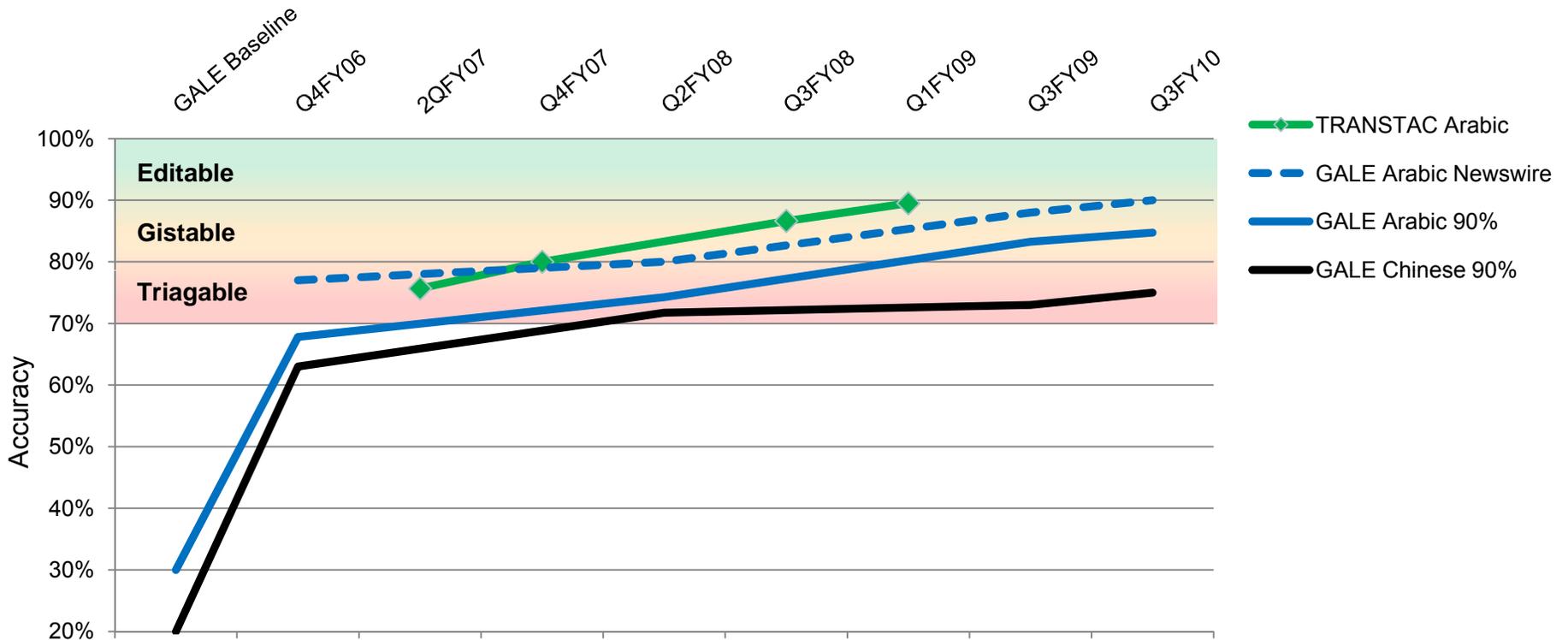
 **GOALS**

BOLT will develop natural language processing capability to enable:

1. Translation of informal language genres
2. Bilingual, multi-turn conversation (text and speech)



DARPA Program Performance



Formal language only (GALE) and single-turn restricted domain (TRANSTAC)

- Editable** – Document is clearly understandable; may require editing for a final product.
- Gistable** – Main idea, substance, or content of document is conveyed.
- Triagable** – Determine whether document is important or not for further analysis.



Key Military Applications

	Medium	Genre	Volume	
Translation	Newswire	Text	Formal	CENTCOM translates 5000 newswire/week
	Broadcast news	Speech	Formal	565 Arabic TV stations
	Blogs and newsgroups	Text	Semiformal	133x10 ⁶ sites
	Talk shows	Speech	Semiformal	657 total Arabic speaking radio stations
	Translate conversation	Speech	Informal	20x10 ⁶ cellular subscriptions (2009) & 1x10 ⁶ land lines (2008) in Iraq
	Translate email and messaging	Text	Informal	298x10 ⁶ internet users in China (2008)
Communication	Restricted speech to speech	Speech	Restricted	~400 checkpoints in Baghdad alone
	Coalition chat and public chat rooms	Text	Informal	31 MNF allies in Iraq not English speaking
	Conversation with coalition partners and local populace	Speech	Informal	US Army has only 2,092* linguist with a total force of 1,112,703* \$1B on contractors (2009)

Current Performance: ■ Good ■ Adequate ■ Poor

* All Components: active, reserve, Guard



Limitations In Present Technology

- Brittle Technology
 - Degrades rapidly for different topics, genres
- Translation of genres such as e-mail, messaging and conversation is inadequate
- Bilingual unconstrained chat and conversation also uses informal language
- Bilingual unconstrained chat and conversations degrade at each dialogue turn



BOLT- Meeting the Needs and Requirements

Flexibility - Ability to translate and retrieve information in multiple languages, regardless of genre or topic

- Analysis of segments longer than single sentences
- Robust syntactic and semantic analysis: Expanded analysis beyond lexical matches, use wildcard for missing or garbled information, co-reference resolution
- Robust language models: non-adjacent words' statistics, dependency modeling
- Handling disfluency
- Dialectal variations

Reliability – Ability to sustain multi-turn conversation and chat across different languages

- Verify accuracy of translation and correct errors
- Develop means of detecting and resolving uncertainty and ambiguity
- Incorporate human-machine dialogue to clarify and disambiguate input to reduce the probability of error



BOLT Program Organization

- Three Technical Areas
 1. Algorithmic Development and Integrated Systems
 2. Data Collection
 3. Evaluation

- Six Activities per Technical Area (Except Data)
 - A. Translation and Information Retrieval
 - B. Human-Machine Dialogue Systems
 - C. Human-Human Dialogue Systems
 - D. Arabic Dialect Translation
 - E. Grounded Language Acquisition
 - F. Basic Technologies



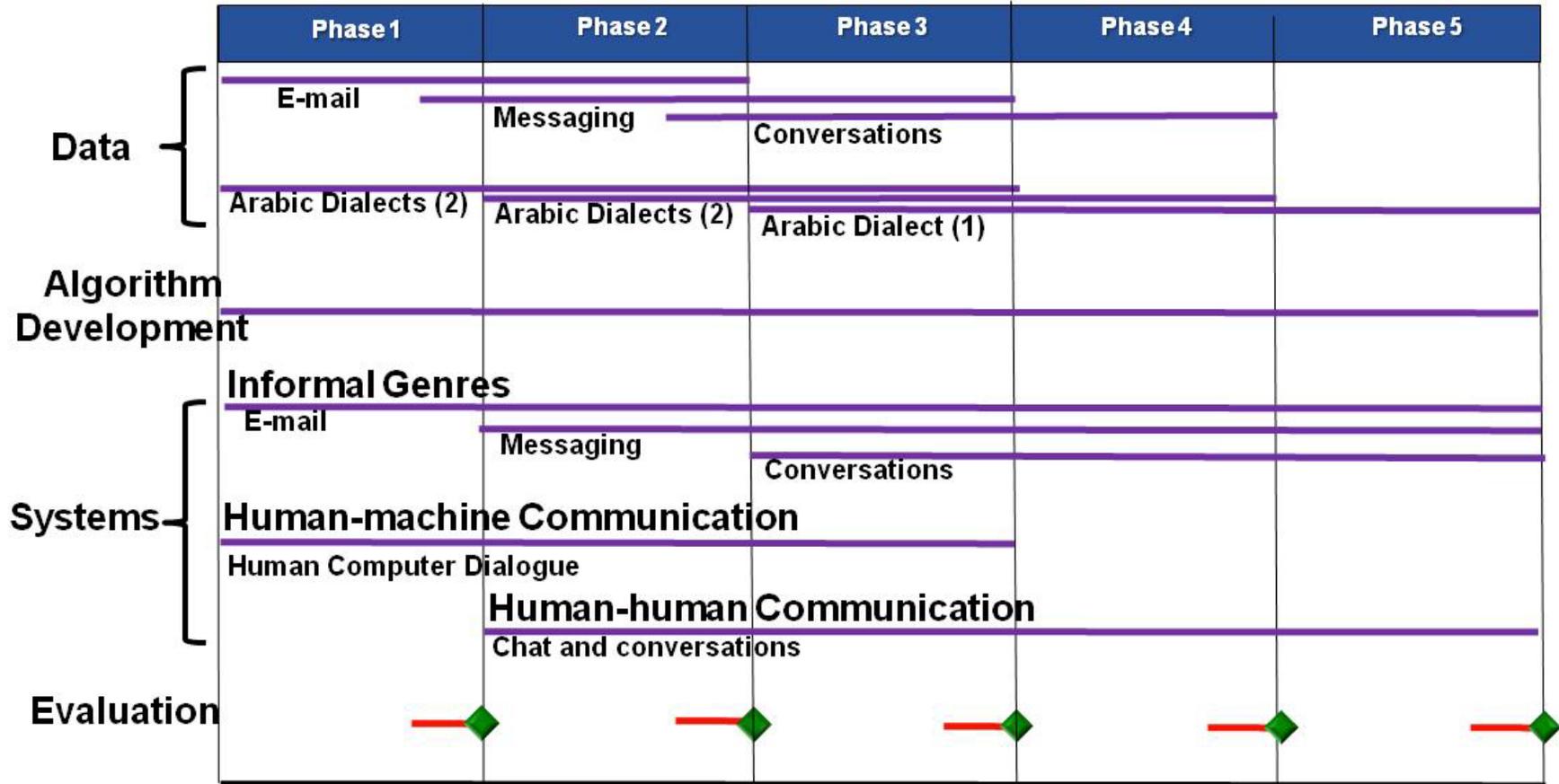
BOLT Program Organization (continued)

- Notionally each area is a 5-phase effort (nominally 12 months)
 - Bidders should develop phasing plans that **realistically** match their efforts
 - Each bidder's proposal must address each proposed phase
- DARPA expects to make multiple awards in each technical area
 - Note: In Technical Areas 2 and 3, there will be one award per activity
- Offerors may submit proposals for all three technical areas

Offerors selected for any activity of Technical Area 1 cannot be selected for that activity in Technical Area 3 and vice versa.



BOLT Notional Schedule





Program Execution – Integrated Systems

- Three different integrated systems
 - Translation and information retrieval (Technical Area 1 Activity A)
 - Human-machine communication (Technical Area 1 Activity B)
 - Human-human communication (Technical Area 1 Activity C)
- Architectures should be configured for ease of incorporating a variety of technologies and algorithms
 - Each algorithm produces annotations synchronized with the input stream
 - Input stream and annotations are available for all subsequent algorithms
 - Algorithms can be repeated after new annotations from different algorithms
 - All decisions are deferred to the end
 - Decisions may consider all or some of the algorithms



Technical Area 1: Algorithms & Integrated Systems

Activity A: Translation and Information Retrieval

Activity A: Genre-Independent Translation and Information Retrieval

System: foreign language translating capabilities to include translation into English and targeted information retrieval in both the source language and English

- **Genre-Independent Translation:** Accurate translation into English of all language genres:
 - Ability to translate informal language (Arabic and Chinese)
 - E-mail – first year; Messaging – second year; Conversation (speech) - third year
 - At least one Arabic dialect (e.g., Levantine or Iraqi)
 - Chosen at the start of the program (first dialect in Activity D)
 - Show capability to use Activity D development
 - Demonstrate methods to eliminate brittleness
 - **Metric:** Overall average accuracy as measured by Human Translation Error Rate (HTER) should exceed **90%**



Technical Area 1: Algorithms & Integrated Systems

Activity A: Translation and Information Retrieval(continued)

Activity A: (continued)

- **Information Retrieval:** Capability to retrieve targeted information from foreign language sources for which machine translation exists
 - Handling of natural language queries
 - Ability to clarify queries and resolve ambiguities
 - Analysis of English documents and documents translated from Arabic and Chinese
 - Methods for annotation of documents (textual or spoken) for fast comprehensive search
 - Details of the response forma
 - **Metric:** the target accuracy for finding all of the relevant information represented as non-redundant entries with complete citations for multiple occurrences is **90% for 90% of the test queries**



Technical Area 1: Algorithms & Integrated Systems

Activity B: Human-Machine Dialogue Systems

Activity B: Human-Machine Communication System: A system providing the capability for human-machine communication in English and one Arabic dialect (dialect to be determined by DARPA at the beginning of the program)

- Demonstrate human control involving complex commands
- Using multi-modal input (language, gesture, gaze, etc.) for robotic control and/or a desk-top application
- Use of dialogue for clarification and disambiguation
- Methods for dealing with OOV's
- Language generation for the dialogue in the language of the interaction
- **Metric:** Complex commands to control a desk-top application (e.g., Microsoft Excel, e-mail, web browsing, etc.) or a robot requiring a minimum of **10** sequential commands with a task completion rate exceeding **90%**



Technical Area 1: Algorithms & Integrated Systems

Activity C: Human-Human Dialogue Systems

Activity C: Human-Human Dialogue System: A system enabling multi-turn, bilingual human-human conversation using English and one Arabic dialect

- Both speech and text
- Unrestricted dialogue
- Language translation
- Monolingual human-machine dialogue in both languages for clarification and disambiguation.
- Means to deal with OOVs
- Methods for language generation in both languages
- **Metric:** English and Arabic human-human conversation (text or speech) requiring **5** complete turns with a successful completion rate exceeding **90%**



Technical Area 1: Algorithms & Integrated Systems

Activity D: Arabic Dialect Translation

Activity D: Arabic Dialect Components: Technological components for accurate translation into English of **five** Arabic dialects, to be determined by DARPA at the beginning of the program

- Arabic dialect(s) need to be integrated with Activities A, B, and C.
- Five Dialects
 - First two dialects will be addressed in Phases 1 and 2
 - Second two dialects will be addressed in Phases 3 and 4
 - Final dialect will be addressed in Phase 5
- We expect the resultant capabilities to be available and used by all performers of Technical Area 1 Activities A, B, and C
- Use and integration of these tools will be discussed in a later slide
- **Metric:** The dialectal Arabic to English translation output should have an overall average accuracy as measured by Human Translation Error Rate (HTER) exceeding 90%

Performance on the first two dialects must reach the 90% metric by the end of Phase 3; the second two dialects by the end of Phase 4; the final dialect by the end of Phase 5



Technical Area 1: Algorithms & Integrated Systems

Activity E: Grounded Language Acquisition.

Activity E: Grounded Language Acquisition: research in deep semantic language acquisition using robotic visual and tactile information as input for experiential learning of objects, actions, and consequences

- Enable robots to acquire language capabilities by grounding objects (nouns and associated adjectives)
- Also enable experiential learning of actions (verbs and adverbs – pre-state, initial state, intermediate state, and final state)
- Also provide robotic abilities to hypothesize and perform automated reasoning in the acquired language
- **Metric:** A robot equipped with vision and tactile inputs will be expected to recognize **250** objects varying in color, shape, size, etc., and understand the consequences (pre-state and post-state) of **100** actions so that it can execute complex commands with **90%** completion rate



Technical Area 1: Algorithms & Integrated Systems Activity E: Grounded Language Acquisition (continued)

Goal: Enable computers to associate real world objects, properties, and activities with linguistic information and use the language facility for reasoning and problem solving:

- Each phase becomes increasingly complex with respect to the objects to be recognized, activities to be accomplished and the surrounding environment
- At the kick-off meeting for each phase, the objects to be recognized, activities to be accomplished and the surrounding environment will be decided upon jointly among the Technical Area 1 (algorithm developers) and Technical Area 3 (evaluators) performers
- Direct perception of the physical world coupled with interaction with an instructor and a planned curriculum used to teach a robot a targeted lexicon, behaviors and consequences
- Approximate language, reasoning, and behavioral competence of a 2 –year-old child



Technical Area 1: Algorithms & Integrated Systems

Activity E: Grounded Language Acquisition (continued)

- Recognition limited to objects, actions on tangible objects, and the physical properties of objects
- Limited generalization of object categories e.g. glass & cup can both be used as liquid containers
 - Object labels could vary by functionality
- No expectation of metaphorical competence (blue – color; not a mood)
- Comprehensive cognitive architectures expected—biological fidelity not required
- Immediate consequences of actions should be learned and understood
- Simulations permitted for training, however, evaluations performed on a **real** robot supplied by the **developer**
- Learning and evaluation in a well defined universe
- Independent evaluations conducted by Technical Area 3 performer(s)



Technical Area 1: Algorithms & Integrated Systems

Activity F: Basic Technologies

Activity F: Basic Technologies: Research in basic technologies (e.g., parsing, SRL, language modeling, discourse analysis, co-reference resolution, dialogue turn analysis, automatic evaluation, etc.) which are essential to the success of the prior Activities

- **It is preferred that offerers team with the offerors of one or more of the previous activities (A, B, AND/OR C)**
- **Stand-alone awards will be given only under special circumstances**
- **Metric:** The metric for this activity is to show progress at the end of each phase. Progress will include both significant improvement in the accuracy of the technology itself (self-assessment) and improvement in any activity that uses the technology (i.e., Technical Area 1 Activity A, Activity B, or Activity C)



Technical Area 1: Algorithms & Integrated Systems Integration requirements for Activities D and F

Integration/Evaluation:

- Extensive consideration will be given to the adaptability of the algorithms and technology components to the various BOLT integrated systems and to the proposed means of support to be provided by the algorithm/technology component developers
- Activity A, B, and/or C performers will accomplish the integration
- Performers in Technical Area 1, Activities D and F are responsible to show significant improvement in the accuracy of the technology itself (self-assessment)
- Technical Area 3 performers for Activities A, B, and/or C will be responsible for determining the effectiveness of the technology at increasing the accuracy of the relevant integrated system to which it has been incorporated



Technical Area 2: Data Collection

- **Genre-Independent Translation and Information Retrieval System:**

Collections will be made in three genres AND in three languages as follows. All data will be translated, aligned, and annotated for parsing, and semantic roles

- English: 2 million words of email, 2 million words of messaging, and 2 million words of conversation
- Arabic (one dialect, e.g., Levantine or Iraqi to be chosen by DARPA): 2 million words of email, 2 million words of messaging, and 2 million words of conversation
- Mandarin Chinese: 2 million words of email, 2 million words of messaging, and 2 million words of conversation

- **Data for Arabic Dialect Components:**

- Data in five dialects of Arabic will be collected at a rate of two dialects per year for the first two years and one in the third year. For each dialect, one million words of text (any genre) and one million words of transcribed speech will be collected. These collections will be annotated to enable natural language processing and translation. Offerors will detail the necessary annotations

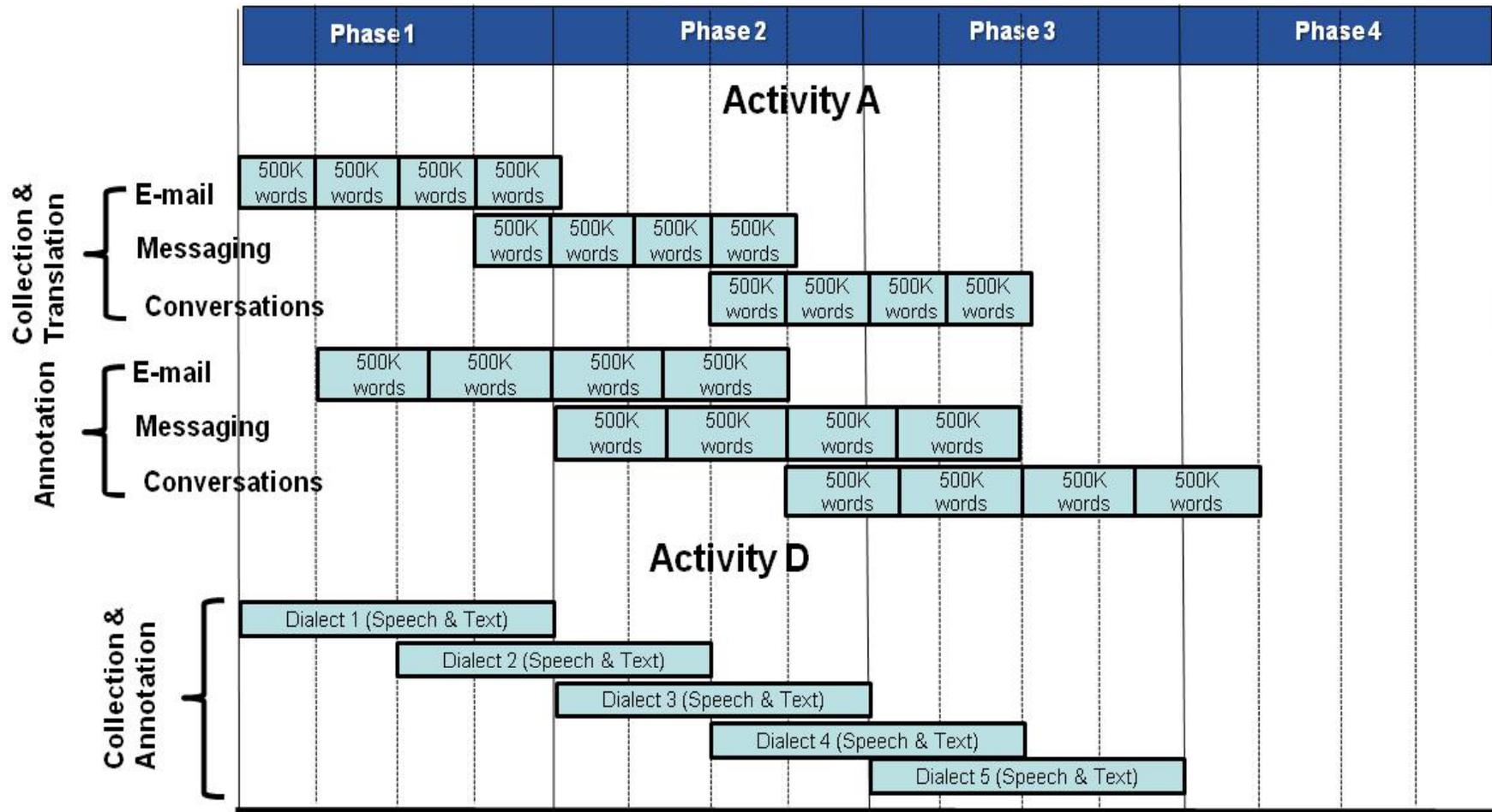


Technical Area 2: Data Collection

- Proposals must address issues of collecting data so it is unclassified as well as issues of privacy, etc.
- Proposals must contain a plan consisting of a data collection matrix which specifies, at a minimum:
 - the source of the data
 - the amount to be collected
 - the languages or dialects of the data
- In addition, offerors must provide an approach and a schedule for collecting, annotating, and furnishing all data. Proposed schedules must be at least as aggressive as the example schedule shown in the figure below



Technical Area 2: Data Collection



Approximately 10% of the data will be held back for the purpose of evaluation (see Technical Area 3) and will not be available for training. This data will be given to the Technical Area 3 performers as applicable.



Technical Area 3: Evaluation

Activity A: Testing of systems developed under Technical Area 1 Activity A will consist of two evaluations:

- (1) the translation quality of the system by comparing system output to carefully-produced human references for Arabic to English and Chinese to English
- (2) the quantity and quality of the information retrieved from documents in English, Arabic, and Chinese in response to queries in English

At the end of Phase 1, translation of e-mail messages will be tested; testing of messaging will be added in Phase 2; Phases 3-5 will include testing on all three genres (e-mail, messaging, and conversation) using the Human Translation Error Rate (HTER) testing method

Activity B: Testing the Human-Machine Communication System(s) for ten sequential commands

Activity C: Testing the Human-Human Dialogue System(s) for five complete sequential dialogue turns



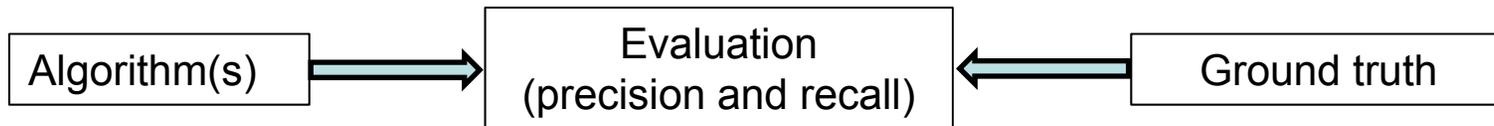
Technical Area 3: Evaluation Activities A, B and C

Proposers for Technical Area 3 Activity A, B, C or any combination of these activities **MUST** also test the performance of the algorithms developed under Activities D and F (Arabic Dialects and Basic Technologies) by evaluating the performance of the respective systems with and without these technologies whenever the Activity D or F technologies are relevant.



Testing Underlying Technologies

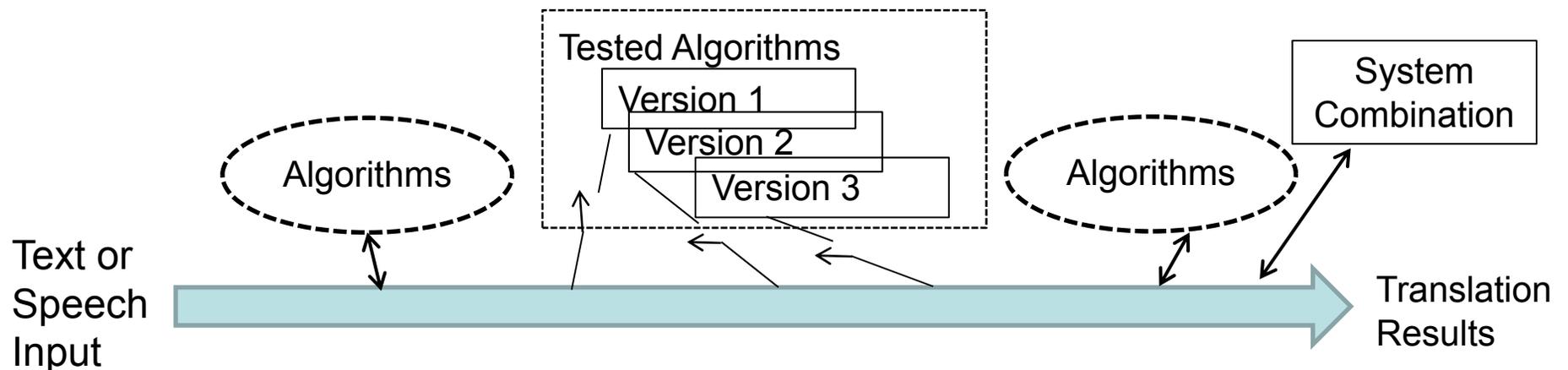
- Testing Progress



- Measure algorithm against ground truth
- Expected progress – 25% error reduction per year

- Testing Contribution to the Overall system

- Connect each version, pair of versions, all versions or none
- Evaluate each translation with a variety of automatic evaluation algorithms





Technical Area 3: Evaluation – Activity E.

Activity E: Testing for competence of Technical Area 1 Activity E (Grounded Language Acquisition)

- Test of robots in a limited environment
 - Tangible objects with physical properties (learned and surprise)
 - Ability to execute complex actions
 - Ability to plan by reasoning



www.darpa.mil