Corpora for Cross-Language Information Retrieval in Six Less-Resourced Languages

Ilya Zavorin, Aric Bills, Cassian Corey, Michelle Morrison, Audrey Tong, Richard Tong
Bluemont Technology and Research Inc, UMD Applied Research Laboratory for Intelligence and Security (ARLIS), MIT Lincoln Lab, UMD ARLIS, National Institute of Standards and Technology, Tarragon Consulting Corporation
Luray, VA, USA, College Park, MD, USA, Lexington, MA, USA, College Park, MD, USA, Gaithersburg, MD, USA, Berkeley, CA, USA
ilya.zavorin@bluemonttechnology.com, abills@umd.edu, cassian.corey@ll.mit.edu, mmorriso@umd.edu, audrey.tong@nist.gov, rtong@tgncorp.com

Abstract
The Machine Translation for English Retrieval of Information in Any Language (MATERIAL) research program, sponsored by the Intelligence Advanced Research Projects Activity (IARPA), focuses on rapid development of end-to-end systems capable of retrieving foreign language speech and text documents relevant to different types of English queries that may be further restricted by domain. Those systems also provide evidence of relevance of the retrieved content in the form of English summaries. The program focuses on Less-Resourced Languages and provides its performer teams very limited amounts of annotated training data. This paper describes the corpora that were created for system development and evaluation for the six languages released by the program to date: Tagalog, Swahili, Somali, Lithuanian, Bulgarian and Pashto. The corpora include build packs to train Machine Translation and Automatic Speech Recognition systems; document sets in three text and three speech genres annotated for domain and partitioned for analysis, development and evaluation; and queries of several types together with corresponding binary relevance judgments against the entire set of documents. The paper also describes a detection metric called Actual Query Weighted Value developed by the program to evaluate end-to-end system performance.

Keywords: Cross-Language Information Retrieval, Less-Resourced Languages, Queries, Speech, Text

1. Introduction

In recent years, deep learning methods revolutionized many areas of Natural Language Processing (NLP) research, including Cross-Language Information Retrieval (CLIR) and Cross-Language Summarization (CLS). CLIR allows users to retrieve relevant content in one or more languages different from the language of the user's query, while CLS provides a way for the user to assess relevance of retrieved foreign-language documents without knowing the language of those documents. While such capabilities are critical to allow monolingual speakers access to foreign data, the amount of training data of sufficient quality often required by deep learning based methods to perform well is simply not available for many less-resourced languages of the world. Even though the amount of digital content the world produces increases tremendously every year, the situation is further complicated by the need to rapidly adapt NLP technologies to new languages, genres and domains.

First conceived in 2015 and launched by the Intelligence Advanced Research Projects Activity (IARPA) in 2017, the Machine Translation for English Retrieval of Information in Any Language (MATERIAL) research program research program is designed to address these challenges (Rubino, 2017). MATERIAL grew out of the Babel program (Harper, 2011) which focused on rapid development of methods to support robust keyword search of large collections of noisy conversational speech. Like Babel, MATERIAL focuses on rapid development of systems for Less-Resourced Languages (LRLs) using limited resources, but it aims at propelling research in a wider array of technologies. MATERIAL performers are tasked with building End-to-End or English-in/English-out (E2E) systems capable of retrieving foreign language speech and text documents relevant to different types of English queries that may be restricted by domain, and providing evidence of relevance of the retrieved documents to both the query string and domain. This evidence is presented in the form of English query-biased summaries.

This paper describes several corpora that were developed for each MATERIAL language. These include (i) annotated Machine Translation (MT) and Automatic Speech Recognition (ASR) build packs, (ii) document sets in three text and three speech genres annotated for domain and partitioned for analysis, development and evaluation, and (iii) queries of several types together with corresponding binary relevance judgments against the entire set of documents. These full-annotation corpora enable exploration of both high-precision and high-recall retrieval of a diverse set of LRLs, unlike more classical IR evaluations such as TREC (Voorhees and Harman, 2005) or CLEF (Ferro and Peters, 2019) that have historically focused on high-resource languages and/or relied on post-hoc annotation.

MATERIAL E2E systems would require high-quality component technologies such as MT and ASR. However, scoring high on standard component quality metrics such as BLEU or WER does not guarantee commensurate performance on downstream tasks (such as retrieval). Therefore, MATERIAL introduced a novel E2E evaluation protocol that combines automatic and human evaluation. While the details of this protocol are beyond the scope of this paper, we present here its central component, which is a detection metric called Actual Query Weighted Value (AQWV) designed to measure quality of both retrieval and summarization.

2. MATERIAL Languages

Table 1 lists MATERIAL LRLs released to date. These languages were selected to create a broad evaluation portfolio consisting of languages with significant Internet

presence from different language families to provide diverse phonotactic, morphological, and syntactic characteristics. To encourage rapid system development, identities of the languages were only known by the Test and Evaluation team until the date of their release. By June 2021, MATERIAL will release three more languages.

	Languages	Release Date
Program Phase I	Swahili (SWA), Tagalog (TGL)	Oct 2017
	Somali (SOM)	Sep 2018
Program Phase II	Lithuanian (LIT) Bulgarian (BUL)	Mar 2019
	Pashto (PUS)	Jan 2020

Table 1: MATERIAL Languages Released to Date

3. MATERIAL Build Packs

For each program language, IARPA provided to the performer teams a build pack that contained MT and ASR training data. It also included a language specific design document (LSDD) which contains information on dialectal variation within the language, related languages, orthographic variation and Unicode codepoint range. The LSDD also describes domains that were both targeted during data collection and those that were specifically excluded due to difficulty of collecting them. For languages that do not use Latin script, the LSDD provided the corresponding Romanization scheme.

3.1 MT Training Data

The bitext portion of each build pack contained 800,000 words of source language text, carefully translated to English at the sentence level and provided to the performers in the form of parallel sentences. Unlike the main collection for the document partitions described in Section 4, bitext sentences were not restricted to a prescribed set of domains. They were, however, collected from sources similar to those defined by the news text genre, i.e. from news stories and articles. The sets of documents collected for the bitext and the main collection were disjoint.

Source sentences were delivered without any type of editing or spelling normalization. Up to five sentences in sequence taken from a single paragraph in an article were marked to indicate their grouping based on the source article. Content was sourced from widely distributed news stories and articles published by major news outlets and local/regional news stories and articles. No pictures, tables or diagrams were included. Content consisting only of fictional narrative, poetry, political comics or drawings was excluded. Scanned newspaper articles were not accepted, and only content of quality consistent with typical published news stories and articles in the target language was collected. A detailed Data Delivery Specification document was created that included translation guidelines used to manually translate the source-language content into English, including handling of idiomatic expressions and metaphors, numbers, foreign/loan words, titles etc.

Because translation into English was performed at the locations where source-language content was collected, additional quality control steps were performed to ensure that the translations are well-formed and fluent. These included:

- Processing of the English side of bitexts by LanguageTool (Miłkowski, 2010) followed by human review of reported errors of four categories: duplication, grammar, misspelling, and non-conformance. Adding this step to the quality control pipeline significantly improved bitext quality.
- Thorough review by US-based native speakers or trained linguists of random samples of parallel sentences to assess their fluency in idiomatic American English.
- Various automatic and manual checks to ensure adherence to the data delivery specification.

3.2 ASR Training Data

The speech portion of each build pack contained a collection of conversational recordings in the form of 8-bit a-law SPHERE (.sph) files and 24-bit WAVE (.wav) files, together with transcription files encoded as UTF8 text.

The speakers involved in the collection of conversational telephony recordings were required to be native language speakers. They were recruited with the goal of obtaining broad coverage of age, gender, and dialect. They were encouraged to talk about topics they felt most comfortable discussing such as family, friends, sports, movies, etc. These topics were not fixed and varied across languages. Speakers showing distinctive speech disorders were excluded from collection or removed if identified later in the transcription process. All speakers were 18 or older. Dialect regions were defined prior to collection for each language. The number of chosen dialects varied across languages (see Table 2), with no dialect representing less than 10% of the collection.

There were no restrictions on acoustic environment (such as whether or not the speaker was indoors, outdoors, driving, etc.) and this information was provided by the speakers themselves. There were also no restrictions on network specifications or telephone models and these values were also noted in the accompanying metadata. Audio was recorded via telephone over an ISDN connection with a terrestrial telephone network. Each speaker was recorded on a separate channel. No post-processing steps were taken to reduce noise or other artifacts of the recording medium at any stage. The total amount of data for each language is shown in Table 2.

For the sake of transcription, the audio files from both channels were programmatically aligned and merged into a single WAVE file. This reduced the burden on transcribers and produced a single transcription file for each conversation that is separated back into channels for the build pack.

Transcription was performed on short utterances in the audio. Each utterance was transcribed on a new line in the transcription file beginning with a time-stamp that indicates the beginning of the utterance. The time-stamp appears in square brackets. In addition to timestamps, the transcription files may also contain tags to represent speech events such as hesitancies, word fragments, overlap, or prolonged periods of silence. Only the time-stamps, tags, and transcription itself appear in the transcription files. Punctuation in the transcription files was at the discretion

of the transcribers who were instructed to abide by natural conventions of the relevant language.

The pronunciation lexicon file provided for each language provides complete coverage of the transcription files in the build pack. The number of terms present in each lexicon is shown in Table 2. This file contains a single term per line with the term in its source language, a Romanized transliteration (where applicable), and a Romanized pronunciation.

Lang	Dialects	Hours	Lex Word Count
SWA	Nairobi	101.92	25289
TGL	North, South, Central	99.94	16129
SOM	Benaadir, Northern	100.29	25874
LIT	Aukstaitian, Samogitian	100.88	32713
BUL	East, West	76.62	22064
PUS	NW, NE, SE, SW	179.77	18745

Table 2: ASR build pack dialects, total hours and word lexicon size per language

4. MATERIAL Documents

For each program language, a document pool was collected of about 15,000 documents with an approximately 3:1 ratio of text to audio documents, in six genres: news text (NT), topical text (TT), and blog text (BT) as well as news broadcast (NB), topical broadcast (TB), and conversational speech (CS). News texts consisted of newswire reports and editorials from national, regional, and local news outlets focusing on news topics and current affairs. These documents targeted a general audience and were presumably highly edited. They were typically around 250-500 words long. Topical texts consisted of articles, reports, non-scientific essays from newspapers or magazines covering a particular topic in-depth. These documents targeted an educated but not specialized audience. They were typically formalized and edited with topic relevant vocabulary and were around 500 words in length. Blog texts were blogs with a single author and did not include discussions or commentaries from other contributors. Blog texts were presumably less edited and more informal with a general vocabulary. They were on average about 500 words. News broadcasts consisted of audio segments of approximately 2.5 minutes from widely distributed broadcasts as well as regional and local news covering news topics and current affairs. The broadcasts were of studio quality while the speech could be formal or informal depending on the segments. Topical broadcasts were similar to news broadcasts in terms of audio quality and speech characteristics but were devoted to in-depth topics and around five minutes in duration. Conversational speech, as described in Section 3.2, consisted of natural conversations between two speakers over the telephone for a duration of approximately 10 minutes on a topic of their choosing from a list of proposed topics. All text and audio documents were five or fewer years old at the time of collection. Outside of checking for the correct language and appropriate content, no additional editing or normalization was performed on these document collections prior to delivery.

Domain	Gloss
Business-And-Commerce (BUS)	All activities and entities associated with economic endeavor.
Government-And-Politics (GOV)	Anything to do with local, regional, national or international government. Includes national level functions such as the provision of national or international infrastructure and capabilities.
Law-And-Order (LAW)	Anything to do with crime, violence or the enforcement of local, regional and national laws.
Lifestyle (LIF)	Anything to do with the lives of families and individuals and the activities they engage in as well as cultural values, norms, practices and expressions.
Military (MIL)	Anything to do with military capability, activity or entities.
Physical-And-Mental-Health (HEA)	Anything to do with the provision of health and wellbeing to a population, as well as causes and correlates that affect health and wellbeing, such as accidents and non-natural disasters. Includes community public health concerns.
Religion (REL)	All aspects of personal and organizational belief systems and practices that relate humanity to what the adherents of that religion consider to be ultimate reality; theological works.
Sports (SPO)	Anything to do with sports activities and entities.

Table 3: Domain names and their glosses

In Phase I of the program, search queries were contextualized by domains, and so the documents were annotated with domain information. While a number of domains were annotated, only eight were eventually released (see Table 3 for their glosses) for the Phase I languages, as listed in Table 4. Annotators were given a gloss for each target domain as well as domain definitions and additional notes to clarify the scope of the domain. Each document had two independent domain annotation passes with a third annotator adjudicating the two previous passes for disagreements.

Lang	Target Domains
SWA	GOV, LIF, BUS, LAW, SPO
TGL	GOV, LIF, HEA, MIL, SPO
SOM	GOV, MIL, BUS, LAW, REL

Table 4: Target domains for the Phase I languages.

The document pool was partitioned into analysis, development, and evaluation sets. The analysis set also included transcriptions and translations of the source documents. It was released to performers months before the official evaluation for glass-box error analysis. Performers were allowed to manually examine the analysis documents in detail and to use it for parameter tuning but were not

allowed to mine for or train language models from the vocabulary in the analysis set for their MT/ASR development. Like the analysis set, the development set was also distributed months in advance of the evaluation for internal testing and had the same restrictions as the analysis set. However, unlike the analysis set the development set did not include transcriptions or translations and performers could not manually examine the development documents. The evaluation set for the official evaluation was not released until the start of the evaluation period. Performers were to treat the test set as a blind test set: no examination of the documents, no tuning, no mining for vocabulary.

In Phase I, the partition was based on having the target domains represented in the analysis and development sets in similar frequencies. The evaluation set then would be the remaining documents not selected as the goal was to ensure the target domains and (combinations of target domains) were adequately represented for system development. Additionally, in order to evaluate language identification capabilities of MATERIAL systems, the evaluation set included some distraction data (text and audio documents in a different language than the language being evaluated).

In Phase II, the document partition was changed to focus on achieving a more balanced P_{Rel} (the probability that a document is relevant to the query) across the document sets without any consideration to the domains as domain was dropped from the focus of the Program. Table 5 gives the document volumes for these datasets for the six languages used in the first two phrases of the Program.

Lang	Analysis		Dev		Evaluation	
	Text	Spch	Text	Spch	Text	Spch
SWA	547	266	449	217	10254 (181)	3267 (1043)
TGL	529	315	460	244	10261 (81)	3191 (1260)
SOM	559	279	482	213	10209 (508)	3259 (1383)
LIT	614	215	433	238	10203	3297
BUL	515	312	416	258	10319	3180
PUS	563	284	470	185	10217	3281

Table 5: Document count for the various datasets for the program six target languages. The number in parenthesis denotes the distraction document count.

5. MATERIAL Queries

5.1 Query Typology

Queries are the means by which users express an information need to the CLIR software developed by the performer teams. In contrast to TREC queries, which consist of multiple sentences restating and delimiting the information need, MATERIAL queries are short, consisting of one or two words or short phrases and optional constraints to reduce ambiguity. These come closer to the kinds of queries one might type into a search engine.

The MATERIAL program targets two kinds of requests for information. The first, a **lexical** request, is a request to find documents containing a specific word or phrase (or a translation equivalent of that word or phrase); queries of this type are used to analyze a system's machine translation, speech recognition and retrieval abilities. The second, a **conceptual** request, is a request to find documents related to a specified concept, regardless of which specific words in a given document touch on that concept; queries of this type are used to analyze a system's information retrieval capabilities.

During Phase I of the program, each query was contextualized by one of the target domains for the corresponding language (see Table 4). This means that in order for a query to be relevant to a document, its domain had to match one of the domains assigned to the document, in addition to the document's content matching the query string. In Phase II domains were dropped to simplify both query development and performance analysis. In the remainder of this paper, we discuss queries and their relevance without any additional domain constraints.

A MATERIAL query can consist of one or two requests for information. In the latter case, a document must satisfy both requests in order to be considered relevant. There need not be any relationship between the two requests. We call these queries **conjunction** queries.

Three "in the sense of" semantic constraints were used for queries with ambiguous words or phrases. A synonym (*syn*) constraint specifies an English word or short phrase that conveys approximately the intended sense of the query term (for example, *star [syn: celebrity]*). A hypernym (*hyp*) constraint specifies a superordinate category of the intended sense of the query term (for example, *bat [hyp: animal]*). An event frame (*evf*) constraint specifies the semantic domain of the intended sense of the query term (for example, *bar [evf: nightlife]*).

A subset of program queries was developed to target specific types of information requests that would be challenging for CLIR systems. Phenomena that were hypothesized to be challenging included polysemy (in particular, cases where a specific word in the document language might be translated into English in multiple ways depending on the context in which it is used), homophony (a word in the document language with the same pronunciation as another word in the language), and homography (a word in the document language with the same spelling as an etymologically unrelated word in the language). Additionally, named entities were targeted because they are more likely to be out of vocabulary than non-names, and could be potentially confused with non-named entities.

Below we present MATERIAL query types with a brief explanation of relevance rules and examples (in English, for demonstration purposes). Some examples of **lexical queries** are given below.

Query: *"herbal medicine"*
Type: lexical, single request
What is considered relevant: documents containing [a translation equivalent of] the phrase in the query, including inflectional variants (e.g., *"herbal medicines"*)
Relevant example: Why not try some herbal medicine?
Non-relevant example: Why not try some medicine?

Query: *prisoner, bribery*
Type: lexical, two requests (aka conjunction)
What is considered relevant: documents containing [translation equivalents of] both words in the query, including inflectional variants, in any order
Relevant example: …two prisoners escaped … In other news, the mayor is accused of bribery…

Query: *fly [hyp: insect]*
Type: lexical with semantic constraint
What is considered relevant: documents containing [a translation equivalent of] the specified sense of the word in the query
Relevant example: There's a fly in my soup!
Non-relevant example: Kiwis can't fly.

The program has also developed a special subtype of lexical queries called **morphological** queries. Queries of this type targeted words with specific marked (e.g., non-default) morphological properties. For example, the query <*will tell*> would match only forms of '*tell'* (or a translation equivalent of this word) in the future tense. (In contrast, the [non-morphological] lexical query *tell* would match forms of '*tell'* in any tense.) Another example of a morphological query is given below.

Query: <*prisoners*>
Type: lexical (morphological)
What is considered relevant: documents containing [a translation equivalent of] the word in the query with the same marked morphological features as the word in the query (in this case, plural number)
Relevant example: prisoners escaped
Non-relevant example: only one prisoner died

A **conceptual** query contains a conceptual request for information. General conceptual requests are marked with a plus sign: *"violence in Sudan"+*. An additional kind of conceptual request was used called **EXAMPLE_OF**. This kind of request was used to test a system's knowledge of basic/natural class hierarchies. A document was considered relevant if it mentioned a subtype of the specified concept. For example, a document would be considered relevant to the request *EXAMPLE_OF(apparel)* if it mentioned sweaters; if the document only contained the word '*apparel'* (or a translation equivalent thereof), it would not be considered relevant. EXAMPLE_OF requests have been discontinued for the third period of performance.

A distinction is made between "pure" conceptual queries, which consist of a single conceptual request, and "hybrid" conceptual queries, which contain/conjoin one conceptual request and one lexical request. For practical reasons, queries consisting of two conceptual requests are disallowed. Some examples of conceptual queries are given below.

Query: *"violence in Sudan"+*
Type: conceptual, single request (pure conceptual)
What is considered relevant: documents that touch on the specified concept
Relevant example: Negotiations in Sudan ended abruptly after violent clashes erupted in the capital.
Non-relevant example: Protesters in Sudan marched outside the presidential palace in Khartoum.

Query: *EXAMPLE_OF(freshwater fish)*
Type: conceptual (EXAMPLE_OF)
What is considered relevant: documents mentioning a subtype of the requested concept
Relevant example: I caught a carp
Non-relevant example: A large catch of cod

Query: *strike+ [evf: labor]*
Type: conceptual with semantic constraint
What is considered relevant: documents that touch on the specified sense of the requested concept
Relevant example: Teachers staged a walkout
Non-relevant example: Threat of a terrorist attack

Query: *"traditional practice", health+*
Type: conceptual, hybrid/conjunction
What is considered relevant: documents that contain [a translation equivalent of] the lexical phrase in the query and touch on the specified concept
Relevant example: A traditional practice in the Philippines is to use guava leaf ointment to expedite healing.
Non-relevant example: Guava leaf tea tastes terrible.

5.2 Query Development and Annotation Process

Queries were developed by teams of three native language speakers per language. They input queries into a web-based tool called the Query Development Tool (QDT) which was developed from scratch to support this effort. This tool was used to develop and test queries against document sets as well as to annotate relevance judgments for individual documents. The QDT also allowed for quality control checks at several stages in the process.

Inspiration for queries came from a variety of sources. Using the QDT, query developers could retrieve a random text or speech document and look for content that might make an interesting query. Some queries were developed from wordlists derived from program documents, such as topical words extracted via Latent Dirichlet Allocation (McCallum, 2002). Often, while annotating one query, a query developer might encounter information in a document that would serve as the basis for their next query. Many queries were based on ideas that came directly from the query developer (for example, the developer might think of a word that happens to be a homograph, and use that as the basis for a query).

Once a query developer had a query concept in mind, they created a list of specially formulated QDT queries that were used to find all documents that could possibly be relevant to the query, including the English query string. QDT searches were intended to achieve 100% recall of relevant documents; precision was not a factor at this stage.

Queries were later reviewed by a second native language speaker, as well as a native English speaker. The vetting process included checks that queries met a number of different criteria, including: 1) Is the query well-formed? 2) Is the query clear for a native speaker of English? 3) Is the query specific enough, or do constraints need to be added? 4) Does the QDT search contain all possible translation

equivalents of the query? 5) Does the QDT search account for all possible inflected forms of words in the query? 6) Does the query correspond to relevant documents? If a query did not meet the quality control targets, it was either further refined or discarded. Once the native speaker and the English speaker reviewers agreed that the query met the vetting criteria, the query was marked as "frozen" in the QDT, and no further query edits were made.

After queries were vetted, the initial query developer annotated documents in the corpus according to their relevance to the query. For lexical queries, if any lexical item in a particular document was a translational equivalent to the query term, the document was marked as relevant. For conceptual queries, a relevant document did not need to contain an exact translation equivalent of the query term(s), but it had to cover the query topic. The QDT showed query developers snippets of documents containing words that matched parts of the search. In some cases, particularly with lexical queries, the snippets provided were not sufficient for the annotator to determine whether the query was relevant; in those cases, annotators could click on the set of snippets and be shown the entire document, with items matching QDT search terms. Following document annotation, a second round of vetting took place before the query was finalized.

5.3 Query Statistics

Table 6 and Table 7 show total counts of various query types developed for each language against text and speech documents, respectively. All queries were partitioned into two disjoint subsets: Q1 was a set of queries annotated against development and analysis document partitions and were released to the performers together with those document partitions. Q2 was a set of queries annotated against the evaluation partitions and were used to evaluate system performance. The two tables list the four basic query types, lexical, morphological, conceptual and EXAMPLE_OF, as well as their conjunctions. The average P_{Rel} (see Section 4) for these query sets is 0.00165.

6. MATERIAL Evaluation Metric

The nominal MATERIAL use case is one in which a user is monitoring a stream of documents for topics of interest. A perfect system would detect all the relevant documents, while rejecting all the non-relevant ones. In practice, some relevant documents will be missed and some non-relevant ones will be falsely detected. Given that scenario, the primary MATERIAL performance metric was designed to allow the program to measure the trade-off that systems are making between miss rates and false alarm rates.

This measure, called QV (Query Value), is defined for a given query as:

$$QV = 1 - P_{Miss} - \beta \cdot P_{FA}$$

where P_{Miss} is the probability that a relevant document for the query will not be detected, and P_{FA} is the probability that a non-relevant document will be incorrectly detected.

		SWA	TGL	SOM	LIT	BUL	PUS
l	Q1	93	54	112	131	56	60
	Q2	351	360	408	495	247	162
m	Q1	6	5	12	29	11	10
	Q2	91	51	75	113	54	22
c	Q1	9	8	3	2	23	3
	Q2	27	111	11	5	67	14
e	Q1	10	1	0	2	2	1
	Q2	8	28	9	21	7	2
l,l	Q1	33	36	34	15	87	93
	Q2	226	149	99	53	307	254
l,m	Q1	0	1	3	6	17	24
	Q2	8	11	28	15	49	69
m,m	Q1	0	0	0	0	0	0
	Q2	1	0	0	1	0	0
l,c	Q1	21	17	35	74	51	79
	Q2	247	76	289	239	179	202
m,c	Q1	0	0	0	0	0	0
	Q2	0	1	0	0	4	0
l,e	Q1	0	2	4	15	23	6
	Q2	10	10	37	38	72	18

Table 6: Number of queries of different types developed against **speech** documents in each language. Q1 and Q2 are query sets against Development+Analysis and Evaluation document partitions, respectively. l, m, c, and e stand for lexical, morphological, conceptual and EXAMPLE_OF query types, respectively.

		SWA	TGL	SOM	LIT	BUL	PUS
l	Q1	67	36	82	74	29	34
	Q2	314	321	346	390	198	116
m	Q1	4	5	13	12	6	5
	Q2	66	44	60	87	36	11
c	Q1	8	3	1	1	10	0
	Q2	22	106	8	2	39	2
e	Q1	9	3	0	2	1	0
	Q2	8	29	8	11	6	2
l,l	Q1	19	20	24	4	46	39
	Q2	194	125	78	34	222	160
l,m	Q1	0	0	1	3	8	7
	Q2	7	11	26	9	34	31
m,m	Q1	0	0	0	0	0	0
	Q2	0	0	0	1	0	0
l,c	Q1	19	10	17	30	22	21
	Q2	159	55	211	133	101	69
m,c	Q1	0	0	0	0	0	0
	Q2	0	0	0	0	4	0
l,e	Q1	0	2	3	9	13	2
	Q2	6	9	22	25	46	5

Table 7: Number of queries of different types developed against **text** documents in each language. Q1 and Q2 are query sets against Development+Analysis and Evaluation document partitions, respectively. l, m, c, and e stand for lexical, morphological, conceptual and EXAMPLE_OF query types, respectively.

The parameter β is defined as:

$$\beta = \frac{C}{V} \cdot \left(\frac{1}{P_{Rel}} - 1\right)$$

where C is the cost of an incorrect detection and V is the value of a correct detection.

In MATERIAL all queries are equally weighted and so the program metric Actual Weighted Query Value (AQWV) is the simple average over all the QVs for a system operating at its actual detection threshold. In any given evaluation, the MATERIAL Test and Evaluation Team specifies β as a constant *a priori*, and performer systems optimize their performance accordingly. A typical value of β is 40 ($V = 1$, $C = 0.0668$, $P_{rel} = 0.0017$). Because of the equal weighting of queries, AQWV is better suited than many traditional information retrieval metrics for the needle-in-the-haystack MATERIAL system use case.

Note that $AQWV = 1.0$ for a perfect system; $AQWV = 0$ for a system that detects no documents at all; and, $AQWV = -\beta$ if all the detected documents are false alarms.

Table 8 shows maximal AQWV CLIR scores achieved by individual MATERIAL performer systems on the speech and text portions of the evaluation sets for five of the six program languages evaluated as of March 2020.

Language	Beta	Mode	AQWV CLIR
SWA	20	speech	0.4556
		text	0.5046
TGL	20	speech	0.5917
		text	0.6408
SOM	40	speech	0.2036
		text	0.2901
LIT	40	speech	0.6093
		text	0.6497
BUL	40	speech	0.6539
		text	0.7244

Table 8: Maximal single-system CLIR AQWV for the MATERIAL languages evaluated as of March 2020.

7. Summary

In this paper we presented several document and query datasets that were created by the IARPA MATERIAL research program for development of CLIR and summarization systems for six LRLs and provided details on document collection and annotation as well as query development, annotation and vetting. The program has propelled research in these areas yielding, as of March 2020, almost 100 publications by the performer teams. The datasets described in this paper are currently being released to US Government entities. It has not been determined if or when they could also be released to a wider research community.

8. Acknowledgements

We thank the anonymous reviewers for helpful comments. This effort is supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contracts 2019-19022100010-005, 2018-1701100005-001, FA8702-15-D-0001, FA8702-15-D-0001, and D2017-1708030008. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copy-right annotation therein.

9. Bibliographical References

Ferro, N., & Peters, C. (2019). From multilingual to multimodal: the evolution of CLEF over two decades. In *Information Retrieval Evaluation in a Changing World* (pp. 3-44). Springer, Cham.

Harper, M. (2011). Babel Broad Agency Announcement. https://www.iarpa.gov/index.php/research-programs/babel/baa

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu

Miłkowski, M. (2010). Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience*, *40*(7), 543-566.

Rubino, C. (2017). Material Broad Agency Announcement. https://www.iarpa.gov/index.php/research-programs/material/material-baa

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval* (Vol. 63). Cambridge: MIT press.