This is the accepted manuscript of the article, which has been published in Ferro N., Peters C. (eds) Information Retrieval Evaluation in a Changing World. Cham: Springer. The Information Retrieval Series vol 41.
ISBN: 978-3-030-22948-1. ISSN: 1387-5264.
https://doi.org/10.1007/978-3-030-22948-1_8

The Challenges of Language Variation in Information Access

Jussi Karlgren, Turid Hedlund, Kalervo Järvelin, Heikki Keskustalo, and Kimmo Kettunen

Abstract This chapter will give an overview of how human languages differ from each other and how those differences are relevant to the development of human language understanding technology for the purposes of information access. It formulates what requirements information access technology poses (and might pose) to language technology. We also discuss a number of relevant approaches and current challenges to meet those requirements.

1 Linguistic Typology

Information access technology—such as information retrieval and related applications—is largely about finding and aggregating meaning from human language, and mostly, so far, from text. On a superficial level, it may seem as if human languages vary a great deal, but they are in fact similar to each other, especially in written form: they share more features than differences. What meaning is and by which means it is encoded in human text is a contentious research topic in itself, but that there is meaning in human utterances and that it is systematically recoverable is not.

Jussi Karlgren
Gavagai and KTH, Royal Institute of Technology, Stockholm

Turid Hedlund

Kalervo Järvelin
University of Tampere, Tampere

Heikki Keskustalo
University of Tampere, Tampere

Kimmo Kettunen
The National Library of Finland

The number of languages in the world is difficult to assess, but is usually put at being around 7 000. More than 90% of those languages are spoken by populations of less than a million and more than half of them by language communities numbering less than 10 000. Many of those languages—primarily the smaller ones—are falling out of use, with some estimates putting about half of the worlds' languages at risk of disappearing. The number of speakers is unevenly distributed: at the other end of the scale the twelve or so largest languages cover half of the population of the world (Lewis et al, 2009; Dryer and Haspelmath, 2011). The details of these facts of course depend crucially on how one language is demarcated from another, which is non-trivial, depending not only on linguistics but also on politics and geography. The variation between human languages is studied in the field of *linguistic typology*, which studies both systematic differences and likenesses between languages (Velupillai, 2012).

Such variation between human languages is first, and most obviously, evident in their writing systems. Some languages use some variation of phonetic writing such as alphabetic or syllabic systems; other systems are based on ideograms; some separate tokens by whitespace, some do not. Some writing systems omit what others require: semitic languages usually do not include vowels, for instance. This type of variation is mostly superficial and is no longer a major challenge for information systems. More importantly, only about half of the world's languages are ever written at all and thus not accessible to most of today's information systems. However, the practical challenge of accommodating various writing systems, character sets, and their encodings, in view of many coexisting and legacy standards may still impact performance.

Secondly, human languages vary in the way they organise the referents the speaker communicates about into a coherent utterance. Some languages impose strict requirements on the order of the constituents of a clause, making use of *word order* an obligatory marker; others allow permutations of constituents within an utterance without much meaning change. Some languages render words in different forms through more or less elaborate *inflection* systems, depending on what role they play in the utterance; others let words appear in more or less invariant form. These two aspects of variation—inflection and word order—are in in the most general sense in a trade-off relation: languages with strict word order tend to have less complex systems for inflection.

Thirdly, many languages combine words or bits of words to make larger words or *compounds* or *derivations*; others prefer to keep words or meaningful units separate.

Fourthly, information that is obligatory to include for some languages may be optional or not mentioned in others.

On another level of abstraction, *genres* and various cultural factors influence which topics are discussed and in which terms. The variation is even more evident with the advent of *new text types*.

We will return to all of the above variational dimensions in turn. More generally, however, all human languages share important features. Languages are *sequential*: they consist of sequences of meaning-bearing units which combine into useful utter-

Table 1 Examples of inflectional variation given for nouns from some languages. Chinese nouns do not inflect. English inflects less than Swedish. Finnish has thousands of possible forms for each nouns.

		SINGULAR		PLURAL	
Chinese		虾			
English		kipper	kipper's	kippers	kippers'
Swedish	INDEFINITE	sill	sills	sillar	sillars
	DEFINITE	sillen	sillens	sillarna	sillarnas
Finnish		muikku	muikun	muikut	muikkujen
	...			...	
	ABLATIVE + "not even"	muikultakaan		...	
	ADESSIVE + our + "also" + EMPHATIC	...		muikuillannekinhan	

ances of salience to their speaker and author, and mostly of interest to their intended or unintended audience. Languages are *referential*: the utterances are composed of expressions which refer to entities, processes, states, events, and their respective qualities in the world. Languages are *compositional*: the constituents of utterances combine to a meaningful whole through processes which to some extent are general and to some extent are bound to situation, context, and participants.

And in the end information access is all about meaning. In the case of text retrieval, about the semantics of a text and the utterances in it.

2 Requirements from Application Domains

The focus of information retrieval experiments has been on the use case of *ad-hoc information retrieval*: the process whereby a concise expression of information need is exchanged for a set or ranked list of documents or other information items. To achieve levels of performance in every or most languages to match the level that systems achieve in English and other widely used languages with large speaker populations, more analysis of the target language is often necessary. This is even more true when the use case is extended to Cross-language Information Retrieval (CLIR), where a query in one language is expected to deliver results in other languages, possibly in combination with results in the target language.

Other related tasks, ranging from media monitoring and routing to sentiment analysis to information extraction often require more sophisticated models and typically more processing and analysis of the information items of interest. Much of that processing needs to be aware of the specifics of the target languages.

Mostly, the various mechanisms of variation in human language pose *recall challenges* for information systems. Texts may treat a topic of interest but use linguistic expressions which do not match the expectations of the system or the expression of information need given by the user: most often due to vocabulary mismatch. This is especially true for users who may know the target language only to some extent, and

who may not be able to specify their information need with as much finesse as native language users would: the benefits of query translation in web search benefits those with poor to moderate competence in the target language more than those who are fluent. Since CLIR will in such cases rely on translating an information need from a source language to a target language, the quality of the translation dictionary or service is a crucial factor for the quality of the end result, whether the translation is done at query time or at indexing time Airio (2008).

Translation is not always possible between arbitrary language pairs, due to lack of resources: see e.g. Rehm and Uszkoreit (2012) for an overview of what resources are available. In such cases, a transitive approach can be adopted, where translation is done from language A to language B by way of translation via a *pivot language* C, if translation resources or services for $A \Longleftrightarrow B$ are unavailable but can be found for $A \Longleftrightarrow C$ and $C \Longleftrightarrow B$. This obviously risks inducing a level of noise and spurious translation candidates, but has been shown to work adequately in many task scenarios Gollins and Sanderson (2001); Lehtokangas et al (2004).

2.1 Cultural Differences and Differences in Genre Repertoire

On the highest level of abstraction, differences between cultural areas are often reflected in how a topic is treated in linguistic data. This may not seem a challenge specifically for information access technology, but awareness of stylistic differences and of acceptability will be a guide to what can be expected to be found in data sources and how much effort should be put into the resolution between similar topics, into sentiment analysis, and other similar tasks.

Many timely and new texts are generated in new media and new genres with little or no editorial oversight: with new, emerging, and relatively volatile stylistic conventions; anchored into highly interactive discourse or into multimodal presentations; incorporating code switching between several languages; characterised by newly minted terms, humourous and deliberate misspellings, topic indicators ("hash tags"), and plenty of misspellings or typing errors. (Karlgren, 2006; Uryupina et al, 2014). This variation does not always follow the same paths across cultural and linguistic areas.

Language processing tools that are built or trained to handle standard language from e.g. news text or academic texts risk being less useful for analysis of new text. Using such tools for multi-lingual material risks skewing results across cultural areas, especially if the reader is less than fluent in the original languages.

2.2 Inflection

One of the first and most obvious differences between human languages is that of *morphology* or inflectional systems: anyone who has made the effort to learn a for-

eign language is familiar with the challenge of learning e.g. verb forms or plural forms, especially irregular ones. The number of different forms of a single lexical entry varies greatly between human languages. Some examples are given in Table 1. Many languages find it necessary to include information about the gender of referents ("elle est fatiguée" vs. "il est fatigué"; "śpiewał" vs. "śpiewała"); others do not. Some require tense or aspect to be marked, some do not. Some allow subjects to be omitted if understood from context ("wakarimasen"); others require subjects even when of low informational content ("es regnet"). The largest languages in the world have very spare morphology: English, Chinese, and Spanish can be analysed using very simple tools (Lovins, 1968; Porter, 1980). Larger languages seem to tend towards simpler morphology, and this observation has been tentatively proposed to have to do with the amount of cultural contact a larger language engages in simply through its dispersal pattern.[1] (Dahl, 2004)

The majority of the world's languages, if not the majority of speakers, have more elaborate morphology. Morphological analysis tools of various levels of sophistication have been developed for languages, often inspired by languages with richer morphological variation than English. These tools have been applied to various tasks such as writing aids, translation, speech recognition, and lately included as a matter of course in many information access systems.

Nouns are in most languages inflected by *number*, to distinguish between one, many, and in some cases pairs of items. In most languages nouns are also inflected by *case*, to indicate the noun's role with respect to other words in a clause. English uses the genitive form to indicate ownership; Latin uses different cases for object and various adverbial functions; Russian adds yet another case to indicate an instrument; Finnish and related languages have a dozen or so cases to indicate various positional and functional roles of nouns. Some languages indicate *definiteness* by inflection (which in English is marked by separate determiners such as *the* or *a*). Verbs in most languages carry information of a temporal and aspectual character of the event, state, or process the clause refers to. In general, adjectives exhibit less complex inflection patterns than do nouns; verbs tend to be more elaborate than nouns.

This variation directly impacts information retrieval performance. If surface variation of terms is reduced through some procedure, the recall of a retrieval system is increased—at some cost to precision—through the system retrieving documents which contain some term in a different surface form than that presented by the user in a query: if a system knows enough to find texts mentioning "festival" when a user searches for texts on "festivals" it will most likely make its users happier (Lowe et al, 1973; Lennon et al, 1981). The process where different forms of a word are collated is variously called *normalisation*, *lemmatisation*, *stemming*, or even *truncation*, depending on which engineering approach is taken to the task.

This variation in morphological systems across languages from the perspective of information access has been addressed in previous literature by e.g. (Pirkola, 2001) who has formulated a description of languages of the world using two variables, *in-*

[1] This would seem to be good news for language technologists with limited resources at their disposal.

dex of synthesis and *index of fusion* and examined how those variables could be used to inform the design of practical tools for both mono- and cross-lingual information retrieval research and system development.

For English, for a long time, it was taken to be proven that normalisation by and large would not help retrieval performance (Harman, 1991). Once the attention of the field moved to languages other than English, it was found that for other languages there were obvious gains to be found (Popović and Willett, 1992), with the cost and utility of analysis varying across languages and across approaches as to how it is deployed (Kettunen and Airio, 2006; Kettunen et al, 2007; Karlgren et al, 2008; Kettunen, 2009; McNamee et al, 2009; Kettunen, 2014).

Not every morphological form is worth normalising. Languages such as Finnish or Basque, e.g., have several thousand theoretically possible forms for each noun. In practice only a small fraction of them actually show up in text. Taking care of the more frequent forms has clear effects on retrieval performance; other forms are more marginal, or may even reduce performance for topical retrieval tasks, if variants which make topically relevant distinctions are conflated.

Today morphological analysis components to normalise terms from text and queries, using a stem or a lemma form instead of the surface form, are used in retrieval systems as a matter of course. For some languages and some tasks, fairly simple truncation-based methods (Porter, 1980, 2001) or n-gram indexing (Kamps et al, 2004) yield quite representative results, but more informed approaches are necessary for the systematic treatment of e.g. languages where affixation can include prefixes or infixes. Most systems today incorporate morphological normalisation by default for some of the larger languages and tools for the introduction of such techniques for languages with less existing technology support.

2.3 Derivation and Compounding

Derivation, the creation of new words by modifying others, and compounding, the creation of new words by combining previously known ones, are productive processes in all human languages. There is no limit to creating new words, but there is a limit in how and to what extent they can be and are included in for example translation dictionaries used in multi- or cross-language information access technology.

Derivational morphology describes how new words can be created through the use of affixes (prefixes and suffixes) combined to a word stem, e.g., *build—builder—building*. Derivation thus affects the part-of-speech and meaning of the word *build* (Akmajian et al, 1995).

Compounds can be closed (such as *classroom*), open (such as *ice cream*), or hyphenated (such as *well-being*). Human languages vary as to how they orthographically construct compounds: German, Dutch, Finnish and Swedish, e.g., favour closed compounds; English orthography is less consistent, but uses open compounds to a much greater extent. The orthographic specification is important in cross-lingual retrieval and is also related to the translation and identification of compounds as

phrases in for example English. Closed compounds are easier to handle in information access technology and in cross-language applications because there is no need for a specific identification of a "phrase" as in open compounds (Lieber and Štekauer, 2009).

Splitting compounds into their constituents may be expedient for the purposes of information retrieval: the compounds may be too specific and splitting them would yield useful and content-bearing constituents, thus increasing recall of an index. This is especially true in a scenario where queries are translated from one language to another (Hedlund et al, 2001).

Doing this is not straightforward, however. A compound may be *compositional*, where the meaning of the compound is a function of some sort of its constituents, or *non-compositional* where the meaning of the constituents is non-relevant or marginally relevant to understanding the compound. Where a compound is compositional, the relation between its constituents may be difficult to predict without world knowledge: most compounds in frequent use have been lexicalised as words in their own right to some extent. In practice, frequently only some or even none of the constituents of a compound are topically relevant (such as in *strawberry*, *Erdbeere*, *fireworks*, or *windjammer*). A compound may also have several possible splits, with typically only one of them being correct (such as in *sunflower*). In languages which make free use of closed compounds these challenges are exarcebated: the Swedish *domstol* (*court of law*) can be split into *dom* and *stol*, the former being *judgment* but also the homograph personal pronoun *they* which trumps the relevant reading by frequency; the latter being *chair*, which is irrelevant; the Swedish 3-way compound *riksdagshus* (*parliament building*) can be reconstructed into *riks*, *dag* and *hus* (*realm*,*day*,*building*) which is less useful than the 2-way split into *riksdag* and *hus*, (*parliament* and *building*).

Many languages make use of *fogemorphemes*, glue components between information bearing consituents, for example, *-ens-* in *Herz-ens-brecher*, the German word for *heart breaker*. Handling these correctly impacts performance noticeably (Hedlund, 2002; Kamps et al, 2004).

Challenges such as these make the application of compound splitting somewhat more difficult than the seemingly simple process the term itself invites (Chen, 2001, 2002; Hedlund et al, 2000; Hedlund, 2002; Cöster et al, 2003; Karlgren, 2005).

In summary, some of the challenges with using constituents from compound splitting are that they may not express a concept similar to that expressed by the compound; may be ambiguous; may not always even be valid words.

2.4 Word Order and Syntactic Variation

Languages vary greatly in how strictly rule-bound the word order of their utterances is, and what that rule order is. In clauses, many languages with strict rule order such as English, require a subject-verb-object order (Example (1-b)) in typical clauses; most languages of the world prefer subject-object-verb order (Example (1-a)) in-

stead, and many languages use verb-subject-object (Example (1-c)). The other three orderings are quite rare in comparison. Languages with comparatively free word order still invariably exhibit a preference for a standard word order which is used when there is no reason to diverge from it, e.g. for reasons of topical emphasis.

(1) *a.* *Caesar aleas amat.*
Caesar dice loves (Latin)

b. *The slow fox caught the early worm.*

c. *Phóg an fear an muc.*
Kissed the man the pig. (Irish)

With respect to single constituents, languages vary in how they organise a head word and its attributes. Adjectives can precede (Examples (2-b) and (2-c)) the noun they modify or come after (Example (2-a)); a language may prefer prepositions to postpositions.

(2) *a.* *Un vin blanc sec*
A wine white dry (French)

b. *An unsurprising sample*

c. *Bar mleczny w Częstochowie*
Bar milk in Czestochowa (Polish)

d. *A hegedű a zongora mögött van*
The violin the piano behind is (Hungarian)

For any information based on more elaborate analyses than bags of words, these variations will impact the results. If e.g. a system automatically recognises multi-word phrases, word order will make a difference; if the tasks move beyond information retrieval to e.g. information extraction, sentiment analysis or other tasks, where more than word counts are instrumental to the analysis, an analysis step to identify head with respect to attribute will be necessary.

2.5 Ellipsis and Anaphora

Elliptic references in human language include omission of words that are obviously understood, but must be added to make the construction grammatically complete. Human language users avoid repetition of referents, replacing something known by a pronoun, and sometimes omit the referent entirely. The ways in which this is done

vary somewhat over languages and genres. Samples (3) are in English, with omitted bits in square brackets.

(3) a. *Kal does not have a dog but Ari does [have a dog]*
b. *I like Brand A a lot. But on the whole, Brand B is better [than Brand A].*
c. *Bertram makes deep-V hulls. It [Bertram] takes sea really well.*

Elliptic references are challenging from the point of view of information retrieval, because search words may be omitted in the text (Pirkola and Järvelin, 1996). Such omissions will impact retrieval efficiency in that the relative frequencies of terms implicitly understood by the author and reader of a text may be underrepresented by an indexing tool. This effect is likely to be marginal, but more importantly, analyses and tasks with more semantic sophistication, which depend on associating a feature or characteristic with some referent will be difficult unless the referent in question is explicitly mentioned. Sentiment analysis (Steinberger et al, 2011) and keyword proximity based retrieval (Pirkola and Järvelin, 1996) are examples.

2.6 Digitisation of Collections and Historical Depth

When originally non-digital material, such as old newspapers and books, are digitized, the process starts with the documents scanned into image files. From these image files one needs to sort out texts and possible non-textual data, such as photographs and other pictorial representations. Texts are recognized from the scanned pages with Optical Character Recognition (OCR) software. OCR for modern text types and fonts is considered to be a solved problem that yields high quality results, but results of historical document OCR are still far from that level (Piotrowski, 2012). Most recently, Springmann and Lüdeling (2017) report high word-level recognition accuracies (ranging from 76% to 97%) based on applying trainable Neural Network-based OCR to a diachronic corpus of scanned images of books printed between 1478 and 1870. This type of corpus is especially demanding for OCR due to many types of variation present in the manuscripts — including linguistic changes (e.g., spelling, word formation, word order) and extra-linguistic changes (e.g., medium, layout, scripts, and technology).

Digitization of old books, newspapers and other material has been an on-going effort for more than 20 years in Europe. Its results can be seen e.g. in large multilingual newspaper collections, such as Europeana (http://www.europeana-newspapers.eu/). Europeana contains 18 million pages in 16 languages (Pletschacher et al, 2015). Scandinavian countries, e.g., have available over 80 million pages of digitized historical newspapers (Pääkkönen et al, 2018). Single newspaper archives, such as Times of London 1785–2012, or La Stampa 1867–2005, can already contain several million or over 10 million pages.

Europeana has estimated word level quality of its contents. For most of the included major languages, word correctness rate is about 80% or slightly more, but

for Finnish, Old German, Latvian, Russian, Ukrainian and Yiddish, correctness rates are below 70% (Pletschacher et al, 2015). Thus smaller languages and content published in more complicated scripts may have a disadvantage in their quality.

OCR errors in the digitized newspapers and journals may impact collection quality. Poor OCR quality obviously renders documents from the collections less readable and comprehensible for human readers but also less amenable to on-line search and further natural language processing or analysis (Taghva et al, 1996; Lopresti, 2009). Savoy and Naji (2011), for example, showed how retrieval performance decreases with OCR error corrupted documents quite severely.

The same level of retrieval quality decrease is shown in results from the confusion track at TREC 5 (Kantor and Voorhees, 2000). The end result effect of OCR errors is not clear cut, however. Tanner et al (2009) suggest that word accuracy rates less than 80% are harmful for search, but when the word accuracy is over 80%, fuzzy search capabilities of search engines should manage the problems caused by word errors. The probabilistic model developed by Mittendorf and Schäuble (2000) for data corruption seems to support this, at least for longer documents and longer queries. Empirical results by Järvelin et al (2016) on a Finnish historical newspaper search collection show that fuzzy matching will help only to a limited degree if the collection is of low quality.

One aspect of retrieval performance of poor OCR quality is its effect on ranking of the documents (Mittendorf and Schäuble, 2000): badly OCRed documents may be quite low in the result list if they are found at all. In practice these kinds of drops in retrieval and ranking performance mean that the user will lose relevant documents: either they are not found at all by the search engine or the documents are so low in the ranking list that the user may never reach them while browsing the result list. Some examples of this in the work of digital humanities scholars are discussed e.g. by Traub et al (2015)

Correcting OCR errors in a historical corpus can be done at access time or at indexing time by filtering index terms through authoritative lexical resources, pooling the output from several OCR systems (under the assumption they make different errors) or using distributional models to find equivalents for unknown words. These are all methods tested and used for OCR correction. As observed by Volk et al (2011), built-in lexicons of commercial OCR systems do not cover 19th century spelling, dialectal or regional spelling variants, or proper names of e.g. news material from previous historical eras. Afli et al (2016) propose that statistical machine translation can be a beneficial method for performing post-OCR error correction for historical French.

3 Reliance on Resources

Languages with few developed language technology resources are sometimes called *low-density languages*. While the concept is somewhat vague, it can be useful in as much as it makes clear that languages with a small number of speakers may be well

served by language technology, whereas widely used languages may or may not be considered low-density. Examples of early studies in African low-density languages in cross-language information retrieval include (Cosijn et al, 2004) (Afrikaans-English) and (Argaw et al, 2004, 2005; Argaw and Asker, 2006; Argaw, 2007) (Amharic-English). Both explored the effectiveness of query translation utilizing topic (source) word normalization, bilingual dictionary-lookup, and removal of stop words as process components. The first study reports the development of a simplified Afrikaans normalizer; the latter used semi-automatic Amharic stemming (prefix and suffix stripping).

3.1 Dictionaries and Lexical Resources

Various types of lexical resources are necessary in Natural Language Processing (NLP). Monolingual dictionaries are used in morphological analysis for producing lemmas, and for decomposing compound words—and as the necessary step for subsequent phases in NLP, e.g., for recognising noun phrases or names; for recognising the target of some expressed attitude; or for extracting emerging topics from a stream of text. Not least in translating queries or other specifications of information needs, dictionaries will form a crucial component (Pirkola et al, 2001; Hedlund et al, 2004).

Synonym dictionaries or thesauri are used for expanding queries, to add recall to a narrowly posed information need. Bilingual dictionaries may be intended either for human readers (and thus contain verbose definitions) or alternatively intended for automatic translation components (transfer dictionaries) either for text translation or for e.g. query translation. It is a non-trivial problem to transfer a bilingual dictionary intended for humans into a transfer dictionary (Hull and Grefenstette, 1996).

3.2 Automatic Machine Translation of Queries and the Challenge of Out-of-Vocabulary Terms

Over the years at CLEF and elsewhere, many researchers have performed and continuously perform experiments to use existing automatic and semi-automatic machine translation resources to translate queries. Various technologies have been tested against each other, against manual human translation, against translated indexes, or against translated target documents (Airio, 2008). The quality of retrieval results, noted by practically all such studies, depends on two factors. Firstly, that publicly available translation resources are primarily intended to provide a *crude* translation designed for human readers, not a *raw* translation optimised for continued editing or use in further processes such as retrieval (Karlgren, 1981). Translations by web resources tend to resolve ambiguities with this in mind, and thus occasionally reducing information present in the original query. This can be ame-

liorated by systems that use other lexical resources to enrich the translated query (Herbert et al, 2011; Leveling et al, 2009; Saleh and Pecina, 2016).

Secondly, and more obviously, coverage of the translation resource. Out Of Vocabulary (OOV) words, i.e., words not found in translation dictionaries, are the major challenge for CLIR, machine translation, and other multilingual language processing tasks and information systems where translation is part of the system. In particular in scientific and technical domains OOV words are often keywords in texts, and if the system is unable to translate the most important words its effectiveness may substantially decrease. Proper names form another word category causing translation problems: while they should not be translated in principle, their surface forms in different languages may differ due to transliteration and inflection. The tools to handle OOV translation include: approximate string matching (fuzzy matching) through methods such as Soundex, character-level n-grams (skip-grams), and edit distance; reverse transliteration e.g. as in Transformation rule based translation in which a word in one language (e.g. Finnish *somatologia* $\longleftrightarrow$ English *somatology* or Finnish *Tsetsenia* $\longleftrightarrow$ English *Chechnya*) based on the regular correspondences between the characters in spelling variants (Pirkola et al, 2003; Toivonen et al, 2005; Pirkola et al, 2007).

4 Challenges

The challenges entailed by cross-linguistic variation can be summarized to be about resources: lack of them, cost of acquiring and maintaining them, and low utility of seemingly relevant tools developed e.g. by computational linguists. Tools built by computational linguists do not always improve results on large scale information processing tasks, since they are built for a different purpose than information access.

While the field of information access research has human communicative behaviour as its main object of study and processing texts and other human communicative expressions to understand their content, linguistic theory has as its goals to explain the structure and regularities of human language. These goals are related but are not perfectly aligned. Obviously linguists would do well to validate their theories by application to information access, but they lack an understanding of what needs are prioritised; information access researchers must formulate requirements for better analyses for computationally oriented linguists to work on, and these requirements need to be formulated at an operationally adequate level of abstraction. These discussions and analyses are what CLEF and other related forums are for; the output could be communicated in clearer terms, in the form of clearly formulated usage scenarios or use cases, for further discussions with application-minded linguists.

Table 2 Challenges in utilizing various resources in information access

Resource or technology	Monolingual Information Retrieval	Crosslingual and Multilingual Information Retrieval
Lexicons or translation dictionaries	-need to create resources per se (especially in low-density languages) vocabulary issues: -insufficient coverage -domain-specific needs (e.g., social media, historical texts, etc.) -OOV words (e.g., proper names) -control and cost of updating	-need to create transfer dictionaries appropriate for CLIR vocabulary issues: -insufficient coverage -excessive number of translations -domain-specific needs -OOV words -control of updating the vocabularies -cost of updating
Stop word lists	-need to create and tune stop word vocabularies for the particular application and domain	-same as in monolingual case (but for both source and target languages)
Normalising and lemmatisation methods	-vocabulary issues (in lemmatisation) -understemming, overstemming, and incorrect processing (in stemming) -linguistically correct processing may be inappropriate from the point of view of IR (generation of nonsense words)	-vocabulary issues (coverage, updating, etc.) -need to detect and translate multi-word phrases (in phrase-oriented languages) -need to decompound and translate compound words written together (in compound-oriented languages)
Fuzzy string matching	-applicability may be language-specific -effectiveness and efficiency issues	-applicability may be language pair-specific -effectiveness and efficiency issues
Generative methods	-need to design and implement the method (in low-density languages) -relatively high number of potential candidate words created (in highly inflectional languages) – efficiency issues -challenges of special domains (e.g., creating expressions matching noisy OCR text)	same as in monolingual case -here the idea is to generate query expansions for the target language in which normalization or lemmatization may not be available or appropriate (e.g., in web domain)
Comparable corpora	(not applicable)	-availability of appropriate corpora for the particular language-pairs in need -appropriateness of the alignment methods

References

Afli H, Qiu Z, Way A, Sheridan P (2016) Using smt for ocr error correction of historical texts. In: 10th international conference on Language Resources and Evaluation, LREC

Airio E (2008) Who benefits from CLIR in web retrieval? Journal of Documentation 64(5)

Akmajian A, Demers R, Farmer A, Harnish R (1995) Linguistics: An introduction to language and communication, 4th edn. MIT press, Cambridge, Massachusetts

Argaw AA (2007) Amharic-English Information Retrieval with Pseudo Relevance Feedback. In: Nardi A, Peters C, Ferro N (eds) CLEF 2007 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1173/

Argaw AA, Asker L (2006) Amharic-english information retrieval. In: Workshop of the Cross-Language Evaluation Forum for European Languages, Springer, pp 43–50

Argaw AA, Asker L, Cöster R, Karlgren J (2004) Dictionary-based amharic–english information retrieval. In: Workshop of the Cross-Language Evaluation Forum for European Languages, Springer, pp 143–149

Argaw AA, Asker L, Cöster R, Karlgren J, Sahlgren M (2005) Dictionary-based amharic-french information retrieval. In: Workshop of the Cross-Language Evaluation Forum for European Languages, Springer, pp 83–92

Chen A (2001) Multilingual information retrieval using English and Chinese queries. In: Workshop of the Cross-Language Evaluation Forum for European Languages, Springer, pp 44–58

Chen A (2002) Cross-language retrieval experiments at clef 2002. In: Workshop of the Cross-Language Evaluation Forum for European Languages, Springer, pp 28–48

Cosijn E, Keskustalo H, Pirkola A, De Wet K (2004) Afrikaans-english cross-language information retrieval. In: Bothma T, Kaniki A (eds) Proceedings of the 3rd biennial DISSAnet Conference, Pretoria, pp 97–100

Cöster R, Sahlgren M, Karlgren J (2003) Selective compound splitting of swedish queries for boolean combinations of truncated terms. In: Workshop of the Cross-Language Evaluation Forum for European Languages, Springer, pp 337–344

Dahl Ö (2004) The growth and maintenance of linguistic complexity, vol 71. John Benjamins

Dryer MS, Haspelmath M (2011) The World Atlas of Language Structures Online. Max Planck Digital Library, München, URL http://wals.info

Gollins T, Sanderson M (2001) Improving cross language retrieval with triangulated translation. In: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 90–95

Harman D (1991) How Effective is Suffixing? Journal of the American Society for Information Science 42:7–15

Hedlund T (2002) Compounds in dictionary-based cross-language information retrieval. Information Research 7(2):7–2

Hedlund T, Keskustalo H, Pirkola A, Sepponen M, Järvelin K (2000) Bilingual tests with swedish, finnish, and german queries: Dealing with morphology, compound words, and query structure. In: Workshop of the Cross-Language Evaluation Forum for European Languages, Springer, Berlin, Heidelberg, pp 210–223

Hedlund T, Pirkola A, Järvelin K (2001) Aspects of swedish morphology and semantics from the perspective of mono-and cross-language information retrieval. Information Processing & Management 37(1):147–161

Hedlund T, Airio E, Keskustalo H, Lehtokangas R, Pirkola A, Järvelin K (2004) Dictionary-based cross-language information retrieval: Learning experiences from clef 2000–2002. Information Retrieval 7(1-2):99–119

Herbert B, Szarvas G, Gurevych I (2011) Combining query translation techniques to improve cross-language information retrieval. In: Proceedings of the 33d European Conference on Information Retrieval, Springer

Hull DA, Grefenstette G (1996) Querying across languages: a dictionary-based approach to multilingual information retrieval. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 49–57

Järvelin A, Keskustalo H, Sormunen E, Saastamoinen M, Kettunen K (2016) Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach. Journal of the Association for Information Science and Technology 67(12):2928–2946

Kamps J, Monz C, De Rijke M, Sigurbjörnsson B (2004) Language-dependent and language-independent approaches to cross-lingual text retrieval. In: Peters C, Braschler M, Gonzalo J, Kluck M (eds) Comparative Evaluation of Multilingual Information Access Systems: Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003) Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 3237, Springer, Heidelberg, Germany

Kantor PB, Voorhees EM (2000) The TREC-5 confusion track: Comparing retrieval methods for scanned text. Information Retrieval 2(2):165–176

Karlgren H (1981) Computer aids in translation. Studia Linguistica 35(1-2):86–101
Karlgren J (2005) Compound terms and their constituent elements in information retrieval. In: Proceedings of the 15th Nordic Conference of Computational Linguistics (NoDaLiDa)
Karlgren J (ed) (2006) New Text—Wikis and blogs and other dynamic text sources. Proceedings of the EACL06 workshop. European Chapter of the Association for Computational Linguistics
Karlgren J, Dalianis H, Jongejan B (2008) Experiments to investigate the connection between case distribution and topical relevance of search terms. In: 6th international conference on Language Resources and Evaluation, LREC
Kettunen K (2009) Reductive and generative approaches to management of morphological variation of keywords in monolingual information retrieval: an overview. Journal of Documentation 65(2):267–290
Kettunen K (2014) Can type-token ratio be used to show morphological complexity of languages? Journal of Quantitative Linguistics 21(3):223–245
Kettunen K, Airio E (2006) Is a morphologically complex language really that complex in full-text retrieval? Advances in Natural Language Processing
Kettunen K, Airio E, Järvelin K (2007) Restricted inflectional form generation in morphological keyword variation. Information Retrieval 10
Lehtokangas R, Airio E, Järvelin K (2004) Transitive dictionary translation challenges direct dictionary translation in clir. Information processing & management 40(6):973–988
Lennon M, Peirce DS, Tarry BD, Willett P (1981) An evaluation of some conflation algorithms for information retrieval. Information Scientist 3(4)
Leveling J, Zhou D, Jones GJF, Wade V (2009) TCD-DCU at TEL@CLEF 2009: Document Expansion, Query Translation and Language Modeling. In: Borri F, Nardi A, Peters C, Ferro N (eds) CLEF 2009 Working Notes, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1175/
Lewis MP, Simons GF, Fennig CD, et al (2009) Ethnologue: Languages of the world, vol 16. SIL international, Dallas, Texas, most recent version at: http://www.ethnologue.com
Lieber R, Štekauer P (2009) The Oxford handbook of compounding. Oxford University Press
Lopresti D (2009) Optical character recognition errors and their effects on natural language processing. International Journal on Document Analysis and Recognition (IJDAR) 12(3):141–151
Lovins JB (1968) Development of a stemming algorithm. MIT Information Processing Group, Electronic Systems Laboratory Cambridge
Lowe TC, Roberts DC, Kurtz P (1973) Additional text processing for on-line retrieval (the radcol system). volume 1. Tech. rep., DTIC Document
McNamee P, Nicholas C, Mayfield J (2009) Addressing morphological variation in alphabetic languages. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 75–82
Mittendorf E, Schäuble P (2000) Information retrieval can cope with many errors. Information Retrieval 3(3):189–216
Pääkkönen T, Kettunen K, Kervinen J (2018) Digitisation and digital library presentation system–a resource-conscientious approach. In: Proceedings of 3d Conference on Digital Humanities in the Nordic Countries, CEUR-WS.org, pp 297–305
Piotrowski M (2012) Natural language processing for historical texts. Synthesis Lectures on Human Language Technologies 5(2):1–157
Pirkola A (2001) Morphological typology of languages for IR. Journal of Documentation 57(3):330–348
Pirkola A, Järvelin K (1996) The effect of anaphor and ellipsis resolution on proximity searching in a text database. Information processing & management 32(2):199–216
Pirkola A, Hedlund T, Keskustalo H, Järvelin K (2001) Dictionary-based cross-language information retrieval: Problems, methods, and research findings. Information retrieval 4(3-4):209–230
Pirkola A, Toivonen J, Keskustalo H, Visala K, Järvelin K (2003) Fuzzy translation of cross-lingual spelling variants. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, pp 345–352

Pirkola A, Toivonen J, Keskustalo H, Järvelin K (2007) Frequency-based identification of correct translation equivalents (fite) obtained through transformation rules. ACM Transactions on Information Systems (TOIS) 26(1):2

Pletschacher S, Clausner C, Antonacopoulos A (2015) Europeana newspapers OCR workflow evaluation. In: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing, ACM, pp 39–46

Popović M, Willett P (1992) The effectiveness of stemming for natural-language access to slovene textual data. Journal of the American Society for Information Science 43

Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137

Porter MF (2001) Snowball: A language for stemming algorithms

Rehm G, Uszkoreit H (2012) Meta-net white paper series: Europe's languages in the digital age

Saleh S, Pecina P (2016) Reranking Hypotheses of Machine-Translated Queries for Cross-Lingual Information Retrieval. In: Fuhr N, Quaresma P, Gonçalves T, Larsen B, Balog K, Macdonald C, Cappellato L, Ferro N (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventh International Conference of the CLEF Association (CLEF 2016), Lecture Notes in Computer Science (LNCS) 9822, Springer, Heidelberg, Germany, pp 54–68

Savoy J, Naji N (2011) Comparative information retrieval evaluation for scanned documents. In: Proceedings of the 15th WSEAS international conference on Computers, pp 527–534

Springmann U, Lüdeling A (2017) OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES corpus. Digital Humanities Quarterly 11(2)

Steinberger J, Lenkova P, Kabadjov MA, Steinberger R, Van der Goot E (2011) Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In: Recent Advances in Natural Language Processing, pp 770–775

Taghva K, Borsack J, Condit A (1996) Evaluation of model-based retrieval effectiveness with OCR text. ACM Transactions on Information Systems (TOIS) 14(1):64–93

Tanner S, Muñoz T, Ros PH (2009) Measuring mass text digitization quality and usefulness. lessons learned from assessing the ocr accuracy of the british library's 19th century online newspaper archive. D-lib Magazine 15(7/8):1082–9873

Toivonen J, Pirkola A, Keskustalo H, Visala K, Järvelin K (2005) Translating cross-lingual spelling variants using transformation rules. Information processing & management 41(4):859–872

Traub MC, van Ossenbruggen J, Hardman L (2015) Impact analysis of OCR quality on research tasks in digital archives. In: Kapidakis S, Mazurek C, Werla M (eds) International Conference on Theory and Practice of Digital Libraries, Lecture Notes in Computer Science (LNCS) 9316, Springer, Heidelberg, Germany, pp 252–263

Uryupina O, Plank B, Severyn A, Rotondi A, Moschitti A (2014) Sentube: A corpus for sentiment analysis on youtube social media. In: 9th international conference on Language Resources and Evaluation, LREC

Velupillai V (2012) An introduction to linguistic typology. John Benjamins Publishing

Volk M, Furrer L, Sennrich R (2011) Strategies for reducing and correcting ocr errors. Language Technology for Cultural Heritage pp 3–22