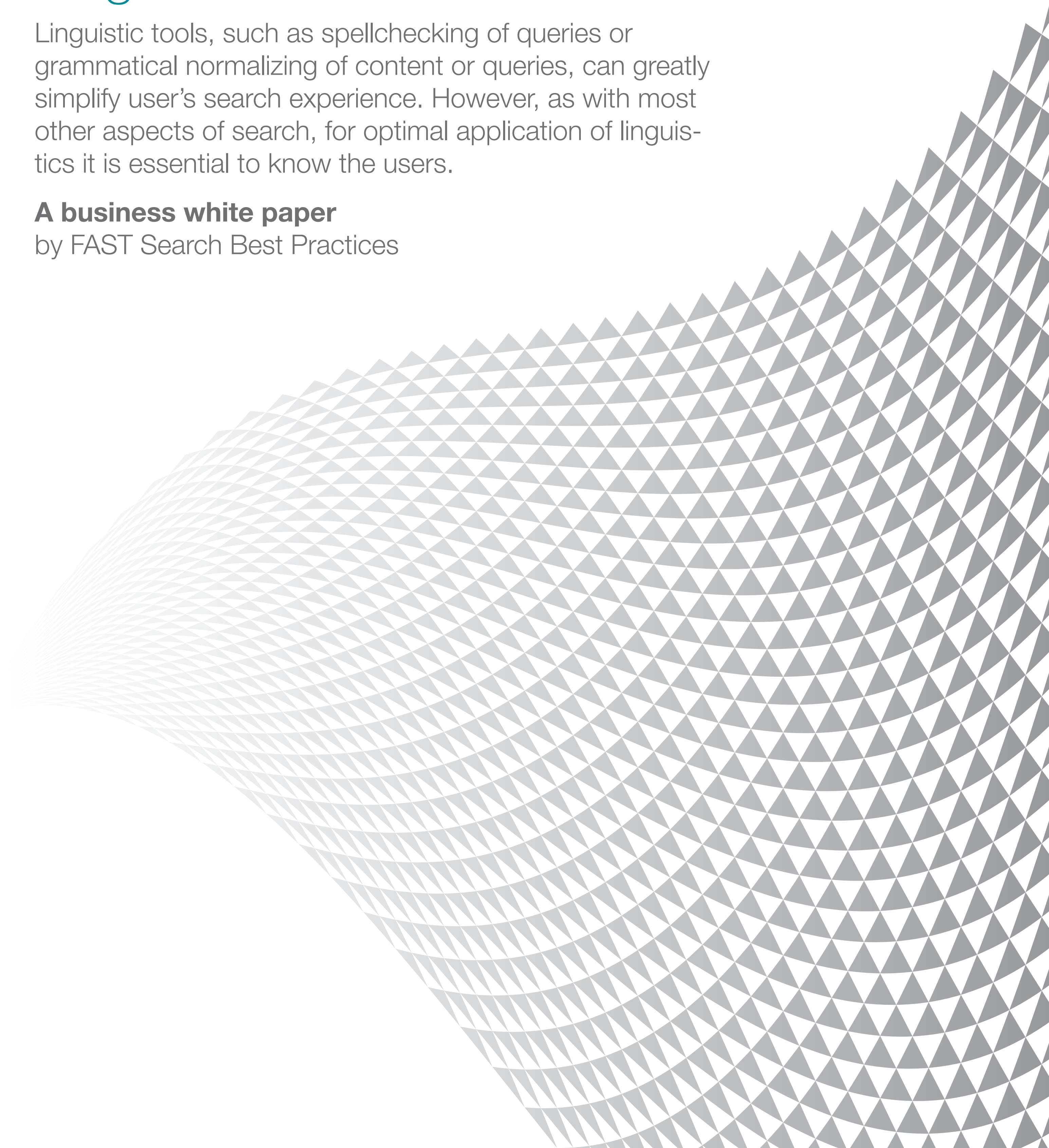
fast SBP™

Linguistics and search

Linguistic tools, such as spellchecking of queries or grammatical normalizing of content or queries, can greatly simplify user's search experience. However, as with most other aspects of search, for optimal application of linguistics it is essential to know the users.

A business white paper
by FAST Search Best Practices



5 things you should know about linguistics

1. Better use of linguistics will improve precision and recall
2. Industry and user knowledge are needed to optimize search systems
3. Linguistic choices can impact hardware and performance
4. Some sites should favour language independence
5. Bad queries can be turned into good queries with the proper linguistic tools

Why is linguistics important for good search?

In the search world, linguistics is defined as the use of information about the structure and variation of languages so that users can more easily find relevant information. This is important for properly using search tools in various "natural" languages; think of the structural differences between English and Chinese. It is also important to industry-specific language usage – for instance, the English used in an American pharmaceutical company versus that used in a Hong Kong-based investment bank.

Another use of linguistic tools is to determine the intent behind keywords. For example, when a researcher enters "When was D-Day?" she is looking for information that resembles a date. Another example: someone who is searching an e-commerce site for "MP3 players" would also be interested in hits that matched "MP3 player" or "iPod." If the site shows only results for the keywords "MP3" and "players," it may easily miss out on a sale. The first example improves search precision by combining semantic analysis (understanding and interpreting the search terms) with entity extraction (isolating known linguistic constructs). The second provides better recall by using lemmatization (associating variations of word forms) and synonym expansion. However, these tools may not always improve the user experience: a stock-market trader searching for the stock quote 'BAB' would just get annoyed if he was asked "Did you mean BAD?" or swamped by hits about "babies" rather than simply being given information on the British Airways share price.

Clearly, a grasp of the linguistic features of advanced search engines, coupled with knowledge about the data and users of the search service, can greatly improve precision and recall and thus yield better business results. This white paper will lay out the main options for those building a search service and bring forward the main questions that must be answered soon.

Q: Five percent of my users and content are non-English. What should I do differently for them?

A: You will need to think carefully about the language-specific features of your search function. If people search only for content in their own language, or there is wide variation in the language types used (English, Polish, and Chinese for example), then it will help to have users specify their language in the query interface. Where there are common linguistic roots - on an e-commerce site featuring English and Dutch content, say - it may be easier to handle everything in the most common language – in this case, English.

The many aspects of linguistic analysis

When designing the linguistic elements of an advanced search engine installation, the designers must consider different levels of sophistication of features in terms of their impact on the user experience, on the administration of the system, and on the complexity of the user interface.

To begin with, there are unilingual and multilingual installations, i.e. those that treat all documents as being either of the same language or of no language, and those that treat them separately. Within the multilingual category there are different degrees to which these languages can be treated separately, and there are systems that combine very different languages - for example, European languages with Chinese, Japanese, or Korean (CJK), all of which have special requirements (there is no symbol used for word delimitation so advanced tokenization algorithms must also be used).

Crucially, given the differences between languages, for some language-specific features to work, it's necessary to distinguish between them before each document enters the index, either manually or by using automatic detection mechanisms. Once that has been done, the document index can be augmented by extracting or adding information. The most common augmentation is lemmatization - the expansion of all known terms in the document to their inflectional forms or their base form. Lemmatization enables searches to match documents with similar meanings but different keywords.

Q: My software features a search engine and my customers worry about the installation footprint. What can I do with linguistics to minimize my hardware usage?

A: Disk usage will vary with the amount of index-time expansion (entity extraction, synonyms, etc.) but it may reduce the QPS (queries per second) or functionality available to end users. There are two classic trade-offs to consider: ingestion rate vs. query speed and recall vs. index size.

The document index can also be augmented by using synonyms, which correlate words in a much broader sense and allow for industry-specific terminology and acronyms. And it can be streamlined by the removal of stop words (words which are frequent but have little meaning to the search), which also aids recall. For example, after stopword removal, a query for "President of United States" would also match documents mentioning "The President of the United States".

For languages that have diacritic characters (like Romance or Scandinavian languages) it often makes sense to consider character normalization (the mapping of diacritic characters to standard characters like {é,è,ê} à e) in order to increase recall. That said, in some search engines it is possible to preserve and search against diacritics.

More advanced interpretation of language can be done through entity extraction (the spotting through the combined use of dictionaries and syntactical patterns of certain entities such as people, places, product codes, prices, etc.), parts of speech detection, and sentiment analysis (the evaluation of the text's sentiment - typically positive or negative - based on the usage of language). Categorization (the identification of documents as being part of an ontology or taxonomy from rule-based or conceptual category definitions) and the unsupervised clustering of documents (grouping related documents on the basis of their content without referring to a taxonomy) also leverage document semantic and conceptual interpretation.

It's important to note that some of the techniques described above can be applied at index time or at query time by modifying the user's search. The choice of where best to integrate lemmatization or synonym expansion, for example, usually comes down to considerations of performance, practicality, and flexibility.

Linguistic features can also be focused on the query side, modifying the search and turning a "bad" search into a "good" one.

An obvious example: spelling correction (either automatic or "did you mean…?"). Other features aim to improve queries too. Phrasing (the recognition and grouping of an idiom such as "home run") and anti-phrasing (identifying word sequences in queries that are irrelevant to the search) are good examples. And the tuning of the relevancy model can be used to increase or decrease the importance of statistical linguistic analysis (i.e. tf-idf, which uses the relative frequency of words in a corpus of data to determine their importance).

Q: I run a mass market classifieds Web site. How can you help me make my interface simpler?

A: Let's say you want people to search for "Two bed flat in midtown Manhattan with two bathrooms" and they find "Midtown - 2 bed apt. 1 en-suite". They key is to get the index to do the work instead of your users having to fill out forms, using custom thesauri, entity extraction, etc. Also you're running a consumer service so don't forget to design the system with high query volumes and do as much processing index-side as possible.

Phonetic search can be applied to structured searches such as name searches. In this case, it isn't the morphological variants or synonyms that are identified; it's the words that are pronounced similarly. For example, phonetic search will detect all possible variants of "Muamar Gadaffi"(Muammar Al Ghaddafi; Muammar Al Qaddafi; Muammar Al Qaddafi; Muammar El Qaddafi; Muammar Gadaffi; Muammar Gadafy; Muammar Gadafy, etc.)

Linguistic methods are also used when performing speech-to-text. By analysing the phonetic models of speech, along with knowledge of the nature and distribution of words within common languages, modern transcription software is able to create an accurate dictation of the audio track of multimedia files. This in turn enables the audio and video files to be searchable alongside other more typical formats such as pdfs or Word documents.

Different industries, different solutions

With such an extensive list of linguistic options, the key to the best user experience in terms of precision and recall is to work out what permutation of them best suits the business needs of the search application.

For instance, in the case of an e-commerce application, the business driver is to make sure the user finds the product he is willing to purchase. It's important to include a dedicated thesaurus specific to the types of products being sold (i.e. "MP3 player" ➔ "iPod"). Automatic spelling correction is another invaluable feature to ward off user frustration: a user who accidentally types "floghts to London" should still be taken to responses to a search for "flights to London" rather than having to click on a "did you mean…?" dialog box (whether the application automatically corrects or prompts is an interface design decision). And lemmatization and spell-checking must be used with great care. Someone looking for "Golf GTIs" should not be steered towards "golfing gifts". Ideally, for e-commerce, custom spelling dictionaries should be maintained and few or no lemmatization and stop-words should be used.

At the opposite end of the spectrum, extensive lemmatization and stop-wording are essential when performing knowledge discovery over verbose unstructured data where meaning is more important than simple keyword matching. Entity extraction will also play a much bigger part to ease navigation through larger amounts of data.

> **"From a search user's point of view, the usefulness of linguistic search assistants will depend on their level of expertise in the field as well as their familiarity with search. Therefore an understanding of users is essential from the content owner's point of view"**

Language-specific functions such as entity extraction and lemmatization require knowing the language of documents and queries to be performed optimally. Therefore OEM integrations of a search application require caution about configurations that assume this knowledge, since the embedding application will be used in a variety of contexts without the opportunity for tuning.

What's the effect of linguistics?

Linguistics aims to leverage the meaning of documents or words outside of the keyword paradigm. It transforms queries – a valuable feature since users type only one or two (often miss-spelt) words on average.

To determine the success of a search application from a linguistics point of view, the most important metric is the number of empty result sets returned. If the number is too high, features such as lemmatization, synonyms, and spelling dictionaries should be investigated or refined. However, this requires that application manager take a proactive interest in the user's search experience and work closely with industry experts and librarians to leverage their knowledge of the content. From a search user's point of view, the usefulness of such search assistants will depend on their level of expertise in the field as well as their familiarity with search.

A novice user or one in an unfamiliar domain will favour an interface with the least number of options to allow natural language searching. A librarian or expert will want to be able to select for themselves whether to use the synonym dictionary, and they will use the date field if searching for a date rather than relying on the natural language processing algorithm to translate "When was…".

This comes back to the key point: an understanding of users and their needs is essential from the content owner's point of view. This will determine what types of entities should be marked up (the "When was D-Day" example relies of the system having marked up all date type content in situ so dates near the word "D-Day" can be identified) and whether to do lemmatization by expansion or reduction.

Lemmatization by expansion will give the best query performance and is required when the language of the searcher is not known (for example, when there is one corporate search service in a multinational enterprise). However, this increases the index size, which is costly for certain languages such as Finnish or Hungarian.

On the other hand, lemmatization by reduction or stemming modifies both the index and query terms so the language of the search must be known to avoid spurious and ambiguous results. This approach is recommended though if storage space is an issue.

Synonym expansion at query time will affect performance, but it will allow the search system's administrators to modify their thesauri when they want without the need to re-index – a big difference from doing the expansion before indexing.

Mini case study

Newspaper network taps advanced linguistic search features to drive growth in classified ads

Who
Major American newspaper network with growing Web presence

Challenge
Make Web site more profitable by generating revenue from classifieds.

Solution
Create a smart interface to allow highly structured searches and alerts for users who know what they're looking for. At the same time, allow natural language search over text with highly specialised acronyms and abbreviations.

Technology
Entity extraction and customized synonyms to power navigators for advanced search, and language processing for query-side novice buyers.

Guidelines and recommendations

Many common mistakes are made when configuring the linguistics of a search engine.

Firstly, there is the failure to use the linguistics features of good search tools to their full potential, especially when the system has a broad and varied user base.

Another common pitfall is attempting to use these features without correctly identifying the language of the documents, and more importantly, the user's first language. It's also necessary to gauge what type of service you are trying to provide: a multilingual search, but "one size fits all", most likely optimized for the most common language; or a targeted system where content is segregated and tweaked according to its different sources.

Yet another area to watch out for is the mismatch of interacting components: for example, when using different processing on the index and query side, or when there are conflicting synonym and lemmatization or stemming methods. It is also important to understand the tradeoffs between hardware (disk, processor), index and query performance, and functionality.

The fundamental steps to improve search

From a language perspective, a poor search installation implies a failure to understand the nature of the user base or the importance of localizing to the industry or country when providing search over unstructured content. To help head off the consequences of such failures, content owners and search providers must ask themselves some simple questions:

What types of queries do my users perform (natural language, keyword, field-based)?

What languages are my documents in? What is their nature (structured, unstructured)?

What languages do my users speak? And do I have enough multilingual users to justify optimizing for them separately?

Is hardware a primary consideration over functionality?

Can I leverage my industry-specific knowledge to improve synonym and spelling and entity dictionaries?

Answering such questions and mapping the appropriate linguistic tools to the subsequent platform profile will help drive the success of search applications in fields such as e-commerce and knowledge discovery. Users will feel that the search engine is working for them; they will feel they've been "listened to" and feel confident that more relevant information is delivered to them. Content providers will see improved visibility for and consumption of their information, giving them greater productivity and increased e-business sales.

Frequently asked questions

Q: Do you need to worry about linguistics with structured data?

A: Usually, most linguistics features can be turned off. However, if your users want to do natural language queries, you will need some query-side processing.

Q: What is an ontology?

A: A specific set of knowledge related to a given domain – for example, pharmaceuticals.

Q: Why does the search tool need to know what language I'm typing in? Can't work it out by itself?

A: To perform synonym expansion or lemmatization at query time, it's necessary to know the language. Although automatic detection tools are used at index time, the queries are normally too short. Sometimes localization information can be used to determine this. If it does not prove useful, multiple dictionaries may be used at query time if the query is limited to a small number of languages.

Q: My data is in different encodings. How will multilingual search work?

A: Everything should be normalised to UTF-8 but with an important caveat. When importing dictionaries, it's valuable to normalize accented characters for simplicity.

Q: What is meant by language morphology?

A: The rules and semantics of the formation and structure of permissible words in a given language.

Q: What is a bad query?

A: A query that is too short and ambiguous for simple keyword matching to bring back relevant results.

About FAST SBP™ (Search Best Practices)

SBP consulting is a highly focused transfer of search knowledge and experience from FAST to its prospects and customers. SBP workshops aim to help enterprises realize the full potential of search, by creating optimal strategic, functional and technical roadmaps, delivered in the form of business model, solution and architecture designs.

Fast Search & Transfer
www.fastsearch.com
info@fastsearch.com

Regional Headquarters

The Americas
+1 781 304 2400

Europe, Middle East & Africa (EMEA)
+47 23 01 12 00

Japan
+81 3 5511 4343

Asia Pacific
+612 9929 7725