



SMART Communications, Inc., New York, offers text mining services to help companies extract hidden technical terminology. The key to *Simplified English* or *Controlled English* is the development of a technical dictionary for a product or service. The SMART team uses the **SMART Text Miner**, an AI Agent that prepares text for analysis. The text mining reports are used to curate the dictionaries.

The results are entered into three dictionaries:

- **Core** dictionary for common words, including a set of verbs. The verbs are based on ASD-STE100 that links to S3000L and S1000D.
- **Technical** dictionary for a product or service.
- **Synonym** dictionary that holds curated suggestions.

(The example below is a text mining report with terms in content.)

cannot be used. It also	extends	the graphite blocks'	25428	(1)
first grate or to the	external	drag chain system via a	25428	(1)
to conform to mating	faces	thus ensuring a tight	25428	(1)
unlikely event of a wire	failure	the graphite blocks will	25428	(1)
-Minimises	false	air	25428	(3)
-Minimises	false	air from entering the	25428	-
system consists of a	fan	and a closed isolated	25428	(2)
system consists of a	fan	a pneumatically operated	25428	-
such as kiln shell,	fans	and hoppers can easily	25428	(1)

Document Selection for Text Mining:

The first step gathers representative samples of existing texts. The text can be a combination of formats, .txt, .docx, .xml and .pdf. The **Text Miner** program ignores .xml <tags> and graphics. The sample size for text mining depends on the product complexity.

Text Mining operation:

The sample pages are automatically converted to plain text. All format codes and graphics are ignored. The **Text Miner** reads the texts and creates a report, shown above. The report shows the term in multiple textual contexts, word frequency and a document reference. The contextual information helps the linguists resolve ambiguity and multiple meanings. A “where found” index shows a reference to the original sample text.

Linguistic analysis and link to the Lexicon Manager

The SMART linguists analyze the text mining reports. Clients are consulted for technical terms. The final list is encoded in the **SMART Lexicon Manager** to create the dictionaries for the **MAXit Checker**.

Notes on the words marked in the sample:

- **extend** - approved verb
- **external** - approved adjective
- **mating face** - consult with engineering (valid or can use mating surface)
- **wire failure** - consult with engineering (be specific)
- **false air** - translation problem (can be unwanted air)
- **fan** - approved technical term

Client subject matter experts make the final decision.

