# Contextualized Usage-Based Material Selection

**Dirk De Hertog, Piet Desmet**
ITEC, KU Leuven, imec
Etienne Sabbelaan 51, 8500 Kortrijk, Belgium
{dirk.dehertog, piet.desmet}@kuleuven.be

**Abstract**

In this paper, we combine several NLP-functionalities to organize examples drawn from corpora. The application's primary target audience are language learners. Currently, authentic linguistic examples for a given keyword search are often organized alphabetically according to context. From this, it is not always clear which contextual regularities actually exist on a syntactic, collocational and semantic level. Showing information at different levels of abstraction will help with the discovery of linguistic regularities and thus improve linguistic understanding. Practically this translates in a system that groups retrieved results on syntactic grounds, after which the examples are further organized at the hand of semantic similarity within certain phrasal slots. Visualization algorithms are then used to show focused information in phrasal slots, laying bare semantic restrictions within the construction.

**Keywords:** Language Learning, Constructions, Corpus-based examples, Distributional Analysis

## 1. Introduction

We present the integration of NLP-functionalities to help with the selection of contextualized usage-based examples to assist language learners and other end users that could benefit from a distributional linguistic analysis. Usage-based material, in the form of examples drawn from corpora, like the Keyword in Context (KWIC) method, are thought of as valuable resources for studying languages. KWIC allows for the user to search for a keyword, retrieves a set of examples from a corpus and presents them ordered alphabetically according to context words. However, from examples alone, it is not always immediately clear which contextual regularities exist at the syntactic, collocational and semantic level. The lack of such additional structure and the required effort to derive it single-handedly have proven demotivational in practical settings.

We propose to use several NLP-methods to provide additional information by which the examples could be further organized in a meaningful way. This would add value as it would clarify linguistic regularities at a glance. Practically this translates in a system that looks for a certain keyword and groups the retrieved examples first according to syntactic grounds, then to semantic similarity within certain contextual phrasal slots.

By doing so, we aim to lessen the gap between corpus-based examples and cognitive understanding of linguistic regularities with the end-user. We start from the basic premise that learning presumes understanding (Krashen, 1981; Chapelle, 1996); in this respect, we argue that full disambiguation facilitates linguistic understanding and thus enhances word knowledge and improves vocabulary retention. The procedure is inspired by collostructional analysis (Anatol and Gries, 2003), which treats syntactic constructions as a disambiguating factor. For the semantic understanding of words, word sense is arguably linked to the actual construction in which it is used. Conversely, the intended word sense presupposes correct syntactic and/or collocational use for it to be understood. A retrieval that links both types of information should therefore prove highly informative.

We first explain what we interpret as cognitive understanding from a pedagogical and linguistic point of view. We then offer technical operationalizations to achieve the intended semantic disambiguation. Section 4 discusses expert opinions and proposes a future validation of the results. We wrap up with a short summary and point to ongoing and future work.

## 2. Towards Linguistic Understanding

Traditional teaching practice often distinguishes between acquisition of vocabulary and syntax. Syntax governs the grammatical rules found in a language, while vocabulary items fill certain syntactic slots following those grammatical rules.

Such a simple dichotomy is being challenged from different viewpoints. For example, within the field of Second Language Acquisition it is accepted that word knowledge is more complex than a simple slot-filler approach and also involves knowledge about lexical collocations, constructions and semantics (Nation, 2001). Within Cognitive Linguistics, the dichotomy is reinterpreted as a cline that ranges from the lexical to the syntactic, with semantic meaning and functional interpretation as a central linguistic principle (Croft, 2001).

From a linguistic point of view, Cognitive Linguistics defends a usage-based model of language. In such a model, all linguistic knowledge flows from the actual linguistic utterances a person encounters. It posits a mechanism of distributional interpretation of language into patterns (Tomasello, 2003), after which the patterns acquire a certain semantic interpretation themselves. Bybee (2006), for instance, exemplifies a mechanism of exemplar-based categorization of lexical items to result in grammatical patterns. The exemplars in their turn relate to the pattern as specific instantiations, called constructs. "The major idea behind exemplar theory is that the matching process has an effect on the representations themselves; new tokens of experience are not decoded and then discarded, but rather they impact memory representations." (p.716)

We aim to stimulate a similar process of linguistic understanding, acquired automatically by native speakers in second language learners, stimulating a distributional

understanding of an unknown word in a fully specified linguistic context (Nagy, Herman and Anderson, 1985; Sternberg, 1987). Moreover, it is widely accepted in Second Language Acquisition studies that exposure to new material stimulates vocabulary knowledge (Nation, 2001). Taking into account that the procedure requires a certain level of proficiency it seems best suited for advanced students that wish to expand on their existing vocabulary knowledge.

## 3. Selection Procedure

We emulate a full contextual disambiguation at the hand of two processing steps. The first uses the (abstract) constructions in which a word occurs and interprets it as a structural context. The second looks within that structure for analogue lexical examples that are semantically related. The double selection procedure thus highlights semantic constraints, shown through positive evidence in the form of examples, within a single construction. The proposed method has been applied to a compiled corpus of freely available literary works of approximately 100 million words in order to exemplify the procedure.

### 3.1. Constructional Constraints

In order to help the language learner discover constructional analogies (syntactic regularities) from examples, we propose to provide language input and enrich it with abstract constructional information. To derive abstract constructional information, we start from POS-tags, a word's syntactic category such as Noun or Verb, as formal superficial word-representations and derive contiguous patterns. Formally, a selected sequence of POS-tags thus stands for a construction, while the lexical word combinations matching that pattern are exemplar-based instantiations.

The selection of patterns is based on an actual example that poses difficulties for the learner. The retrieved examples from the corpus are matched to the syntagmatic axis of the problematic example on two levels: the target word (as a lexical center) and the syntactic constructions (constructional patterns) in which the lexical item functions.

Take for instance Ex. 1, in which *obvious* is set as the target word for which we want more distributional information.

Ex. 1. He made a rather **obvious** remark.

**Pron Verb Det Adv** Adj **Noun**

To get a better understanding of the word, we use corpus-based examples that share the keyword *obvious* and the constructions in which it participates. As such, rather than taking context-words such as *rather* and *remark*, as lexical collocations, we use them to determine the partly abstract constructions *Adv obvious* and *obvious Noun*. The corpus is then used to retrieve paradigmatically analogous constructs, resolving the abstract slot, as seen in Table 1.

The target word participates in a number of constructions that are increasingly complex. Using the number of different lexical types (as opposed to tokens) for a certain construction is a direct quantification of how prominent a particular structural context is for the word at hand. It is used for our purposes as an objective measure to quantify which constructional slots are most interesting to provide semantic structure for. From a practical perspective, those slots largely coincide with immediate dependencies of the target word. Ex. 2 shows a single construction, for which the numbers in brackets signify how many different word-types occur in the slot.

Ex. 2. Det (3) Adv (6) OBVIOUS Noun (14)

| | | | |
|---|---|---|---|
| | the | perfectly | fact |
| | the | apparently | consideration |
| | an | equally | conviction |

| Lexical Example | Derived Constructions | Analogous Constructs | Lexical Types |
|---|---|---|---|
| Obvious remark | Obvious Noun | Obvious point<br>Obvious reason<br>… | 290 |
| Rather obvious | Adv obvious | Fairly obvious<br>Very obvious<br>… | 58 |
| Rather obvious remark | Adv obvious Noun | Fairly obvious fact<br>Very obvious reason<br>… | 28 |

Table 1: Lexical Examples, Derived Constructions, Analogous Constructs and Lexical Types (output Step 1)

### 3.2. Semantic Similarity

To help the learner discover semantic structure in the retrieved constructions, we order the words according to their semantic similarity with the initial example. Because we see semantic restrictions as being related to certain constructional patterns, we aim to highlight the imposed paradigmatic restrictions with the given example in an unsupervised fashion.

We use a self-implemented version of Semantic Vector Spaces (SVS) (Lund and Burgess, 1996; Turney and Pantel, 2010; Mikolov, 2013) to order words according to certain criteria. A vector space model is defined by three distinct parameters: the contexts included as features, a weighting function for the collocational corpus counts, and a similarity metric to compare the vectors. Each parameter has an impact on the resulting measure of similarity. For instance, choices for the first parameter include the inclusion/exclusion of function words as context features and the window size that states how many context words surrounding the target are taken into account. Small context sizes that include function words will directly shift the vector space to capture a more syntactic interpretation,

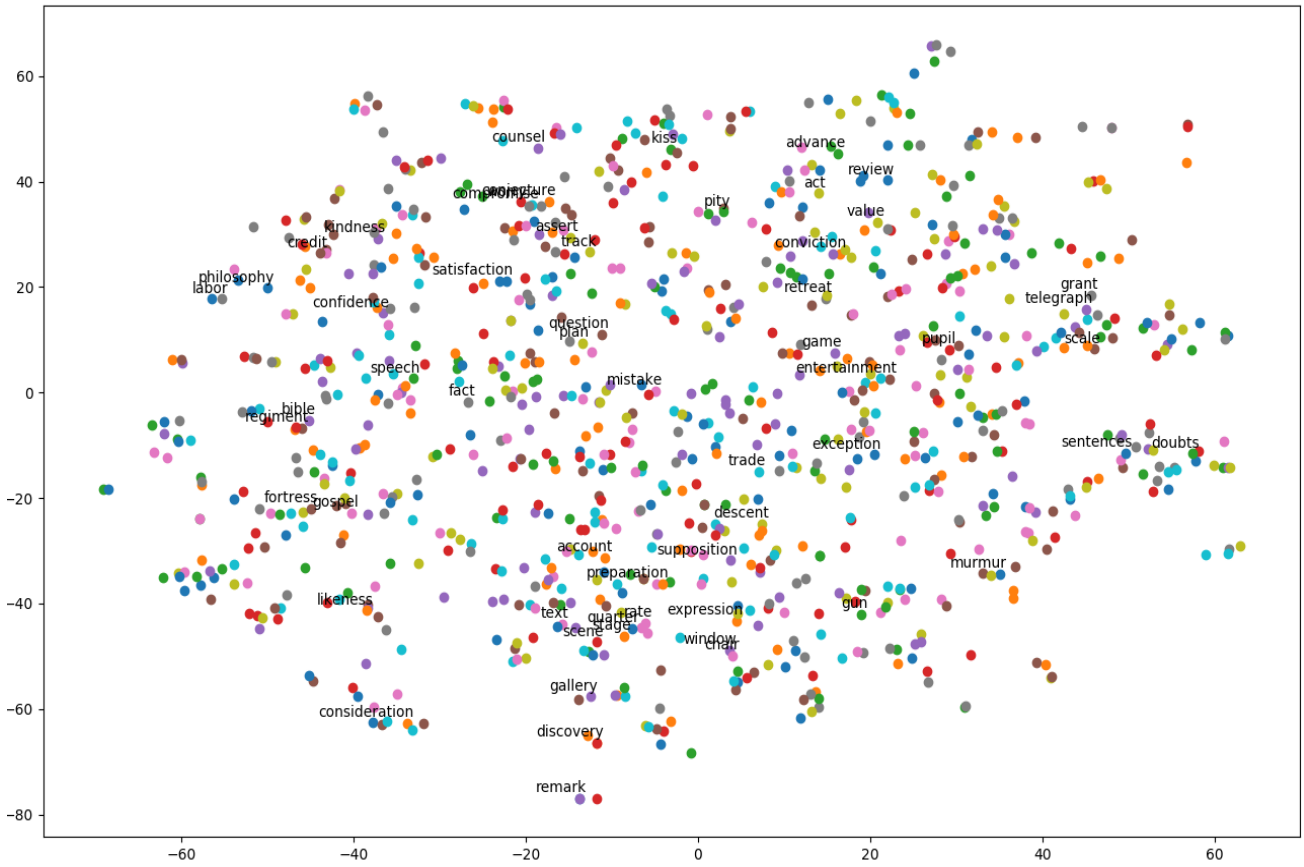


Figure 1: t-SNE visualization of 'remark'

while large context vectors that focus on content words will shift the interpretation towards a topical one.

We choose the parameter settings in such a way that the model captures both topical and syntactic word information. To achieve this, we use a word-order maintaining context window of three at either side of the target word, we include function words and bigram features, and we use as a weighting function a binary metric. The latter downgrades the importance of (highly frequent) function words, while it puts more weight on words with moderate frequencies. Also, it partially overcomes any data sparsity that might evolve from the inclusion of bigram context features. The chosen distance function is the Jaccard coefficient, which measures the number of shared features of each vector (the overlap).

After the identification of the Noun and the Adverb slot as having the most variation, we apply the described procedure to Ex. 2, which yields the results as seen in Table 2. The learner could infer from this that *rather* in combination with *obvious* specifies a degree, while *remark* can be interpreted as some sort of statement, both contributing and leading to a better distributional understanding of *obvious* and how it functions in this specific context.

### 3.3. Visualization of semantic clusters

As a further refinement to the above ranking, we also carried out experiments to visualize the word relationships at the hand of vclust (Suzuki and Shimodaira, 2006) and t-SNE (Van der Maaten and Hinton, 2008). We hypothesize that such visualizations will trigger a better understanding of which words within the phrasal slot are highly related to the relevant word, but also highlight which words have a set of distinctive features within the semantic space.

| Target Word | Paradigmatic Alternatives | Jaccard coefficient |
|---|---|---|
| rather | very | 0.429 |
| | pretty | 0.283 |
| | apparently | 0.234 |
| | equally | 0.220 |
| | fairly | 0.200 |
| | perfectly | 0.199 |
| remark | question | 0.256 |
| | reason | 0.248 |
| | fact | 0.229 |
| | point | 0.227 |
| | expression | 0.218 |
| | plan | 0.203 |
| | consideration | 0.190 |
| | conviction | 0.151 |
| | tendency | 0.126 |
| | begging | 0.112 |
| | monument | 0.104 |
| | supposition | 0.085 |
| | inference | 0.078 |

Table 2: Paradigmatic alternatives, ordered by semantic similarity to the target word (output Step 2)

Vclust is a hierarchical clustering algorithm based on multiscale bootstrap resampling. It overcomes the arbitrary input bias from a normal clustering algorithm by randomly

shuffling the provided input, effectively checking for consistency and robustness of the achieved clustering. T-SNE is a low-dimensional visualization of the achieved semantic vectors. Both visualizations are attractive as they offer an (implicit) unsupervised clustering which are poised to stimulate human interpretation.

Figure 1 shows the results of a 2D-representation of the semantic space for 'remark' using the t-SNE algorithm. The feature vectors of the 1000 nearest-neighbours (according to the Jaccard distance) have been selected to create a sufficiently large and meaningful space for the subsequent algorithm to work on. We reduce the high dimensionality of the selected vectors to 50 using Truncated Singular Value Decomposition (Halko, 2011). To account for usability, we try to avoid visual information overload by randomly labeling 90 words from the initial set of 1000 in addition to the 13 paradigmatic alternatives of 'remark' from Table 2. These pieces of information are given as input to the t-SNE algorithm; perplexity is set at 10 motivated by empirical evidence that our vector space is good at determining close neighbours, but that the effect wears off quickly. The number of iterations is set to 2000.

## 4. Expert Opinion and Validation Proposal

We asked 3 language teaching experts for their opinion on the usefulness of such techniques for possible applications in language learning. The experts pointed to applications that enhance receptive (comprehension) and productive vocabulary knowledge. The techniques could be integrated in for instance a reading aid, an activity mainly focused on receptive understanding. However, all experts expressed doubt whether language learners, especially high school students, would put in the extra effort to solve the linguistic puzzle laid out before them simply to understand a word, while easier alternatives, such as linked dictionary could be made available. They thus dismissed the usefulness of adding more structure from a didactic point of view for a mechanism (concordancing) which remains largely unused. One expert however thought of the technique as an excellent feedback mechanism to supplement the correction of errors as it would improve understanding why using a certain word is not correct in a specific context.

In a future validation scheme, we will therefore focus on the mechanism to serve as feedback for written exercises and tests in conjunction with automatic error analysis. Computations can in this case be performed offline and sent as a report. For language acquisition, it is of particular interest to investigate the influence of focused corpus-based information on the acquisition of certain lexical and syntactic patterns. As such the technique could prove a valid substitute, or at least a welcome addition for teacher feedback.

## 5. Summary and future work

We presented a set of NLP-functionalities that further structure examples retrieved for keyword-based searches. The examples are first grouped based on syntactic grounds, using POS-tags as an abstract representation of words from which constructions are derived. After this the examples are further organized at the hand of semantic similarity within certain phrasal slots. We use a self-implemented vector space model that emphasizes similarity both on a functional and a topical level. Visualization algorithms are then used to present focused information in phrasal slots, laying bare semantic relatedness of words, but also semantic restrictions within the construction.

We asked 3 experts for their opinion on the procedure's usefulness in language learning. While they were hesitant for its qualities to increase repective knowledge, they were more positive for it to be used as an automatic form of feedback.

While we intend to apply the procedure in language learning tools, we would like to point towards its utility for linguists, translators, lexicographers and other language professionals. The tools are generally intended to enable the study of language; the semantic interpretation of words and the proper understanding how to use words in context, both on a constructional, lexical and semantic level.

## 6. Bibliographical References


Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language* 82(4), 711-733.

Anatol, S. and Gries, S. Th. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8 (2), 209-43.

Chapelle, C. (1998). Multimedia call: lessons to be learned from research on instructed SLA. *Language Learning & Technology*, 2(1), 21–36.

Chinkina, M. and Meurers, D. (2016). Linguistically Aware Information Retrieval: Providing Input Enrichment for Second Language Learners. Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), 188-198.

Croft, William A. (2001) Radical Construction Grammar: Syntactic Theory in Typological Perspective. Oxford University Press.

Goldberg, Adele E. (2006). Constructions at Work. The Nature of Generalization. Oxford University Press.

Halko, N., Martinsson, P.-G. and Trop, J.A. (2011). Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *Society for Industrial and Applied Mathematics,* 53(2), 217-288.

Krashen, S. (1981). *Second Language Acquisition and Language Learning*. Oxford: Pergamon Press

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28 (2), 203–208.

Mikolov T., Sutskever I., Chen K, Corrado GS., Dean J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111-3119.

Nagy W., Herman P., and Anderson R. C. (1985). Learning words from context. *Reading Research Quarterly*, 20(2), 233–253.

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge, UK: Cambridge University Press.

Sternberg R. J. (1987). Most vocabulary is learned from context. *The nature of vocabulary acquisition*, 89–106.

Suzuki R. and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22(12), 1540-1542.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition.* Harvard University Press.

Turney, P.D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37(1), 141–188.

Van der Maaten, L.J.P. and Hinton, G.E. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research* 9:2579-2605.