University of Barcelona

MA in Cognitive Sciences and Language: Research Practicum

The Formulaicity of Language: A Computational Proposal to Detect Linguistic Constructions

Raquel Garrido Alhama

Advisors: M.Antònia Martí and Xavier Carreras

July 26, 2012

Contents

1. Introduction

Engineering is the discipline of applying scientific knowledge in order to build solutions. Therefore, Natural Language Engineering (NLE) should be the discipline that applies Natural Language knowledge to develop practical resources and applications that solve practical problems. However, the theoretical issues underlying NLE applications nowadays are not always clear.

Wintner [2009] points out that NLE used to be based in Linguistics. However, this paradigm became gradually abandoned. The main reasons were, on the one hand, the non-scalability of systems based on explicit knowledge; and in the other hand, the difficulty of the main trend in Linguistics -the Generative Grammar- to be related to other domains, such as computation or even psychology, due to its self-centeredness.

However, Wintner follows, there is still hope for NLE. Psycholinguistics and Cognitive Linguistics came up with new proposals that are general enough to be applied to other domains. These schools derive linguistic knowledge from language events, favouring statistics and probability theory rather than models based on *aprioristic* knowledge. These and other models should provide theoretical background to back up the NLE research, and at the same time, NLE should try to shed light on new linguistic phenomena we were unaware of.

This dissertation tries to make progress in this direction. In chapter 2 I review the theoretical issues that should be taken into account in a language model. I consider three dimensions of language: language as a cognitive faculty, language as human behaviour and language as a product itself. I survey some of the current findings and mental frameworks proposed in neuroscience to get an insight of how we represent language in our minds. After that, I examine the research carried out by developmental psychologists in relation to language acquisition. I finally relate the proposals of both disciplines with Construction Grammar and the Usage-Based models proposed by the Cognitive Linguistics school. These three axes reveal important properties of language and cognition in general that motivate my work in NLE.

One of the most important properties of human cognition and language that the disciplines above acknowledge is that we tend to group the units of information we process frequently in order to obtain faster access to them [Hawkins, 2005]. This implies that all the units of information that go together are stored as part of a larger unit or *chunk* that is accessed as a whole, with a progressive loss of conciousness of the smaller units that integrate it. These *chunks* of information are not necessarily concrete but may leave open some parametrizable abstract *slots* in order to adapt the *chunk* to different situations. Depending on how fixed or variable the *chunk* is, it will be more entrenched (i.e., with a greater tendency for the subject to lose conciousness of its integrating parts) or more productive (i.e., more parametrizable and hence useful on a greater range of situations) [Bybee, 2010]. These complex units have received many different names, some of them

referring to the general phenomenon of *chunking* and others to extremely prototypical or idiosyncratic cases (such as idioms): schemas, formulaic expressions, multiword expressions, multiword units, constructions, holophrases, collocates, chunks, amalgams, frozen phrases, etc. [Wray and Perkins, 2000].

The identification of linguistic schemas is crucial to enrich computational models of language with these cognitively plausible units. Chapter 3 surveys the state of the art on the identification of some of these schemas. Since the notion of schema is quite heterogeneous, some authors restrict their work to the identification of specific types of schemas; usually idioms, named entities, noun compounds, light verb constructions, and so on.

Chapter 4 presents my own approach to the identification of light verb constructions as a first experiment. A Vector Space Model is used to compute the semantic similarity between candidate constructions and single words: constructions that are similar enough to a single word have a cohesive meaning that does not actually need several words to be expressed, so the hypothesis is that they will be stored in our memories as a single unit with a single functional meaning.

I conclude this dissertation with reflections on the work done. I propose some lines of work to for my future research in order to derive some valuable considerations for my future work.

2. Cognitive and Linguistic background

This chapter reviews the theoretical issues that motivate my interest in linguistic schemas. It is divided into several sections that cover approaches to the study of language from different disciplines, focusing on those aspects that are convergent within the different perspectives.

Section 2.1 comprises the biological dimension of language as approached by neuroscientific studies. It reviews the Memory-Prediction Framework [Hawkins, 2005], a neuroscientific model that aims to unify different cognitive abilities and brain functions in a single coherent framework for human intelligence. A glimpse of the Mirror-Neuron System [Rizzolatti and Craighero, 2004] is also presented, since it plays a fundamental role in imitative learning, which is a crucial aspect for language acquisition, and in the evolution of language.

The study of language acquisition focuses on the individual evolution of the use of language. It is empirically based on the observation of human behaviour, and it can provide some intuitions about underlying psycholinguistic processes. Section 2.2 reviews the relevant issues related to language acquisition from the discipline of Developmental Psychology.

Finally, section 2.3 reviews the theoretical proposals from Cognitive Linguistics, a school that approaches the study of language in a bottom-up fashion that induces knowledge from the linguistic product. Several linguistic frameworks have been proposed, usually closely related to each other. All of them share the basic assumptions that make them part of the same school: language is seen as a dynamic system that underlies general mental processes, and it is studied in a holistic manner that does not tear apart grammar, semantics and pragmatics.

2.1. A neuroscientific perspective

Crick [1979] claims that in spite of the great accumulation of knowledge we have about the brain, how this organ really works still remains a mystery. According to Crick, neuroscience needs a broad framework to interpret conjointly what we know about different brain areas.

Inspired by this idea, Hawkins [Hawkins, 2005] present a comprehensive framework that explains human intelligence as a whole. The model is known as Memory-Prediction Framework, and its central idea is that we memorize patterns of the regularities we detect in the world and use them to make predictions about what we are going to experience next.

Hawkins' notion of the human brain is supported by Karl Friston's mathematical brain theory of the Free Energy Principle [Friston, 2010]. Friston claims that several of the current theories about how the brain works can be unified under his proposal, which

states that the brain tries to minimize surprise while maximizing utility. The underlying idea is that the brain creates hypotheses related to the patterns it detects in the world and makes predictions taking into account the tension between surprise and utility. Although Hawkins does not provide a mathematical apparatus to make predictions, his model is consistent with the models unified under the Free Energy Principle framework, which include, among others, the Bayesian Brain Hypothesis.

Section 2.1.1 examines the relevant aspects of Memory-Prediction Framework. Section 2.1.2 briefly introduces the Mirror-Neuron System to provide a neurophysiological basis of imitative learning and the evolution of language.

2.1.1. The Memory-Prediction Framework

The study of the brain is traditionally divided into the study of different functional areas that have been identified in the cortex. This is supported by the fact that some mental illnesses are localized in specific regions, and functional areas seem to be specialized in certain aspects of perception or thought. However, Mountcastle [1978] observes that the anatomy of the neocortex is in fact very regular. He suggests that, although the cortex is divided into task-specific functional areas, the operation they are performing is the same.

Our daily world has structure: objects can move as a whole, retaining their size and shape, and they have surfaces, color, smell, etc. The brain deals with sensory information coming from different senses, such as visual, auditory or tactile information, and it also controls motor actions. In spite of that, the brain is executing the same algorithm regardless of the type of information it is processing. Crick [1979] claims that whatever processing is done on the information coming to the brain will be related to the invariants or semi-invariants of the external world that is accessible to the senses.

The Memory-Prediction framework is based on Mountcastle's principle. According to Hawkins, hearing, seeing or moving are the same in an abstract level of sensory inputs. The brain only deals with spatio-temporal patterns of neural signs -action potentials or spikes- that are identical regardless of what caused them. Consequently, spoken language and written language are perceived in a similar way, despite being completely different at the sensory level.

We know so far that the neocortex treats all the information in the brain in an homogeneous way, and that it is represented as sequences of spatio-temporal patterns that store the invariance detected in the world. Our memory is based on this representations, and it has some important properties that will enable us to predict our next experiences.

Memory

The memory-Prediction framework states that human intelligence relies on memory to solve problems. The entire cortex is a memory system from where solutions to problems are retrieved. The difference between computing a solution and retrieving it from memory relies on the use of rules or exemplars, respectively.

If we tried to catch a ball computing each muscle movement on the fly we would be terribly slow. But retrieving from memory the muscle commands that are necessary to catch a ball works faster: the appropriate memory is recalled; then it is subdivided into a temporal sequence of muscle commands and finally it is adjusted to the actual parameters -the path of the ball and the position of your body. After years of practice, our brain has grouped and stored each muscle command necessary to catch a ball, hence it does

not need to calculate them from scratch. That is why we train ourselves through the repetition of exercises: we want our memory to group and store together the sequence of commands we use to complete a task.

As Crick anticipated, our memory represents information as flexible (semi-)invariant sequences of patterns that are suitable to handle variations automatically. In order to solve any problem we rely on solutions we have already stored to solve previous problems, and we adapt them to the actual situation. When the same problem has been solved repeatedly, we routinize the solution: we package it and store it in an invariant form that is adapted to new situations when needed.

Our memory has some important properties that make this processes possible. To begin with, the neocortex stores sequences of patterns. We have a vast capacity of memory, so we do not consciously access all our memories at a time but sequentially. This explains why we find it difficult to recall some memories in opposite order, such as the alphabet. The stream of memories we are consciously aware of follow a pathway of association.

Secondly, human memory is auto-associative. We can recall complete patterns from partial or distorted inputs. That is how we recognize familiar objects even though we perceive them under different conditions. As an example, we are not confused when we see our friends' faces under different light conditions, make up, glasses or even if they are partly hidden: the input is different all the time, but we manage to recognize the invariant. In fact, it is well known that we do not actually hear all the words we perceive; nevertheless, our mind completes the missing parts. The cortex is constantly completing the input patterns using stored memories. In other words, a piece can activate the whole.

The third important feature of our memory system is that it stores patterns in an invariant form instead of collecting the exact perceived information. The auto-associative memory system needs of an invariant form that gets rid of the details in order to recognize entities. In the neurophysiological level an invariant representation is the stability of a group of cells firing. Whenever some known entity enters our field of vision, some particular cells remain active.

Finally, the neocortex stores the invariant patterns in a hierarchical structure that is isomorphic with the world's nested structure. The regions of the brain that are lower in the hierarchy deal with detailed information that comes from sensors, while higher regions treat the information in a more abstract way. Higher regions are also responsible for the association of data that comes from different sensory inputs.

Our memories of objects are distributed over nested hierarchies. For instance, our memory of a tea cup is not stored in a single spot of our brain, but distributed over the cortical hierarchy. Its parts and details are stored in the lower regions while a more abstract pattern of its silhouette will be stored in a higher region, facilitating the recognition of another tea cup that differs in details such as size or colour. At the same time, the whole memory of the cup is nested in the memory of the kitchen along with all the memories of all the objects present there, and the memory of the kitchen is nested inside the memory of our house, and so on.

Objects that are stored lower can be reused to be part of different higher objects. This property is nicely reflected in language: words can be reused in different sentences, so that we do not need to memorize them repeatedly for each sentence. Phonemes can be reused to be part of different syllables, which can be part of different words, which can be part of different sentences, and so on.

Whenever a new sequence of patterns enters the neocortex, the information flows

from lower to higher regions. Each region tries to recognize the input and assign it a 'name'. If the region is not successful, it passes the information to the next region in the hierarchy, which at the same time attempts to recognize it. If the input arrives at the top of the hierarchy without being recognized, it is stored as a new memory. The flow of information goes in the opposite direction when we perform some action. Higher regions recover a pattern of the action to perform and unfold it into lower regions that deal with the details.

As we will see next, these properties of our memory system are crucial to make predictions about the world we experience.

Prediction

The Memory-Prediction framework states that the seat of intelligence is prediction. The input we are receiving combined with our stored memories allow us to predict what we are going to experience next. Since our memories are sequences of patterns, once we recognize the sequence we are perceiving we predict the next pattern. When we are unsure, we assign tacit probabilities to more than one prediction at a time. For instance, when someone starts the sentence "Pass me the..." while having dinner, one would usually predict that the next word will refer to one of the objects on the table, as we have previously experienced the situation of asking for unreachable objects during a meal. Hawkins does not provide a mathematical model of probability as Friston [2010], but implicitly suggests that we make predictions minimizing our surprise and maximizing what is consistent with our experience.

According to Mountcastle, all cortical areas perform the same operations. So the brain makes predictions ubiquitously, in all levels of representations in the hierarchy - from the lowest more-detailed levels to the highest more-abstract ones. From the neurophysiological point of view, predicting an experience means that the neurons that would be involved in that experience become active before they actually receive sensory input. Once the sensory input arrives, it is compared with the expected input. Correct predictions result in understanding, while the incorrect ones result in surprise.

When the flow of information that comes from our sensors goes up the cortical hierarchy, each region tries to recognize it; if it fails, it passes the message up in the hierarchy. Once the input has been recognized, we have retrieved its correspondent sequence of patterns from memory, so we have a prediction of what will be next. If we have several sequences in memory that correspond to the input, we will have several predictions with a tacit probability. At this point, information about what we are going to experience next flows down the hierarchy.

Behaviour works under the same principles. We predict what we are going to experience because we are aware of the actions we are going to do. In that case our prediction is also the cause of our next sensation. As the prediction unfolds down the hierarchy, it generates the motor commands to fulfill the prediction. Predicting and acting become part of the same process.

Our perception usually comes from different sensors at the same time: we see, hear and maybe touch what is around us. All this information flows up in the cortical hierarchy until an association area makes a prediction that integrates the information coming from all senses. Multisensory information is flowing up and down the cortex to create a unified sensory experience.

To sum up, our cortex stores sequences of patterns. When a new sequence of pattern

enters the cortex, the auto-associative memory tries to retrieve a similar sequence already stored to make predictions of future events. Memories need to be stored in an invariant form so that the knowledge of past events can be applied to new situations that slightly vary with the ones already experienced.

Learning

Since the moment when we are born, we start absorbing and creating a model of the structure we perceive in the world. That is a learning process, and it is ubiquitous in all regions of the cortex. Hebbian Learning establishes that when two neurons fire at the same time, the synapses between them get strengthened. Learning is based in classifying and grouping spatio-temporal patterns into sequences.

Patterns are stored together whenever they are part of the same object. One way to do this is to group the patterns that occur contiguously in time. For instance, when an object moves, the input patterns change but there is an invariant part since the object keeps its properties. Other sequences of patterns need of a guide, such as distinguishing whether an apple belongs to the same set as a banana or as celery.

As a region of cortex builds sequences, the input to the next region changes from representing mostly individual patterns to representing groups of patterns: from apple to fruits, or from phonemes to words. Thanks to that, the higher region can focus on learning sequences of higher-order objects. For instance, it can take words from the lower region to build sentences.

During repetitive learning, representations of objects also move down the cortical hierarchy. In our early years of life, our memories will form in higher regions of the cortex, but as we learn they move down to lower and lower parts of the hierarchy. The brain does not literally 'move' the sequences but reforms them in the lower regions. In that way, higher regions are free to learn more complex and subtle patterns. For instance, a child needs great effort to recognize individual letters, but after repetition, he gets proficient in the task an focuses on recognizing the higher-order objects (words). Once he is proficient in that task, he gets to the point where he instantly recognizes entire words at a glance. The recognition of letters is occurring fairly low in the cortical hierarchy, closer to the sensory input. As the memory of objects move down the hierarchy, higher regions have the ability to learn higher-order objects.

Language in the Memory-Prediction Framework

According to the Memory-Prediction framework, language does not need any dedicated machinery or any special treatment in the brain. Linguistic constructions, whether written or spoken, are represented as sequences of patterns. Syntax and semantics reflect the same hierarchical structure of other objects in the world.

Every level in the cortex tries to 'name' sequences of patterns when recognized. If successful, the name is passed on to the next region above in the hierarchy. Each higher region experiences a less variable input, since it is more abstract and less detailed. Thanks to that, a pattern stored in a higher region can have different representations at the lower ones. As an example, a concept stored at one level can be linked to both phonological and orthographic patterns at lower levels. The association between form (whether phonogical or ortographical) and meaning works exactly the same as the multisensory association.

Thanks to this correspondence, language can invoke memories in other humans and create different juxtapositions of objects in the listener's mind that lead him to new asso-

ciations. Language has also the property of causing other humans to learn about things they may never experience, working like pure analogy. The language faculty requires of a large neocortex capable of handling the nested structure of syntax and semantics, as well as a fully developed motor cortex and musculature to enable the sophisticated articulation of phonemes.

Nygren and Wu [2011] observes that, according to Hawkins, language is learnt from scratch and with no difference to other skills. The human brain is sensitive to the frequencies of the events it experiences, and creates form-meaning associations by abstracting regularities from the input. As we will see in section 2.3, these are the main assumptions of Usage-Based models of Language.

Crucial to this thesis is the role of repetition in the general learning process, and hence also in language. Getting proficient in some skill means that objects are moved down in the cortical hierarchy, so that higher regions can work with higher-order objects and focus in more complex and subtle relations. The point of this thesis is the observation that, since this process is ubiquitous, it should not only work for letters and words, but also for larger linguistic constructions.

We are constantly receiving linguistic input. If we observe that some words go frequently together, the cortex may store them lower in the hierarchy as a higher order object. Since the cortex uses invariant representations, this complex unit may still be parametrized when used; that is to say, the linguistic construction may still permit morphological inflection and the insertion of other elements within. This general notion of complex parametrizable units corresponds to the notion I introduced as linguistic schemas (see chapter 1), and since all information in the brain is processed the same, linguistic schemas can be found in every form - from phoneme constructions to complete recurrent sentences.

2.1.2. The Mirror-Neuron System

The discovery of mirror neurons in the cortex of primates has been considered one of the most exciting recent breakthroughs in neuroscience. These neurons discharge both when the primate performs a particular action and when he observes an individual doing a similar action [Rizzolatti and Craighero, 2004]. They account for action understanding and imitation learning, and they are of great importance for the evolution of language.

The understanding of actions performed by others is solved by a rather simple mechanism in the Mirror-Neuron System. Each time an individual sees an action done by another individual, the neurons that would activate if the former individual did that action himself are activated in his premotor cortex. This automatically induced motor representation of the observed action corresponds to that which is spontaneously generated during active action and whose outcome is known to the acting individual. Thus, the mirror system transforms visual information into knowledge [Rizzolatti et al., 2001].

Understanding others' actions allows primates to learn by imitation. There are, broadly speaking, two types of behaviours that are acquired by imitation: substitution of a motor pattern for the observed one, if it is more adequate to fulfill a given task, and the capacity to learn a new motor sequence for a specific goal. Even if the observer is shown the same action under very different visual stimuli he is able to generalize and activate the same motor neurons, proving that the Mirror-Neuron System has great degree of generalization[Rizzolatti and Craighero, 2004].

The process of learning by action understanding can be seen as a communication

system, where the individual performing an action is sending a message -the action itself- to the observer, who understands the action and acquires new knowledge without any cognitive mediation. Mirror neurons are the neural basis that create a direct link between the sender of the message and the receiver. On the basis of this property, Rizzolatti and Arbib [1998] propose that the Mirror Neuron System provides the neurophysiological mechanism that enables a common semantic between individuals, making possible the evolution of language from gestural to oral communication.

In gestural communication, semantics is inherent to the gestures used to communicate. In speech, however, the phono-articulatory actions necessary to pronounce words are not necessarily related to their meaning. At some point in evolution there must have been a step of transferring gestural intrinsic meaning to symbolic sound meaning.

A number of studies prove that hand/arm gestures share a common neural substrate with speech gestures [Rizzolatti and Craighero, 2004]. The transfer from hand/arm to oro-laryngeal gestures is thought to occur thanks to the activation of audio-visual mirror neurons related to ingestive behaviour. A further step is also needed to generate the sounds originally accompanied by a specific action without doing the action, probably thanks to improved imitation capacities. Neurons developed able to generate the sound and discharge in response to that sound (echo-neurons). Therefore, when an individual listens to verbal stimuli, there is an activation of the speech-related motor centers.

Gallese and Lakoff [2005] observe that the sensory-motor system also characterizes abstract concepts that constitute the meanings of grammatical constructions. As in the Memory-Prediction Framework, they claim that language makes use of the same brain structures used in perception and action, without a dedicated 'language module', and grammar resides in the neural connections between concepts and their expression via phonology. Hierarchical grammatical structure is isomorphic to conceptual structure, while linear grammatical structure is just phonological. Neither semantics nor grammar consist of rules for manipulating meaningless symbols.

2.2. Developmental Psychology: Imitative Learning and Language Acquisition

Most researchers agree that human beings are endowed with an innate capacity that allows them to learn one or more languages. The nature of this capacity, however, is a matter of controversy. The different theoretical proposals about this endowment can be situated somewhere between two extreme positions: those that consider language acquisition as genetically determined by innate representations and specific mechanisms and those that consider linguistic knowledge as the result of general learning processes.

Proposals that fall into the first extreme usually assume that we humans are not capable of inducing all aspects of language just with our exposure to the linguistic environment. This is known as the Poverty of the Stimulus argument [Chomsky, 1965]. Researchers that hold this position claim that we are born with innate linguistic principles and a set of parameters that need to be set during our exposure to language.

However, empirical research challenges the innatist position. Clark and Lappin [2011] develop formal learning models to demonstrate that it is possible to explain linguistic acquisition through general methods of induction that extract patterns and structure from data. From the field of developmental psychology, Tomasello [2001, 2003, 2008] proposes a theory of language acquisition that is based in imitative learning and general

domain processes, such as analogy and abstraction. This proposal is consistent with the idea of a single cortical operation in the Memory-Prediction Framework, as well as with the role of mirror neurons in action understanding and imitative learning (see section 2.1). This section reviews the main principles of Tomasello's theory of language acquisition.

Tomasello is interested both in the phylogenesis and the ontogenesis of language. He tries to identify the cognitive and cultural aspects that distinguish humans from their nearest primate relatives -the great apes- in order to discriminate which characteristics are relevant for language acquisition. He realizes that one of the crucial differences between humans and apes is the understanding of communicative intentions, which is non-existent for monkeys: there are no observations of primates directing the attention of another ape towards a third entity. Prelinguistic infants, however, do realize after their first year of life that people who produce sounds directed to them are in fact trying to manipulate their attention towards something else. At that age, children begin to understand actions as intentional.

Adults express their communicative intentions with utterances. Therefore, Tomasello claims utterances must be the basic psycholinguistic unit, understood as *a linguistic act in which one person expresses towards another, within a single intonation contour, a relatively coherent communicative intention in a communicative context* [Tomasello, 2001]. This definition is broad enough to cover linguistic expressions of different forms and complexity.

Tomasello hypothesizes that children attempt to understand and reproduce the entire utterances they hear from adults in a process of imitative learning. At early stages they are only successful in reproducing some of its linguistic elements, such as 'That!' meaning 'I want that'. These productions are known as *holophrases*, since they are linguistic symbols that function as whole utterances.

The holophrases come in many different forms. Earlier holophrases are just frozen phrases. At some point during the learning process they are segmented into their elements, such as 'Lemme-see' or 'Gimme-that'. This is a process of moving from the whole to the parts, thanks to the progressive understanding of the functional role of each element.

The early productions of most children exhibit an asymmetry between constituents. A single element seems to be more prominent, and the whole utterance revolves around it. Some examples are 'Where is the X' or 'I'm X-ing it', where 'X' stands for a slot to be filled with other linguistic elements. These are the first examples of linguistic schemas showing some degree of abstraction; the child has identified some linguistic structure in the utterances he previously perceived as lineal, and hence the beginning of grammar.

So grammar emerges from specific linguistic examples in a process that goes from specific lexical pieces to abstract categories to be filled. These categories need not be similar to adults' morphosyntactic categories, but may have some semantico-pragmatic form, such as 'I want THING-WANTED'. Tomasello hypothesizes that the first syntactic categories created in verb-islands (i.e., utterances structured around a verb) are lexically-based roles such as 'hitter', 'thing hit' and 'thing hit with' for the 'hit' verb-island. Anyhow, the linguistic behaviour of children can be characterized as a progressive induction of linguistic schemas from lexical exemplars where slots are created as type variation is observed. This involves a process of observation, detection of statistical regularities and variability, and the final abstraction of higher-order categories.

Linguistic schemas with slots allow for the creativity of the speaker. They function as tools that let the child fulfil his communicative intentions. The more number of slots

a linguistic schema has, the more adaptable it becomes and therefore, the more useful to achieve different communicative intentions. Linguistic schemas function as local syntactic rules that combine form and meaning.

As in the Memory-Prediction Framework, linguistic schemas can be regarded as solutions to the problem of communicating particular intentions, and the same linguistic schema can be adapted to similar but slightly different situations. When some linguistic elements of a utterance come frequently together, the schema becomes more entrenched and it is more difficult for the child to adapt it to new situations, although he becomes more fluent with the production of entrenched items.

Tomasello also contemplates the domain-general mental process of analogy. For instance, verb-islands such as those around 'give', 'tell' and 'show' share a meaning of transference, and they all appear in the structure NP+V+NP+NP. Children may be creating relations by mapping across constructions that share meaning, transferring the structural knowledge they already have to the newly acquired.

To sum up, young children imitate our linguistic behaviour, reproducing utterances to express communicative intentions. In their attempt to understand adults, they come to discern patterns of language usage (including tacit knowledge of token and type frequency) and these cues lead them to memorize a redundant lexicon composed of linguistic schemas that vary in complexity and abstractness. The linguistic memory is an heterogeneous inventory of constructions that go from single words to item-islands or partially abstract utterances stored together with their communicative intention. The creativity of language comes from retrieving linguistic schemas and adapting them to new situations.

This theory of language fits very well with proposals coming from Cognitive Linguistics. Usage-based models (see section 2.3.3) also consider grammar as emergent from linguistic events, thanks to general domain processes of detecting patterns in the stream of sound we perceive in speech. In construction-based theories (see section 2.3.2), linguistic constructions are the basic unit of grammar. A construction is conceived as a pair of form and meaning that may vary in complexity and abstractness, so this notion is compatible with the unit of utterance considered by Tomasello.

2.3. Cognitive Linguistics

Cognitive Linguistics is a branch of linguistics that was consolidated around the 1980's. Two publications are specially important to establish the bases of this school: *Foundations of Cognitive Grammar: Theoretical Prerequisites* [Langacker, 1987], where the author presents a semanticocentric theory of grammar, and *Women, Fire and Dangerous Things. What Categories Reveal about the Mind.* [Lakoff, 1987], where the philosophical and psychological foundations of cognitivism are explained.

The branch subsumes a great variety of frameworks: typological studies (Talmy [1985], Slobin [2004]), prototype theories (Prototype Model [Rosch, 1975], Radial Model [Lakoff, 1987], Schematic Network Model [Langacker, 1987]), embodied cognition and conceptual metaphor [Lakoff and Johnson, 1980], construction-based grammars (Cognitive Grammar [Langacker, 1987], construction grammar (Lakoff [1987], Goldberg [1995]), Construction Grammar [Kay and Fillmore, 1999], Radical Construction Grammar[Croft, 2001]) and usage-based models [Bybee and Hopper, 2001, Bybee, 2010], among others. This section will focus on the general principles that are shared by most of the frameworks, but I will put special emphasis in construction-based grammars and usage-based models, since they

can be related to the proposals coming from neuroscience and developmental psychology previously reviewed.

2.3.1. The commonalities

Cognitive Linguistics appeared as a reaction towards some of the basic principles of the main trend in linguistics at that time, Generative Grammar. Contrary to that trend, this new school of linguistics does not assume that we are born with an innate linguistic module that determines our linguistic knowledge. The specificity of language in our minds is challenged with the proposal that linguistic knowledge should not be conceived separately from general cognition. Our general-domain mental processes determine language acquisition and usage in the same way they operate with any other cognitive ability, such as general reasoning.

So this approach considers language as a capacity whose acquisition process follow the same parameters as other cognitive skills. This broad perspective takes into account the communicative act where the linguistic experience takes place, as well as the mental processes underlying language perception and production. Therefore, it is open to the integration of principles derived from other disciplines related to human cognition and communicative behaviour, such as psychology, cognitive neuroscience or anthropology.

From the principle of the domain-generality of language it follows that linguistic categories follow the same rules as our general conceptualization of the world. Because of this reason, grammar models proposed in this branch are semanticocentric, in the sense that syntax arises as the result of structuring conceptual knowledge. This view is also contradictory to Generative Grammar: syntax is not an autonomous module, so it cannot be formalized with abstract categories that are isolated from semantics. Grammar is considered a continuum that ranges from isolated concepts in the lexicon to complex grammatical assemblies.

This holistic view of grammar is also manifested in the units of language. All linguistic elements are considered a pair of form and meaning; morphemes, for instance, are meaningful symbols that only differ with words in complexity. But the lexical pieces which are closer to grammar in this continuum include structural information. These lexico-grammatical assemblies are the basis of several grammar proposals that are usually referred to as Construction-based Grammars. Section 2.3.2 reviews the main issues of these theoretical proposals.

Finally, it is important to remark that Cognitive Linguistics approaches the study of language in a bottom-up fashion that completely deviates from the Generative Grammar tradition. Rather than proposing an aprioristic top-down theory, cognitivists collect empirical data of communicative acts from where induce the theoretical models. In this way, linguistic competence and performance are not analysed as separate things; moreover, usage-based models, which are reviewed in section 2.3.3, emphasize that linguistic knowledge is being constantly shaped by language usage.

2.3.2. Construction-based Grammars

Construction-based theories describe language from a synchronic perspective. They consist on grammars that rely on the holistic paradigm described above, which states that lexicon and syntax are not separated by clear-cut boundaries but part of a continuum. The basic unit of this lexico-grammatical continuum is the *construction*.

Constructions are commonly defined as grammatical assemblies that combine a specific form with a specific function or meaning, exhibiting both general grammatical properties and idiosyncratic features. This notion is general enough to apply to all grammatical assemblies: the formal side comprises phonological, morphological and syntactic features, while the functional side subsumes semantic, pragmatic and discourse-pragmatic features.[Diessel, 2004]. Therefore, the notion of construction is ubiquitous to all linguistic elements, ranging from morphemes to abstract constructions such as the passive.

One of the main characteristics of constructions is the variation in syntagmatic complexity. Morphemes are constructions that are combined forming words, which are larger constructions that at the same time are combined in constituents and finally in sentences, both of which are constructions as well. Hence constructions can be constituted of different number of elements, and they are nested in a hierarchical structure.

Since the form of constructions can comprise syntactic features, constructions also vary in abstractness. This implies that the form of a construction may be either lexically specified or provided by morphosyntactic features. For instance, *She sneezed the napkin off the table* is a construction, but *NP sneezed NP off NP* is a construction as well. In the lexico-grammatical continuum, fully lexicalized constructions are in one extreme and completely abstract constructions such as *[SUBJ [V OBJ OBL]]* are in the other. It is important to notice that fully abstract constructions are not equivalent to formal rules as found in Generative Grammars, since constructions are not only correlated with form but also with meaning. Hence, *[SUBJ [V OBJ OBL]* is stored in our memories along with its semantic function of expressing caused motion.

Contrary to what had been traditionally thought, the verb does not project an argument structure. It is provided by the construction, as well as the general meaning. Once the verb fills the construction, it refines the meaning with its own contribution. For instance, the caused-motion construction can be instantiated such as in *She forced the ball into the jar* or in *He wiped the mud off his shoes*; both express caused motion but in different manners. Resultative constructions such as *She sneezed the napkin off the table* provide evidence that the meaning cannot be entirely derived from the verb, since *sneeze* is not usually considered a motion verb [Goldberg, 1995].

The variation in abstractness also subscribes a hierarchical arrangement of the pieces of language. It is clear that specific constructions are lower in the hierarchy and inherit properties from the abstract constructions above them. However, the amount of information that lower-level constructions store has provoked some divergences among cognitive linguistics: while Goldberg [1995] hypothesizes that all constructions are stored the same way in a redundant fashion, Fillmore [1976] proposes a model where specific constructions are not stored with the properties that are common to its abstract counterparts but directly inherit them. While the latter proposal is more economic in terms of memory, the former is optimal in terms of computation time. As Hawkins proposes in the Memory Prediction Framework, our brain is not powerful in computational speed; instead of that, it has a vast memory from where solutions to problems are retrieved.

Abstract constructions, usually referred to as constructional schemas, allow for the creativity of language. The productivity of a constructional schema is determined both by similarity with other schemas and by the number of expressions that are related to a particular schema: the more types of expressions linked to a constructional schema, the more it is used. However, as a product of habituation, speakers tend to draw on prefabricated specific formulas whenever they are available and suitable for the social situation [Bybee, 2010]. Thanks to a grammar that combines both abstract and specific pieces

of all sorts, speakers manage the tension between being creative and minimizing computation effort. That is one of the reasons why spontaneous speech often abounds with formulaic expressions and semi-productive phrases that are organized around concrete lexical expressions [Diessel, 2004].

Construction-based grammars are relevant to the topic of this thesis because they are consistent with the proposals coming from neuroscience and developmental psychology, and they provide the linguistic model in which the experimental method presented in chapter 4 relies. The information is organized in a hierarchical fashion that mirrors our conceptual representation of the world, in the same way proposed by Hawkins. The variation in specificity and abstractness reminds of the organization of our memory proposed in the Memory-Prediction Framework: exemplars are stored lower in the hierarchy, while abstract categories that are less specified are above, related to all the exemplars that can instantiate them. Our learning paradigm is based on the abstraction of regularities from exemplars, but also in grouping elements into higher order units. The syntagmatic variation of constructions correlates with this capacity, but as we will see next, the usage-based models explain how frequency creates differences in the strength of syntagmatic relations.

As advanced in section 2.2, Tomasello adopts Goldberg's Construction Grammar since it is consistent with his observations on the acquisition of language: children acquire lexical constructions and learn to find structure on them, creating this more abstract and productive assemblies that compose grammar. Holophrases are nothing more than early meaningful constructions in which the child does not yet differentiate the parts from the whole.

2.3.3. Usage-based models

Usage-based models approach the study of language with the aim of explaining both its variation and its regularities. Their central idea is that each linguistic experience has an immediate effect on the linguistic representations in our memory, to the extent that linguistic structure arises from language usage as the result of multiple applications of domain-general mental processes along time. These are the processes behind the construction-based structure of language seen in section 2.3.2.

From this perspective, grammar is seen as a dynamic adaptative system that is constantly changing depending on the subject's exposure to language (Bybee and Hopper [2001], Bybee [2010]). Mental representations are unstable states of affairs that are sensitive to usage, constantly adapting themselves. Consequently, emergent structures are manifested stochastically.

The constant adaptation of language can be nicely illustrated with the phenomenon of grammaticalization. Linguistic expressions are usually divided into two general types: symbolic expressions (nouns, verbs and adjectives) and grammatical markers (prepositions, conjunctions, etc.). The former have a denotative function, while the latter are structural expressions. Grammaticalization theory posits that all grammatical markers are in fact derived from symbolic expressions. The reason is the role of frequency of use: linguistic expressions that are frequently used tend to be reduced in structure and meaning, eventually leading to the development of new grammatical markers. To put an example, the conjunction *because* evolved from the adpositional phrase *by cause*.

The dynamism of language is crucial to this thesis specially because of the effects of frequency in linguistic representations. Usage-based proponents agree that frequent

constructions are more deeply entrenched in our memory than rare constructions. The representation of constructions in our memory is reinforced with each use, so at the same time, the expression is more readily available to next uses. Thus, the use of linguistic expressions has an immediate effect on the representation and activation of linguistic knowledge [Diessel, 2004].

Frequency has another effect related to our general-domain process of routinization. Lexically specific constructions become highly entrenched in our mental grammar if they occur with high token frequency, that is, if they are experienced a large amount of times. In that case, the linguistic elements composing the construction may develop a syntagmatic relation that reinforces their use together, so they start to function as a single piece of language that is directly accessed. Bybee [2010] refers to this expressions as *chunks*, and she provides examples such as *I don't know X* or *why don't you X*. This process, which is ubiquitous at all levels of grammar (from phonetics to syntax) contributes to fluency and ease both in production and perception: the longer the string that can be accessed together, the more fluent the production and comprehension.

Chunking is the mechanism behind the formation of prefabricated expressions. It is the process responsible of the formulaicity we perceive in language, but also of the formation of constructions and the hierarchical structure of language. All sorts of prefabricated expressions -multiword expressions, idioms, or even constructional schemas- can be analysed as chunks [Bybee, 2010].

It is worth pointing out that chunks need not be continuous: they can be interrupted by open classes of items. For instance, the *drives X mad* construction can have either noun phrases or pronouns in the X slot, although some choices are more frequent than others. In fact, this is equivalent to saying that chunks may include abstract categories that are not lexically specified.

The status of a chunk in memory falls along a continuum: from words that have been experienced together only once, which will constitute a weak chunk whose internal parts are stronger than the whole, to recurrent chunks which are accessed as a whole while still maintaining connections to their parts (e.g. *lend a hand*). Chunks resulting from grammaticalization would be in the extreme high-frequency end of the continuum, since they are markers that have lost their internal structure and the identifiability of their constituent parts.

Finally, let us compare the process of chunking and the effects of linguistic experience in general with the learning paradigm presented in the the Memory Prediciton Framework. Hawkins emphasizes that we learn through repetition of experiences: we train ourselves repeating exercises until we routinize all the steps of the exercise in a manner such that they are stored as a whole routine in our memory. Next time we do the exercise, we will access to the whole representation directly, adapting it to the particular situation. The effects of frequency in language experience works just the same: the more we perceive that some linguistic elements go together, the more likely they will be stored in our memory as a chunk that will be later accessed as a whole, although it will be parametrized for the particular communicative act. This permits accessing to stored solutions to achieve communicative purposes in the same way we memorize solutions to other problems. The computational cost of recovering linguistic pieces is reduced if we store larger recurrent pieces, so the formulaicity of language improves fluency in the same way that we get better at other cognitive and neuromotor tasks with practice.

Usage-based proposals are also consistent with the observations in developmental psychology. As Tomasello affirms, children are constantly adapting their linguistic knowl-

edge depending on their linguistic environment. The frequency of linguistic experiences determines the entrenchment or the productivity of the constructions that the child is acquiring; that is why, for instance, they do not break frozen phrases until they observe type frequency. The structure of language progressively becomes apparent to children when they collect enough linguistic knowledge to perceive regularities and variation. At that stage, they create abstract representations in the same way that structure emerges in usage-based models.

2.4. Summarization

The theoretical background previously reviewed presents an integrated view of language that involves different perspectives. Each discipline focuses in different aspects of language, but all of them coincide when defining the basic units we store and retrieve from our linguistic memory.

From the point of view of neuroscience, we represent the world in sequences of patterns that at some point we may group together in a higher-order sequence, storing a complex object that behaves as a unit. In developmental psychology, the unit of language attested in children's speech is the utterance, an early production that gradually becomes more schematic. In the Cognitive Linguistics field, Construction-based grammars propose that the units of language are constructions that have different levels of complexity and abstractness, and Usage-based models analyse the process of chunking in order to explain how speakers create assemblies of linguistic elements as the result of frequency effects and domain-general processes. All the proposals emphasize that these units include information both of its symbolic form and its semantic-pragmatic function.

So all these theoretical proposals have something in common: they define the units we store and retrieve from memory as complex assemblies consisting on smaller elements that at some point are grouped together in order to create a higher-order object composed of form and function. Whether referred to as sequences of patterns, utterances, constructions or chunks, the underlying idea is the same.

This convergence attests for a reliable and interdisciplinar framework that explains how do we perceive and produce language, specially concerning the identification of its basic units. The next chapter surveys how researchers have tried to detect these units automatically to benefit applications in NLE.

3. Computational Approaches: State of the Art

This chapter presents a state of the art of computational approaches that try to identify linguistic constructions of several types. Most of the approaches focus on linguistic constructions that are highly conventionalized or that have developed idiosyncratic properties: fixed expressions (*ad hoc*, *by and large*, ...), idioms, compound nominals (*car park*, *part of speech*), proper names or named entities, verb-particle constructions (*look up*, *fight on*...), light verb constructions (*give a demo*, *make a mistake*) or institutionalized phrases (*traffic light*, ...). These are extreme cases of the formulaicity of language I have argued on in the latter chapter, and in the NLE community they are usually grouped under the name of *multiword expressions*.

Multiword expressions are a key problem for the development of large-scale language processing technology. Treating them by general compositional methods leads to overgeneration problems, because the formulaicity of language is not captured and hence the combinations generated do not sound natural. Furthermore, compositional methods do not predict idiosyncratic meanings such as those of idioms, in which the meaning is not derived from the combination of the parts. Another traditional approach that is problematic is treating multiword expressions as words-with-spaces. This technique suffers from a flexibility problem, since it does not adapt to the insertion of modifiers or changes in word order. Moreover, listing each multiword provokes a loss of generality, and its cost makes these techniques difficult to extend and maintain [Sag et al., 2002].

The next sections review recent work that tries to make progress in the identification of multiword expressions or other linguistic constructions in order to overcome the problems of processing them compositionally or listing them in a word-with-spaces fashion.

3.1. Frequency and Association Measures

Multiword expressions can be identified or validated using statistical information. Raw frequency and association measures account for the lexical association between words in the candidate tuples. The quality of association measures and their appropriateness to identify multiword expressions is reviewed by several authors [Evert and Krenn, 2001, Ramisch et al., 2008, Pecina, 2008]).

Watrin and François [2011] apply association measures to identify nominal structures. They define an n-gram frequency database in order to improve the efficiency of computing such measures, and to ensure that lexical resources do not lack any of the necessary data. The candidates are filtrated using part-of-speech patterns, and then they are validated with the association measures. Since the common association measures are not directly applicable to candidates larger than bigrams, they select three measures that go beyond

that limitation: the Fair Log-likelihood Ratio [da Silva and Lopes, 1999], the Symmetrical Conditional Probability [da Silva and Lopes, 1999] and the Mutual Expectation [Dias et al., 1999]. The latter is useful even in the case of non-contiguous multiword expressions. Instead of the usual selection of a threshold to decide whether a candidate is a multiword expression or not, they use the LocalMax algorithm, which selects the multiword expressions whose Association Measures are higher than those of their neighbourhood.

Although not focusing so much in association measures, Wible and Tsao [2010] extract n-grams at different levels of abstraction, which are, from concrete to abstract: the lexical item, its lemma, a rich part-of-speech tag (taken from the CLAWS set) and a rougher part-of-speech tag. They only select n-grams that have at least one lexicalized item, and that appear with a minimum frequency of five. After that, a pruning process is applied. Vertical pruning deletes a candidate n-gram when it appears as a parent of another n-gram (i.e., in a more abstract realization of it) in a proportion above a threshold. In a similar manner, horizontal pruning deletes a candidate n-gram when it is included in a longer n-gram, and the proportion of the shorter n-gram appearing as part of the longer n-gram is above a threshold. The authors implement a web interface (namely, the StringNet website[1]) to browse the extracted multiword expressions, and they cross-index the n-grams so that the visitors of the web page can navigate through the different levels of abstraction of the multiword expressions.

Attia et al. [2010] present three approaches. Two of them are reviewed in the Machine Translation and Paraphrase section. The third one applies Pointwise Mutual Information and Chi-square formula to all bigrams and trigrams with frequency above a threshold in the Arabic Gigaword corpus. This approach differs from the main trend of finding part-of-speech patterns and then applying Association Measures because it works the other way around, computing first the Association Measures of frequent bigrams and trigrams and filtering the most promising results so that unlikely part-of-speech patterns are excluded.

Stefanowitsch and Gries [2003] present a framework for analysing the interaction of words and meaningful constructions, the so-called Collostructional Analysis. Inspired by the classical approach of examining collocates, their new framework extends this analysis to include grammatical structure. They provide a family of statistical methods that characterize lexico-grammatical constructions.

Within this family, the method of Collexeme analysis investigates the association between a construction and the words occurring in a particular slot in the construction. The Distinctive Collexeme Analysis focuses on the association between a word and (one member of) two or more semantically or functionally equivalent constructions. The Covarying Collexeme Analysis quantifies the association between pairs of words occurring in two different slots in the same constructions.

All methods start with a manually chosen grammatical construction that may have different degrees of specificity. They conclude with a ranking of most attracted and repelled words in a particular slot of the construction. The analysis does not rely on raw frequency, but computes a distribution table over which the Fisher-Yates exact test is (typically) used. The Association Measure applied is the p-value or the negative base-10 logarithm of the p-value, although others can be chosen [Gries and Stefanowitsch, 2009].

[1] http://nav.stringnet.org/

3.2. Machine Learning

Tu and Roth [2011] identify light verb constructions using a Super Vector Machine to classify six manually chosen verb candidates. The features are extracted automatically, and they include Association Measures, statistics such as the Deverbal v/n ratio (computed using WordNet and NomLex), the phrase size and some contextual linguistic features (for instance, the part of speech of the word before the verb).

Constant and Sigogne [2011] approach the identification of multiword expressions as a problem of part-of-speech tagging. They implement a finite-state lexical analyzer and a Conditional Random Field decoder, and compose them using weighted transducers composition. Features in the Conditional Random Field use the information added by the part-of-speech tagger.

Crysmann [2011] extracts relational nouns training different classifiers. Interesting unigrams and bigrams are extracted from the deWaC corpus [Baroni and Kilgarriff, 2006] -which is annotated with the TreeTagger [Schmid, 1995]- and frequency lemma-based counts are calculated. These counts are useful to compute different Association Measures that will be used as feature information by the classifiers. The paper presents an extensive evaluation of the different classifiers and the particular contribution of each Association Measure.

3.3. Machine Translation and Paraphrasing

When a single word can be paraphrased with a longer expression, the expression is probably a multiword expression: since the semantics can be expressed with a single word, the longer expression will probably be non-decomposable. The same principle holds for translation in different languages: an expression that is translated into a single word in other language is a candidate to be a multiword expression for the same reason.

Attia et al. [2010] make progress in this direction. They find assymetries between the Arabic and 21 different languages using Wikipedia titles. The authors also present another approach based on translation that extracts the multiword expressions of an English corpus (Princeton WordNet 3.0) and automatically translates them into Arabic with Google Translate tool. The Arabic results are validated using frequency counts in order to conclude whether the translation remains a multiword expression or not. The latter approach underlies the insight that a multiword expression in a language is likely to be translated as a multiword expression in another.

3.4. Semantic Approaches

Schone and Jurafsky [2001], Baldwin et al. [2003] and Katz and Giesbrecht [2006] use Vector Space Models to compute the semantic similarity between a whole multiword expression and each of its components, with the intuition that low similarity indicates that the candidate is a linguistic construction whose meaning has deviated from the meaning of its parts.

In the same line, Van de Cruys and Moirón [2007] apply clustering techniques to nouns in order to classify them according to their semantic similarity. They compute measures of selectional preferences between nouns in the cluster and verbs, following the intuition

that a noun within a multiword expression cannot be easily replaced by a semantically similar noun.

Chakraborty et al. [2011] identify noun compounds with the assumption that the semantics of each noun integrating a multiword expression is diminished from their original meaning. Using the *WordNet::Similarity* module they look for the synsets of each component of the bigram. They intersect both synsets, and the words in the intersecting set act as features of a Vector Space Model. The Cosinus and the Euclidean distances are computed in order to obtain similarity measures of the isolated words and the words occuring in the compound. A threshold is established to decide which words are too similar to their original meaning to be part of a multiword expression.

3.5. Segmentation Approaches

The identification of multiword expressions can be viewed as a segmentation problem in which we aim to find which are the units of language, without taking into account the notion of word as a unit separated by a delimiter (e.g. the space character). Chinese language is a good starting point, since it does not have a textual word delimiter. Xu et al. [2010] use the Tightness Continuum Measure to quantify the degree of compositionality between Chinese characters. The authors take four-grams as the starting point, using the heuristics that Chinese most usual compounds are made up of four characters. Taking into account all possible segmentations of a four-gram, they compute a ratio based on the frequencies of the segmented patterns. Therefore, the lack of certain patterns in the data is regarded as evidence for the tightness of the units.

Duan et al. [2006] approach the problem in a similar manner, but in a bio-inspired fashion. They notice that the identification of multiword expressions is a problem similar to that of gene sequence alignment. They apply the Longest Common Subsequence theory, and experiment with a purely mathematical bioinformatic method and another enriched with linguistic heuristics. Similar to Xu et al. [2010], this approach regards the problem also as a uniform stream that must be segmented to identify the units that compose it.

3.6. Data Oriented Parsing

Data Oriented Parsing (DOP), first developed by Scha [1990], is a framework for exemplar-based statistical parsing and language modelling. It uses fragments of trees of arbitrary size and shape as elementary units of combination, so it provides a good environment to identify constructions. The model has been used, among other applications, to extract linguistic patterns without prior candidate filtering, with the assumption that a construction can be considered linguistically relevant if there is some empirical evidence about its reusability.

One example is the proposal by Zuidema [2006]. The method searches the best statistical grammar that describes a corpus, allowing its productive units to surpass word boundaries. He uses the Stochastic Tree Substitution Grammars (STSG) formalism, which is useful to represent single words, contiguous and noncontiguous multiword expressions, context-free rules or complete parse trees. The 'push-and-pull' algorithm finds linguistic constructions based on occurrence and co-occurrence frequencies in the corpus: when the expected frequency of a tree is higher than the observed one, the difference is

substracted from the tree's score and added to the derived trees; in the opposite case, it is substracted.

Later on, Zuidema [2007] proposes P-DOP, an approach to the same problem with the aim of improving the efficency of DOP. He applies statistical restrictions in order to prune all the possible subtrees that could be identified as productive units. Borensztajn et al. [2009] extends this work to identify multiword expressions in children's language in order to confirm the progression from concrete towards abstract constructions proposed by Tomasello (see section 2.2).

Using a kernel-based method in the DOP framework, Sangati et al. [2010] also identify syntactic subtrees (*fragments* and *partial fragments*) that contain a specific lexical piece and the most frequent subtrees it participates in. The system selects those constructions which are recurrent in a treebank by iteratively comparing every possible pair of structures in the corpus.

3.7. Environments and toolkits

Martens [2010] and Martens and Vandeghinste [2010] implement the *Varro* toolkit with the purpose of identifying unordered syntactic subtrees in treebanks. They introduce *condensed canonically ordered trees* as a data structure for discovering frequently recurring unordered subtrees, and they use the Apriori discovery algorithm, which is known as an algorithm used to discover frequent items in datasets.

Sanches Duran et al. [2011] define the part-of-speech patterns that are common in Brazilian Portuguese and extract them using the *mwetoolkit*[2]. This toolkit is developed by De Araujo et al. [2011], and it provides and environment to filter part-of-speech patterns, extract n-gram frequencies and compute association measures.

Kulkarni and Finlayson [2011] develop the *jMWE toolkit*[3] to identify multiword expressions in general. It allows for the manual definition of part-of-speech patterns to be searched, so it may be combined with *mwetoolkit* for that purpose. The software also provides indices to retrieve patterns associated to a specific word and part-of-speech. Additional tools are provided, such as filters and different ways of solving conflicts when multiword expressions overlap.

Pedersen et al. [2011] implement another software package to extract predefined patterns. Regular expressions can be used to identify them, and it provides several tools to calculate frequencies of n-grams and a range of association measures.

[2] www.sf.net/projects/mwetoolkit

[3] http://projects.csail.mit.edu/jmwe/

4. A computational proposal

This chapter presents my first computational approach to the identification of linguistic constructions. As follows from chapter 3, there are many different techniques to approximate the problem. This thesis presents an approach based on Vector Space Models of Semantics.

Vector Space Models of Semantics (VSMs) are computational models that characterize lexical meaning in terms of syntagmatic context of occurrence. These models project a multidimensional space where semantically related words are geometrically close. Hence, these models provide a way to characterize words in a bottom-up fashion that derives the model entirely from linguistic data, as well as a similarity function that quantifies the semantic relation between words.

VSMs can be built in different forms depending on how context is characterized or how the similarity measure is computed. With a good VSM and a similarity measure there are many ways to approximate the problem of finding linguistic constructions. Linguistic insights about semantic properties of linguistic constructions can be used to decide if candidates are formulaic pieces of language; furthermore, a study of composition over these models would provide knowledge about how the composition of linguistic pieces inside prefabs differs from the 'casual' compositionality among less frequent expressions.

This chapter is divided into two sections. Section 4.1 presents an overview of Vector Space Models of Semantics, as well as the research I have done over several VSMs to find the best similarity measure. Section 4.2 explains how I tried to take advantage of VSMs in order to detect one particular type of linguistic construction -the Light Verb Construction- and provides an analysis of the obtained results.

4.1. Vector Space Models of Semantics

Vector Space Models of Semantics (or VSMs) are computational models of word meaning that make use of distributional information of words collected over a large corpora to represent semantic similarity between these words in terms of spatial proximity [Sahlgren, 2006]. The models can be also used to model document similarity, taking into account the distribution of words occurring in each documents. VSMs have been extensively used in linguistic research, as well as on applications such as document retrieval or document classification (Turney and Pantel, 2010; Baroni and Lenci, 2010; Padó and Lapata, 2007; Erk, 2010).

The distributional information obtained from corpora is arranged in a matrix. Each row corresponds to a word and each column is an event. The observed co-occurrence of the expression in the row with the event in the column is stored in the correspondent cell. The events in the columns correspond to relevant information that characterizes the words in the rows: either words that co-occur with the word in the row in a window

of a predetermined size, or syntactic paths in which the expression appears in, or even documents that contain the word in the row.

VSMs are based on the hypothesis that near points correspond to words that tend to appear in the similar context [Harris, 1954]. The meaning of a word is modelled with the n-dimensional vector that is formed by taking only its correspondent row. The positions of the vector are interpreted as coordinates of a point in an n-dimensional space. Points that are closer in that space correspond to semantically related expressions, since the distance between points is less when the expressions have similar values in the same coordinates, or in other words, when the expressions tend to appear in the same contexts.

4.1.1. Resources

We depart from the the Araknion corpus, a subset of the 300 million words EsPal corpus[1]. The corpus has been automatically tagged, and the words that were not recognized by the tagger have been semi-automatically reviewed: some of them have been substituted with an equivalent word with its correspondent part-of-speech tag, and when there was a disambiguation problem, it has been substituted with the most probable word. Furthermore, several HTML tags that were part of the corpus have been removed. Araknion is a subset of 96 million words that corresponds to the part of the corpus that has been most benefited from the cleaning process.

After that, the corpus has been enriched with syntactic informaiton. This task will be accomplished by the Treeler dependency parser[2]. Treeler has been trained using the Ancora corpus[3], a Spanish corpus of 500.000 words.

4.1.2. The Models

The matrices are the basic component of any VSM. They contain the co-occurrence information that is later exploited to calculate the proximity measures. Depending on the information they contain and how the notion of context is defined, each matrix defines a different semantic model.

The context of a focus word *fw* is defined as the window containing the words surrounding *fw*, including *fw*. A window of size *s* contains the *s* words appearing in the left of the focus word, the focus word itself, and *s* words appearing in the right of the focus word. In matrices that include syntactic information, the context of a word is collected from the dependency tree that corresponds to the sentence rather than from the actual plain text. The notion of window is changed to that of a syntactic path, namely, a sequence of edges and nodes extracted from the correspondent dependency tree.

Once the matrices are built, a similarity measure is computed comparing each of the words in the rows among them. As explained in Jurafsky et al. [2000, section 20.7.3], there exist several measures, such as the Euclidean distance, the Manhattan distance, Jaccard measures (including variations such as Greffenstete or Dice) and Cosinus. Panchenko [2011] compares several distance measures, and his study shows that Cosinus is the corpus-based measure that performs best. It is also known that distances from the Minkowsky family, such as the Manhattan or the Euclidean, are too much sensitive to extreme results. Taking this into account, the similarity measure used in this project is the Cosinus.

[1] http://clic.ub.edu/ca/espal
[2] http://treeler.lsi.upc.edu/treeler/
[3] http://clic.ub.edu/corpus/ancora

The *term-term* matrix

In this matrix, columns correspond to terms that appear around a focus word in a window of size 3. The rows contain the represented words, which are all type of words except punctuation, numbers, dates, determiners and conjunctions. This same filter is applied to the terms in the columns. Each cell in the matrix records the number of times that the word in the corresponding column is part of a window of which the word in the row is the focus word. The exact position that a word occupies in a context window is not taken into account, and the frequency of the focus word itself is considered zero.

Table 4.1 shows a toy example of how a matrix in the *term-term* format. The matrix captures the distributional information from a couple of sentences: *El gato come pescado* and *La gaviota come pescado*. As we can see, *gato* and *gaviota* are represented by equal vectors because they appear in the same context.

	gato	come	pescado	gaviota	...
gato	0	1	1	0	
gaviota	0	1	1	0	
come	1	0	1	1	
pescado	1	1	0	1	
...					

Table 4.1: Toy example of a term-term matrix.

The *term-dep3.5* matrix

In this model, contexts are formed by syntactic paths that contain either words that are directly connected to the focus word, or words connected to the focus word through a preposition. Contexts are expressed in the form $[< \mid >] : * : finalterm$. The symbol $<$ indicates that the focus word is a daughter of the following word in the path, and the a symbol $>$ indicates that the focus word is parent of the following syntactic path.

Table 4.2 shows the contexts derived from the sentence *El gato come pescado del bueno*.

The *term-dep4* matrix

This model is similar to *term-dep3.5*, but enriched with dependency relations. Contexts are formed by syntactic paths that either contain words that are directly connected to

Term	**Context**
gato	<:*:come
come	>:*:gato
come	>:*:pescado
pescado	<:*:come
pescado	>:*:bueno

Table 4.2: Example of contexts generated by the model *term-dep3.5*.

the focus word, or words connected to the focus word through a preposition. They are expressed in the form $[< \mid >] : [rel|*] : [finalterm|*]$. These syntactic paths include information about direction: a symbol $<$ indicates that the focus word is a child of the following word in the path, and the a symbol $>$ indicates that the focus word is parent of the following syntactic path.

Each context is generated in three versions: one with the dependency relation included, another where it is not considered (to avoid a bias towards the default relation assigned by the parser, namely, '_') and a third one where the term itself is not considered (only the direction of the path and the dependency relation). When some information is not considered, it is substituted with a wildcard symbol '*'.

Table 4.3 shows the contexts derived from the sentence *El gato come pescado del bueno*.

Focus Word	**Context**
gato	<:*:come
gato	<:suj:come
gato	<:suj:*
come	>:*:gato
come	>:suj:*
come	>:suj:gato
come	>:*:pescado
come	>:cd:*
come	>:cd:pescado
pescado	<:*:come
pescado	<:cd:come
pescado	<:cd:*
pescado	>:*:bueno

Table 4.3: Example of contents generated by the model *term-dep4*.

4.1.3. Evaluation of the models

In order to evaluate if these models are providing a good representation, I have extracted synonymy relations from the Spanish Wordnet [4]. Each pair of synonyms retrieved from WordNet is expected to have a high similarity measure. In order to proof to what extent this is happening, I compute the Rank measure.

Consider a pair of synonyms, such as *acentuar* - *enfatizar*. To compute the Rank, *acentuar* is listed in combination with each of the 3000 most similar words. This list of pairs is ordered in decreasing order, so that similar terms are on the top. Then, a number is assigned to each pair, so that the most similar pair receives 1, the second 2, etc. The number corresponding to the pair *acentuar* - *enfatizar* is saved, and the same process is applied to *enfatizar*. The final rank of the pair is the average between the two numbers obtained in that way. The Rank measure of the whole model is the average of the ranks computed for each pair.

[4] http://www.illc.uva.nl/EuroWordNet/

The Rank measure is better when lower. However, the pair obtaining the number one rank is that composed of the same word (e.g., *acentuar* - *acentuar*), so the best result is a rank of two, and the worse is 3000 (because of the filter previously applied).

Table 4.4 lists the Rank measure of the models. All the models have a good performance taking into account that the number is far from 3000. Nonsurpisingly, the best performance is for *term-dep4*, the model which is enriched with more precise syntactic information. The *term-term* model is probably better than the *term-dep3.5* model because the *term-dep3.5* incorporates very few syntactic information (only the direction of the paths) and restricts the number of words appearing in a context in comparison with the 3-sized window of the *term-term* model.

Model	**Rank Measure**
term-term	364.54
term-dep3.5	645.16
term-dep4	236.77

Table 4.4: Ranking measure for all the models.

In order to improve the similarity measure, the models can also be submitted to a dimensionality reduction, so only relevant columns are considered. Singular Value Decomposition or SVD is a technique that allows such a reduction. The columns are weighted depending on their relevance in the distance between points. Less relevant columns can be cut out in order to reduce the space where words are represented, so different models are obtained depending on the number of columns reduced.

Figures 4.1 and 4.2 show the performance of the models with different number of columns left. While for *term-term* the model improves whenever it has more columns, making the reduction unnecessary, *term-dep4* achieves an improvement when reduced to 850: it seems that the model gets rid of redundant columns, which are probably motivated for introducing each context three times (due to the different levels of abstraction represented).

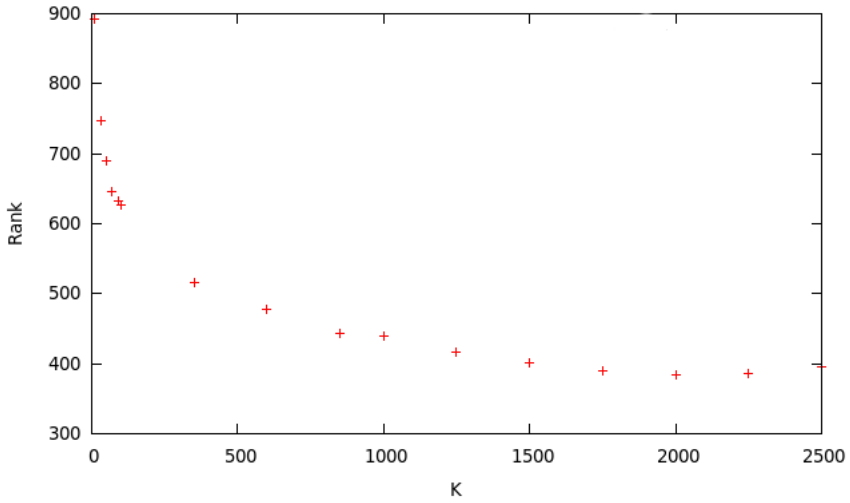

Figure 4.1: SVD for *term-term* model.

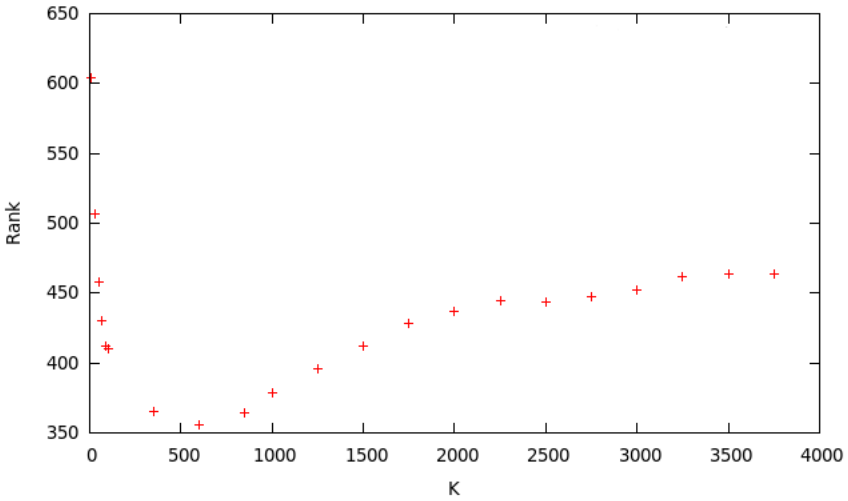

Figure 4.2: SVD for *term-dep4* model.

4.2. Using VSMs to find linguistic constructions

This section presents my first approach to the identification of linguistic schemas using VSMs. The method I have developed focuses on Light Verb Constructions (LVC). These constructions are well known to linguists: they are characterized by the fact that the semantic contribution of the verb is reduced, hence the meaning of the construction comes mainly from the noun object.

I center my attention on two specific types of LVCs: those consisting on a commonly used verb and a noun phrase in its direct object position (e.g. *dar miedo*), and those that follow this same pattern plus a preposition (e.g. *tener necesidad de*). These constructions usually have a semantic equivalent in the same language, expressed by a single verb. For instance, *dar miedo* is equivalent in meaning to *asustar*. Therefore, in order to detect the LVC, it should suffice to check whether they have an equivalent coded in a single word.

We know that semantic equivalents can usually be interchanged in the same context [Harris, 1954]. Since Vector Space Models of Semantics (VSM) take advantage of this assumption to quantify semantic similarity based on context of occurrence, it seems feasible to use them in order to identify LVCs.

4.2.1. Related Work

Chapter 3 already surveys the state of the art on the identification of linguistic schemas. However, it is important to emphasize the approaches that are more closely related to the one I am proposing.

Baldwin et al. [2003] use a VSM approach to detect multiword expressions in corpora. They are interested in non-decomposable multiword expressions. Their insight is that VSM should reflect the non-decomposability of an expression: a multiword expression as a whole should appear in the model as distant to its constituents. Katz and Giesbrecht [2006] follow the same intuition. Their goal is to confirm that the local context of a known idiom can reliably distinguish idiomatic uses from non-idiomatic uses. Furthermore, they attempt to determine, for each use of an expression, whether its context can help in

distinguishing whether it is being used as an idiom or as a compositional expression. Therefore, these approaches take advantage of VSMs in terms of similarity between the whole and the parts. The difference with my approach is that the LVC candidates are considered as a whole, and they are not compared to its parts but to other words in the corpus.

Ingram and Curran [2007] model the similarity between words and bigrams, and within bigrams themselves. This approach is closely related to mine, but I provide a more extensive study of the results and how they could be applied to the detection of LVCs.

4.2.2. The method

The goal of this approach is to identify LVCs assuming that for each LVC exists a single verb that is equivalent in meaning to the whole construction. To find this semantic similarity, the matrices need to represent constructions, that is, the expressions in the rows of the matrices need to incorporate LVC candidates.

In order to have a list of candidates, I have extracted bigrams and trigrams from the Araknion corpus, following the patterns [V N] and [V N P]. The latter try to account for LVCs such as *tener necesidad de* (*to be in need of*). In order to solve the conflict between a trigram and the bigram formed by the verb and noun in the trigram, I have established the heuristics to choose the trigram only if it appears in the corpus at least 80% of the occurrences of the bigram. For instance, *tener necesidad de* appears 642 times in the corpus, while *tener necesidad* appears 765 times; the trigram appears the 83% of the times that appears the bigram. The 80% threshold has been chosen after some tests: a higher threshold left out many interesting cases, and a lower threshold had the tendency to permit trigrams containing the same combination of verb and noun but different prepositions.

In addition to the threshold for trigrams, I have applied a frequency filter of 50, which seemed fair in order to leave out underrepresented cases. The resulting list is formed by 1311 candidates, of which 59 are trigrams.

These candidates need to be represented in their own VSM. The evaluation provided in section 4.1.3 demonstrates that *term-dep4* provides the best performance. However, I rejected the idea to use that model for LVC candidates precisely because of the syntactic tags. The noun in a LVC usually responds to a dependency relation of direct object; for instance, in *El general dio la orden de atacar* (*The general gave the order to attack*) *orden* is the direct object of the verb *dar*, so *el ataque* must have another syntactic relation. However, in *El general ordenó atacar* (literally, *The general ordered to attack*), the direct object is *el ataque*. So using the dependency relations would be misleading, since the direct object relation would cause *dar orden* and *ordenar* to be considered as appearing in different contexts.

Due to the inadequateness of *term-dep4*, I model the LVC candidates following the format of *term-term* and *term-dep3.5*. The resulting matrices, *vnvnp-term* and *vnvnp-dep3.5*, provide models over which the similarity measure between LVC candidates and words is computed. In order to finally classify which LVC candidates are truly LVCs, a similarity threshold must be chosen, so to accept as positive LVC those candidates whose similarity with some other verb is above the threshold. The following sections explain the performance of the models for several thresholds, as well as an analysis of the results of the best performing model.

4.2.3. Evaluation

Each model is expected to identify true LVCs from among the candidates. In order to evaluate if this goal is achieved, it is necessary to know which candidates are really LVCs according to our linguistic knowledge. Inspired by Evert and Krenn [2001], I decided to ask a person outside the experiment to annotate the whole list of 1311 candidates, indicating whether she would consider the candidate a LVC or just a 'casual' combination of verb and noun (and preposition).

Each of the two models is evaluated with 10 different thresholds that go from 0.5 to 0.95 in steps of 0.05. So there are 20 evaluations to be done. For each of them, I have computed the confusion matrix and the precision and recall measures (see table 4.5 and 4.6).

According to the confusion matrix of *vnvnp-deps3.5*, the lowest threshold obtains the best recall, 0.980. With that configuration, the model recovers 347 true positives out of 454. However, the precision is very low, since with such a low threshold we also identify 898 false positives. The best precision, which is only 0.325, is achieved with threshold 0.80. Of the 565 candidates classified as LVCs, 184 are true positives. But 170 candidates are incorrectly identified as not being LVCs, so the recall decreases to 0.519.

As for the *vnvnp-term*, both thresholds of 0.60 and 0.65 achieve the best recall, namely 0.988. The number of true positives identified is 350, and only 4 false negatives are left out. However, the precision is less than 0.3 for both thresholds: around 900 candidates are incorrectly predicted to be LVCs. The best precision achieved, which is 0.4, needs of a higher threshold, namely 0.90. With this threshold the recall decreases to 0.480, since the model lefts out 184 LVCs.

Figure 4.3 shows the precision/recall curves of each model. None of the model achieves high precision, which means that the models have a tendency to identify a large number of false positives. Nevertheless, it is clear that *vnvnp-term* performs better, since its precision is higher than *vnvnp-dep3.5*'s regardless of the recall. Therefore, *vnvnp-term* is the model for which I provide further analysis in the next section.

Threshold	**True Positives**	**False Positives**	**False Negatives**	**Precision**	**Recall**
0.50	347	898	7	0.278	0.980
0.55	342	868	12	0.282	0.966
0.60	333	820	21	0.288	0.940
0.65	320	747	34	0.299	0.903
0.70	293	649	61	0.311	0.827
0.75	247	528	107	0.318	0.697
0.80	184	381	170	0.325	0.519
0.85	108	238	246	0.312	0.305
0.90	42	130	312	0.244	0.118
0.95	10	48	344	0.172	0.028

Table 4.5: Confusion matrix for the *vnvnp-deps3.5* model.

Threshold	True Positives	False Positives	False Negatives	Precision	Recall
0.50	351	939	3	0.272	0.991
0.55	351	925	3	0.275	0.991
0.60	350	908	4	0.278	0.988
0.65	350	882	4	0.284	0.988
0.70	347	847	7	0.290	0.980
0.75	341	789	13	0.301	0.963
0.80	326	705	28	0.316	0.920
0.85	271	508	83	0.347	0.765
0.90	170	253	184	0.401	0.480
0.95	49	87	305	0.360	0.138

Table 4.6: Confusion matrix for the *vnvnp-term* model.

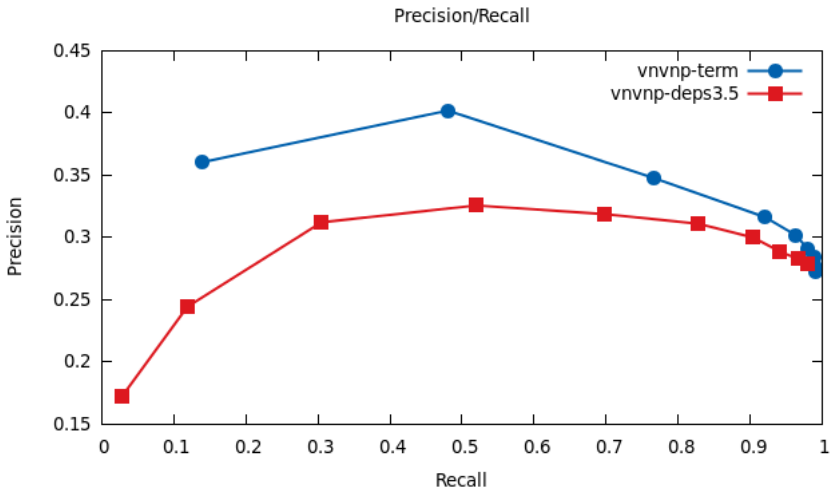

Figure 4.3: Precision/Recall curves for both models.

4.2.4. Analysis of the results

The previous evaluation presents numbers related to the correct and incorrect predictions of the models. However, it is also necessary to review which are the verbs that the model considers to be similar to a candidate, and what is the similarity measure assigned to each one. It may happen that even good predictions are based on an improper similarity measure. This section aims to review the similarity measures for the best model, *vnvnp-term*, set to the most exigent threshold, 0.95.

Table 4.7 presents a list of some true positives identified by the method. Each candidate is related to the most similar verbs provided by the model, its ranking in the ordered list of similar verbs and the similarity measure. I have also added some supposedly similar verbs according to my linguistic criteria, also with its ranking and similarity measure.

True positives are identified because the method assigns them very high similarity to other verbs. However, table 4.7 shows that most of these verbs are not semantically related to the LVC candidate. The exceptions are *poner énfasis* with *insistir*, *sacar provecho* with *disfrutar* and *poner límite* with *renunciar*; *contraer matrimonio* and *convivir*

are not synonyms but are related in terms of encyclopaedic knowledge. The verbs I proposed as similar according to my linguistic knowledge appear very far in the ordered list of similar verbs, being 35 the best result. The distances for these verbs are generally low, except for the cases of *mencionar* and *limitar*.

Table 4.8 presents some candidates that are identified as LVCs although not being so. As before, each candidate is presented with its five most similar verbs according to the model, along with their ranking and similarity measure. As it can be seen, the similarity measure for the most similar verbs is very high, over 0.9 in all cases. In spite of that, the candidate as a whole is not semantically related to the verbs, since it does not have a cohesive meaning. So a possible guess would be that these verbs are similar to the verb inside the candidate, but a look to the table reveals that this is not the case.

The LVCs that were not detected by the model are listed in table 4.9. The similarity measure is obviously less than 0.95 in all cases, but the problem is not due to this quantity, but to the quality of the verbs that are deemed to be more similar: they are semantically unrelated in all cases except for *emprender viaje* with *regresar* and -loosely- *hacer burla* with *renegar*.

The analysis of these particular cases shows that even when the method is successful, it is not due to a good semantic measure. However, the *term-term* model in which *vnvnp-term* is based has shown to perform well in the ranking (see section 4.1.3). In order to compare both models, table 4.10 presents particular examples.

Table 4.10 is focused on an arbitrary LVC, particularly *dar miedo* (literally, *to give fear*, which means *to scare*). For this construction, the three most similar verbs are presented along with its ranking and similarity, as well as the ranking and similarity of two well known true synonyms of this construction (*asustar* and *espantar*). The rest of the columns account for these true synonyms and also for two similar verbs according to the model, *gustar* (*to like*) and *perdonar* (*to forgive*). For each of these verbs, the table lists the ranking and similarity of the three most similar verbs according to the *term-term* model, along with the ranking and similarity between themselves.

The verbs appearing in the first three positions on the ranking are semantically related to the verb heading the column, with the exception of the verbs related to *gustar* and the particular case of *reponer*. This demonstrates, at least for these cases, that the *term-term* model works as expected. However, *asustar* and *espantar*, which are synonyms, score low in the similarity measure between each of them and *gustar* or *perdonar*. This result is not surprising taking into account our linguistic knowledge, but since *gustar* and *perdonar* are the verbs considered similar to *dar miedo* by the *vnvnp-term* model, this data shows that the models do not agree in the similarity measure.

To further analyse this fact, I have calculated the quantity of common columns in the matrices - which stand for contexts- related to *dar miedo* and its synonyms. *Asustar* appears in 2357 contexts according to the *term-term* model, and *espantar* in 1170. However, the *vnvnp-term* model only records 380 contexts for *dar miedo*. *Asustar* and *espantar* share 480 contexts, while *dar miedo* shares 129 with *espantar* and 231 with *asustar*. The intersection between the three expressions is 113.

So it seems that there is something going on when the same model is computed for representing multiword units rather than single words. Although the contexts are retrieved in the same way, the similarity measure relating multiwords and words is completely deviated in most cases. A possible explanation for this fact may be lack of data: while the corpus is big enough to model single words, multiword units have less frequency in text, so the model may be lacking more occurrences in order to have representative contexts.

Finally, it is worth to mention that there is some noise in the data that deviates a few results. Among the 1311 candidates selected, 25 are verbs combined with the word *don*, which is a title placed before a man's name. This word plus the man's name it accompanies should be recognized as a named entity, and actually there are some cases in the corpus where the words are joined and assigned a part-of-speech of proper noun (e.g. *don_Manuel NP00000*) . However, there are 29033 occurrences of *don* in the corpus where it is identified as a common name. The candidates that contain *don* behave in a different manner: their most similar verb according to the models is related to the verb in the candidate. For instance, *contestar don* (literally, *to answer 'don'*) is related to *exclamar* (*exclaim*) with a similarity measure of 0.97 in *vnvnp-deps3.5*. This case is strange taking into account that the similar verbs are not usually related to the verb inside de construction; it seems that for the particular case of candidates whose noun has little semantic content the model simply works as if this noun was omitted.

LVC	**Most similar verbs**	**Supposedly similar verb(s)**
contraer matrimonio	(1, 0.964) coincidir	(1953, 0.075) casarse
	(2, 0.951) tropezar	
	(3, 0.942) convivir	
	(4, 0.930) colaborar	
	(5, 0.923) pactar	
dar cita	(1, 0.956) fundir	(761,0.565) citar
	(2, 0.948) diluir	
	(3, 0.948) revolcar	
	(4, 0.945) alistar	
	(5, 0.944) refugiar	
hacer mención	(1, 0.952) excluir	(35, 0.911) mencionar
	(2, 0.942) partir	
	(3, 0.938) exceder	
	(4, 0.937) recelar	
	(5, 0.936) copiar	
poner énfasis	(1, 0.969) insistir	(315, 0.549) enfatizar
	(2, 0.951) consistir	
	(3, 0.951) radicar	
	(4, 0.946) tomar_parte	
	(5, 0.946) incurrir	
poner límite	(1, 0.964) renunciar	(42, 0.899) limitar
	(2, 0.956) aludir	
	(3, 0.952) recurrir	
	(4, 0.952) apelar	
	(5, 0.951) hacer_referencia	
tomar nota	(1, 0.977) hacer_uso	(1235, 0.543) anotar
	(2, 0.970) renegar	
	(3, 0.965) prescindir	
	(4, 0.965) carecer	
	(5, 0.965) formar_parte	
sacar provecho	(1, 0.973) hacer_uso	(127, 0.847) aprovechar
	(2, 0.967) prescindir	
	(3, 0.965) disfrutar	
	(4, 0.960) carecer	
	(5, 0.958) formar_parte	

Table 4.7: Some true positives identified by the method with threshold 0.95, with the most similar verbs provided by the model along with its ranking in the ordered list of similar verbs and the similarity measure, and some supposedly similar verbs according to my linguistic criteria, also with the ranking and the similarity measure.

LVC	Most similar verbs
dar vista	(1, 0.965) aludir
	(2, 0.963) hacer_referencia
	(3, 0.963) equivaler
	(4, 0.954) pertenecer
	(5, 0.953) instar
determinar momento	(1, 0.975) radicar
	(2, 0.966) consistir
	(3, 0.962) insistir
	(4, 0.960) irrumpir
	(5, 0.958) estribar
haber momento	(1, 0.956) desembocar
	(2, 0.949) residir
	(3, 0.946) imperar
	(4, 0.943) incidir
	(5, 0.941) tener_lugar
ir camino	(1, 0.959) calificar
	(2, 0.952) tildar
	(3, 0.950) formar_parte
	(4, 0.948) proveer
	(5, 0.944) colmar
poner cara	(1, 0.954) hacer_uso
	(2, 0.948) encuestar
	(3, 0.948) prescindir
	(4, 0.947) renegar
	(5, 0.945) despojar
ser patrimonio	(1, 0.961) carecer
	(2, 0.960) formar_parte
	(3, 0.956) hacer_uso
	(4, 0.956) provenir
	(5, 0.946) prescindir
tener cargo	(1, 0.954) brotar
	(2, 0.953) emanar
	(3, 0.953) datar
	(4, 0.951) pender
	(5, 0.949) emerger

Table 4.8: Some false positives identified by the method with threshold 0.95, with the most similar verbs provided by the model along with its ranking in the ordered list of similar verbs and the similarity measure.

LVC	Most similar verbs	Supposedly similar verb(s)
adoptar decisión	(1, 0.838) reservar	(39, 0.763) decidir
	(2, 0.836) combatir	
	(3, 0.832) prevenir	
	(4, 0.827) habilitar	
	(5, 0.803) comprar	
caer enfermo	(1, 0.876) partir	(35, 0.827) enfermar
	(2, 0.865) sugerir	
	(3, 0.862) recibir	
	(4, 0.857) sondear	
	(5, 0.855) sacar	
dar aviso	(1, 0.940) subir	(20, 0.919) avisar
	(2, 0.936) obedecer	
	(3, 0.934) destinar	
	(4, 0.933) recurrir	
	(5, 0.933) apelar	
dictar orden	(1, 0.856) tildar	(307, 0.691) ordenar
	(2, 0.853) calificar	
	(3, 0.852) renegar	
	(4, 0.851) hacer_uso	
	(5, 0.850) formar_parte	
ejercer presión	(1, 0.833) reflexionar	(33, 0.471) presionar
	(2, 0.688) meditar	
	(3, 0.651) versar	
	(4, 0.627) recaer	
	(5, 0.579) cerner	
emprender viaje	(1, 0.880) regresar	(22, 0.835) viajar
	(2, 0.875) ascender	
	(3, 0.865) apelar	
	(4, 0.860) asistir	
	(5, 0.858) vencer	
hacer burla	(1, 0.946) renegar	(124, 0.846) burlar
	(2, 0.944) hacer_uso	
	(3, 0.943) despojar	
	(4, 0.942) proveer	
	(5, 0.941) tachar	

Table 4.9: Some false negatives that the method did not identify when establishing a threshold of 0.95. Each candidate is related to the most similar verbs provided by the model and a supposedly similar verb according to my linguistic criteria. All the verbs are presented along with their ranking in the ordered list of similar verbs and the similarity measure.

Most similar according to linguistic criteria		Most similar according to the method		
dar miedo	**asustar**	**espantar**	**gustar**	**perdonar**
(1, 0.946) gustar	(1, 0.960) estremecer	(1, 0.837) espanto	(1, 0.617) me	(1, 0.857) perdón
(2, 0.876) perdonar	(2, 0.919) exclamar	(2, 0.761) aterrar	(2, 0.548) le	(2, 0.725) disculpar
(3, 0.872) agradar	(3, 0.784) temblar	(3, 0.756) asombrar	(3, 0.503) preguntar	(3, 0.694) reponer
(45, 0.689) asustar	(16, 0.730) espantar	(6, 0.730) asustar	(78, 0.339) asustar	(21, 0.419) asustar
(77, 0.655) espantar	(981, 0.339) gustar	(2499, 0.260) gustar	(400, 0.260) espantar	(303, 0.393) espantar
	(430, 0.419) perdonar	(562, 0.393) perdonar	(98, 0.328) perdonar	(726, 0.328) gustar

Table 4.10: Ranking and similarity measure of some verbs related to the LVC *dar miedo*.

5. Conclusion

This dissertation has been motivated by the aim of making progress in NLE research, specifically in the identification of the linguistic units of language postulated from disciplines that are related to the study of language. These units are linguistic constructions of different complexity and abstractness.

In order to develop a computational method for the automatic identification of these units, my first step has been to carry out research in cognitive neuroscience, developmental psychology and linguistics. Thanks to this research, I have been able to present an integrated framework that allows me to understand language from different perspectives.

Since linguistic constructions provide language its characteristic formulaicity, NLE needs to identify them to enrich its models and applications. The theoretical proposals do not include a collection of the constructions that have been conventionalized in our communication. So researchers in the field of NLE try to automatically detect them taking their properties into account. Due to the heterogeneous nature of linguistic constructions, many different approaches exist, usually focusing on specific types of constructions. My second step has been the creation of a state of the art of recent proposals, providing a classification based on how the problem is conceived and approached.

Finally, I have done experimental research. I have presented my own method for the identification of a specific type of construction, the LVC. I have chosen to use VSMs because they are rich linguistic models that take into account distributional information. VSMs have been traditionally used to model single words, so my approach also has made progress in extending the modelled object to that of a linguistic construction.

However, when applying the similarity measure between constructions and words, the results are not as good as expected. This fact does not imply that research in this direction is not feasible, but probably that there are problems that I did not anticipate, and also that more research may be needed in order to improve the representations.

To begin with, I have to investigate whether the unexpected results are due to data problems or to the model itself. It is not unreasonable to suspect that there might be a lack of data, since the number of occurrences of LVCs is less than that of single words, and it may be insufficient to have a proper representation.

As for the model itself, another possibility is to experiment with retrieving the contexts of occurrence in other ways, taking other events into account. A further step in this project could be adapting the parser so that dependency relations can be used in a way such that the results of the VSM are not biased. The same idea can be applied in the other way around, adapting the VSM to make a different use of the information provided by the parser (for instance, detecting that in *dar una orden de ataque* and *ordenar atacar*, *de atacar* and *atacar* should be treated as having the same dependency relation with the construction and the verb over them, respectively).

Furthermore, researchers working with VSMs are concerned with the problem of com-

positionality: once words have a proper representation in a VSM, how should they be combined? This problem is closely related to that of representing linguistic constructions, since constructions typically include several linguistic elements. I think one possible future line of research is to study to what extent these problems are interrelated and how do linguistic constructions differ from the regular operation of compositionality.

Lastly, this problem could also be approached from a broader perspective: instead of searching for specific constructions, they could emerge while looking for a linguistic model that is based on units above the words level.

The research presented in this dissertation has provided me with invaluable experience. I have had the possibility to delve into ideas coming from different areas and also to relate them with research in NLE. The implementation of the computational method has taught me about the methodology of experimental research, and this final dissertation has allowed me to put all the research together in a coherent work that establishes the beginning of my career as a researcher.

Bibliography

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, and Josef van Genabith. Automatic extraction of arabic multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 18–26, Beijing, China, August 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W10/W10-??03`.

T. Baldwin, C. Bannard, T. Tanaka, and D. Widdows. An empirical model of multiword expression decomposability. pages 89–96, 2003.

M. Baroni and A. Lenci. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721, 2010.

Marco Baroni and Adam Kilgarriff. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, EACL '06, pages 87–90, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1608974.1608976`.

G. Borensztajn, W. Zuidema, and R. Bod. Children's grammars grow more abstract with age—evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science*, 1(1):175–188, 2009.

J. Bybee. *Language, usage and cognition*. Cambridge University Press, 2010.

Joan Bybee and Paul Hopper. Introduction to frequency and the emergence of linguistic structure. In Joan Bybee and Paul Hopper, editors, *Frequency and the emergence of linguistic structure*, chapter 1, pages 1–24. John Benjamins Publishing Company, 2001.

Tanmoy Chakraborty, Dipankar Das, and Sivaji Bandyopadhyay. Semantic clustering: an attempt to identify multiword expressions in bengali. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 8–13, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-0803`.

N. Chomsky. *Aspects of the Theory of Syntax*, volume 119. MIT Press (MA), 1965.

A.S. Clark and S. Lappin. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley Online Library, 2011.

Matthieu Constant and Anthony Sigogne. Mwu-aware part-of-speech tagging with a crf model and lexical resources. In *Proceedings of the Workshop on Multiword*

Expressions: from Parsing and Generation to the Real World, pages 49–56, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W11-0809.

F.H. Crick. Thinking about the brain. *Scientific American*, 1979.

W. Croft. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press, USA, 2001.

Berthold Crysmann. A machine learning approach to relational noun mining in german. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 65–73, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W11-0811.

J.F. da Silva and G.P. Lopes. A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Sixth Meeting on Mathematics of Language*, pages 369–381, 1999.

Vitor De Araujo, Carlos Ramisch, and Aline Villavicencio. Fast and flexible mwe candidate generation with the mwetoolkit. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 134–136, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W11-0822.

G. Dias, S. Guilloré, and J.G.P. Lopes. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. *Traitement Automatique des Langues Naturelles, Institut d'Etudes Scientifiques, Cargèse, France*, pages 333–339, 1999.

Holger Diessel. *The acquisition of complex sentences*, volume 105, chapter A dynamic network model of grammatical constructions, pages 13–40. Cambridge Univ Press, 2004.

J. Duan, R. Lu, W. Wu, Y. Hu, and Y. Tian. A bio-inspired approach for multi-word expression extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 176–182. Association for Computational Linguistics, 2006.

Katrin Erk. What is word meaning, really? (and how can distributional models help us describe it?). In *Proceedings of the workshop on GEometrical Models of Natural Language Semantics (GEMS)*, 2010.

S. Evert and B. Krenn. Methods for the qualitative evaluation of lexical association measures. pages 188–195, 2001.

C.J. Fillmore. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32, 1976.

K. Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

V. Gallese and G. Lakoff. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *The Multiple Functions of Sensory-Motor Representations*, 22 (3/4):455, 2005.

A.E. Goldberg. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, 1995.

Stefan Gries and Anatol Stefanowitsch. *Corpus Linguistics: An International Handbook*, volume 2, chapter 43. Corpora and grammar, pages 933–951. Walter de Gruyter, 2009.

Z.S. Harris. Distributional structure. *Word*, 10:146–162, 1954.

Jeff Hawkins. *On Intelligence*. Holt Paperbacks, 2005.

L. Ingram and J.R. Curran. Distributional similarity of multi-word expressions. In *Proceedings of the Australasian Language Technology Workshop*, pages 146–148, 2007.

D. Jurafsky, J.H. Martin, A. Kehler, K. Vander Linden, and N. Ward. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, volume 163. MIT Press, 2000.

Graham Katz and Eugenie Giesbrecht. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia, July 2006. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W06/W06-1203`.

P. Kay and C.J. Fillmore. Grammatical constructions and linguistic generalizations: the what's x doing y? construction. *Language*, pages 1–33, 1999.

Nidhi Kulkarni and Mark Finlayson. jmwe: A java toolkit for detecting multi-word expressions. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 122–124, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-0818`.

G. Lakoff. *Women, fire, and dangerous things: What categories reveal about the mind*. Cambridge Univ Press, 1987.

G. Lakoff and M. Johnson. *Metaphors we live by*, volume 111. Chicago London, 1980.

R.W. Langacker. *Foundations of cognitive grammar: Theoretical Prerequisites*, volume 1. Stanford Univ Pr, 1987.

S. Martens and V. Vandeghinste. An efficient, generic approach to extracting multi-word expressions from dependency trees. In *CoLing Workshop: Multiword Expressions: From Theory to Applications (MWE 2010)*, 2010.

Scott Martens. Varro: an algorithm and toolkit for regular structure discovery in treebanks. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 810–818, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1944566.1944659`.

V. Mountcastle. An organizing principle for cerebral function: The unit model and the distributed system. 1978.

T.I. Nygren and Y. Wu. Language acquisition, emergentism, and the brain-changing norms of unilateral interventionism. *TCNJ Journal of Student Scholarship*, XIII, 2011.

S. Padó and M. Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

Alexander Panchenko. Comparison of the baseline knowledge-, corpus-, and web-based similarity measures for semantic relations extraction. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 11–21, Edinburgh, UK, July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-2502`.

P. Pecina. A machine learning approach to multiword expression extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 54–61. Citeseer, 2008.

Ted Pedersen, Satanjeev Banerjee, Bridget McInnes, Saiyam Kohli, Mahesh Joshi, and Ying Liu. The ngram statistics package (text::nsp) : A flexible tool for identifying ngrams, collocations, and word associations. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 131–133, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-0821`.

C. Ramisch, P. Schreiner, M. Idiart, and A. Villavicencio. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 50–53, 2008.

G. Rizzolatti and M.A. Arbib. Language within our grasp. *Trends in neurosciences*, 21 (5):188–194, 1998.

G. Rizzolatti and L. Craighero. The mirror-neuron system. *Annu. Rev. Neurosci.*, 27: 169–192, 2004.

G. Rizzolatti, L. Fogassi, V. Gallese, et al. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2(9):661–669, 2001.

E. Rosch. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General; Journal of Experimental Psychology: General*, 104(3):192, 1975.

I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. Multiword expressions: A pain in the neck for nlp. *Computational Linguistics and Intelligent Text Processing*, pages 189–206, 2002.

M. Sahlgren. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Stockholm University, 2006.

Magali Sanches Duran, Carlos Ramisch, Sandra Maria Aluísio, and Aline Villavicencio. Identifying and analyzing brazilian portuguese complex predicates. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 74–82, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-0812`.

F. Sangati, W. Zuidema, and R. Bod. Efficiently extract recurring tree fragments from large treebanks. 2010.

R. Scha. Taaltheorie en taaltechnologie; competence en performance. *Computertoepassingen in de Neerlandistiek, Almere: Landelijke Vereniging van Neerlandici (LVVN-jaarboek)*, 1990.

H. Schmid. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer, 1995.

P. Schone and D. Jurafsky. Is knowledge-free induction of multiword unit dictionary headwords a solved problem. In *Proc. of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, pages 100–108, 2001.

D.I. Slobin. The many ways to search for a frog. *Relating Events in Narrative. Vol*, 2: 219–257, 2004.

Anatol Stefanowitsch and Stefan Th. Gries. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2):209–243, 2003.

L. Talmy. Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description*, 3:57–149, 1985.

M. Tomasello. First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11(1-2):61–82, 2001.

M. Tomasello. *Constructing a language: A usage-based theory of language acquisition*. Harvard Univ Pr, 2003.

M. Tomasello. *Origins of human communication*, volume 2008. The MIT Press, 2008.

Yuancheng Tu and Dan Roth. Learning english light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 31–39, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-0807`.

P.D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

T. Van de Cruys and B.V. Moirón. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 25–32. Association for Computational Linguistics, 2007.

Patrick Watrin and Thomas François. An n-gram frequency database reference to handle mwe extraction in nlp applications. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 83–91, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-0813`.

David Wible and Nai-Lung Tsao. Stringnet as a computational resource for discovering and investigating linguistic constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 25–31, Los Angeles, California, June 2010. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W10-0804`.

Shuly Wintner. What science underlies natural language engineering? *Comput. Linguist.*, 35(4):641–644, December 2009. ISSN 0891-2017. doi: 10.1162/coli.2009.35.4.35409. URL `http://dx.doi.org/10.1162/coli.2009.35.4.35409`.

A. Wray and M.R. Perkins. The functions of formulaic language: an integrated model. *Language and communication*, 20(1):1–28, 2000.

Ying Xu, Randy Goebel, Christoph Ringlstetter, and Grzegorz Kondrak. Application of the tightness continuum measure to chinese information retrieval. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 54–62, Beijing, China, August 2010. Association for Computational Linguistics.

W. Zuidema. What are the productive units of natural language grammar?: a dop approach to the automatic identification of constructions. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 29–36. Association for Computational Linguistics, 2006.

W. Zuidema. Parsimonious data-oriented parsing. pages 551–560, 2007.