8

Morphosyntactic Linguistic Wavelets for Knowledge Management

Daniela López De Luise
Universidad de Palermo
Argentina

1. Introduction

Morphosyntactics studies grammatical categories and linguistic units that have both morphological and syntactic properties. In its proscriptive form, morphosyntactics describes the set of rules that govern linguistic units whose properties are definable by both morphological and syntactic paradigms.

Thus, morphosyntactics establishes a commons framework for oral and written language that guides the process of externally encoding ideas produced in the mind. Speech is an important vehicle for exchanging thoughts, and phonetics also has a significant influence on oral communication. Hearing deficiency causes a leveling and distortion of phonetic processes and hinders morphosyntactic development, particularly when present during the second and third years of life (Kampen, 2005).

Fundamental semantic and ontologic elements of speech become apparent though word usage. For example, the distance between successive occurrences of a word has a distinctive Poisson distribution that is well characterized by a stretched exponential scaling (Altmann, 2004). The variance in this analysis depends strongly on semantic type, a measure of the abstractness of each word, and only weakly on frequency.

Distribution characteristics are related to the semantics and functions of words. The use of words provides a uniquely precise and powerful lens into human thought and activity (Altmann, 2004). As a consequence, word usage is likely to affect other manifestations of collective human dynamics.

1.1 Words may follow Zipf's empirical law

Zipf's empirical law was formulated using mathematical statistics. It refers to the fact that many types of data studied in the physical and social sciences can be approximated with a Zipfian distribution, one of a family of related discrete power law probability distributions (Figure 1)[1] (Wolfram, 2011).

[1] In the English language, the probability of encountering the r^{th} most common word is given roughly by $P(r)=0.1/r$ ($r>1000$).

There is no theoretical proof that Zipf's law applies to most languages (Brillouin, 2004), but Wentian Li (Li, 1992) demonstrated empirical evidence supporting the validity of Zipf's law in the domain of language. Li generated a document by choosing each character at random from a uniform distribution including letters and the space character. Its words follow the general trend of Zipf's law. Some experts explain this linguistic phenomenon as a natural conservation of effort in which speakers and hearers minimize the work needed to reach understanding, resulting in an approximately equal distribution of effort consistent with the observed Zipf distribution (Ferrer, 2003).

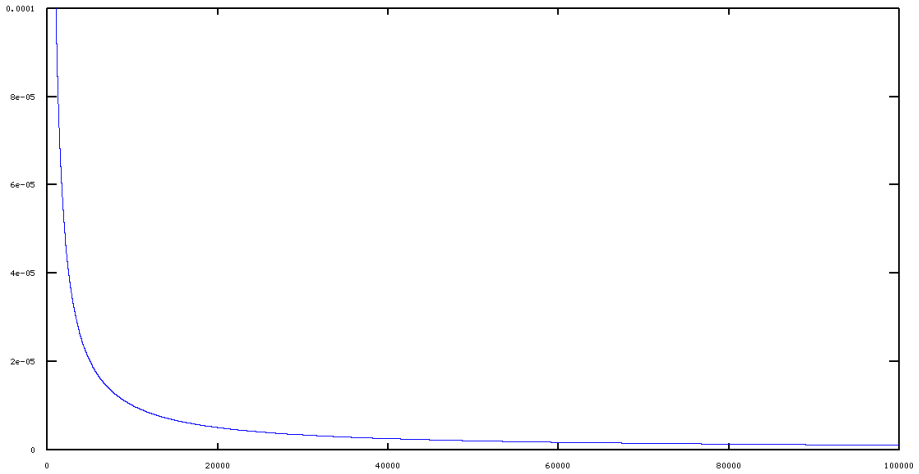


Fig. 1. Zipf's law for English

Whatever the underlying cause of this behavior, word distribution has established correspondences between social activities and natural and biological phenomena. As language is a natural instrument for representation and communication (Altmann, 2004), it becomes a particularly interesting and promising domain for exploration and indirect analysis of social activity, and it offers a way to understand how humans perform conceptualization. Word meaning is directly related to its distribution and location in context. A word's position is also related to its thematic importance and its usefulness as a keyword (López De Luise, 2008b, 2008c). This kind of information (recurrence, distribution and position) is strongly correlated with morphosyntactic analysis and strongly supports "views of human conceptual structure" in which all concepts, no matter how abstract, directly or indirectly engage contextually specific experience tracing language in the ever larger digital databases of human communications can be a most promising tool for tracing human and social dynamics". Thus, morphosyntactic analysis offers a new and promising tool for the study of dynamic social interaction. (Altmann, 2004).

1.2 Why morphosyntactic wavelets?

The evidence that wavelets offer the best description of such morphosyntactic decomposition is revealed by comparing the details of both traditional and morphosyntactical analyses.

ID	**Topic**	**Traditional wavelet**	**MLW**
1	application	just for signals[I]	any text[II]
2	type of transformation	mathematical	heuristic/statistical
3	goal	highlight, reinforce and obtain further information that is not readily available in the raw signal	extraction of information that highlights and reinforce knowledge that is not readily available in the raw text
4	time-domain signals	measured as a function of time. They have not undergone any transformation	they are analogous to the knowledge structure model (Hisgen, 2010). The sequence of the sentences is essential for contextualizing spoken/written words
5	frequency-domain signals	processed to transform then into a useful representation	are the E_{ci}, that represent sentence content and retain its main features
6	unit	Frequency: the number of the oscillations per seconds in a signal, measured in Hertz (Hz, cycles per second)	E_{ci} symbolizes morphosyntactic representations of sentences
7	domain	any type of data, even with sharp discontinuities[III]	any text
8	type of information	can represent signal in both the frequency and time domains[III]	also represents the time and frequency dimensions[IV]
9	scaling role	important. Can process at different scales and resolutions	represents knowledge at different levels of abstraction and detail
10	data decomposition result	decompose data x(t) into a two-dimensional function of time and frequency	decompose data into E_{ci} (representation of concrete/specific knowledge) and E_{ce} (abstract knowledge)[V]
11	data decomposition procedure	decompose x(t) using a "mother" wavelet W(x)	decompose using morphosyntactic rules and "mother sequence" of filters

I. Detectable physical quantity or impulse by which information may be sent
II. Although this theory is explained in general, it has only been proved in Spanish
III. This is an advantage over the FFT alternative
IV. This is true within the MLW context, given the statements in rows 4 and 5
V. The knowledge derived from the filtering processing is called E_{ce} in the MLW context

Table 1. Traditional wavelets versus MLW

Figure 2 shows a graphical comparison between a signal and its FFT. Figure 3 is a linguistic version: E_{ci} and ER. The graphics in Figure 2 represent the original signal (time-domain) and the resulting FFT decomposition (Lahm, 2002). The images in Figure 3 represent a translated original Spanish text (content from wikipedia.org, topic Topacio) transformed into an E_{ci} (López De Luise, 2007) that models dialog knowledge. (Hisgen, 2010) Statistical modeling of

knowledge is beyond the scope of this chapter, but additional information is available in (López De Luise, 2005, 2008, 2008b, 2008c, 2007b, 2007c).

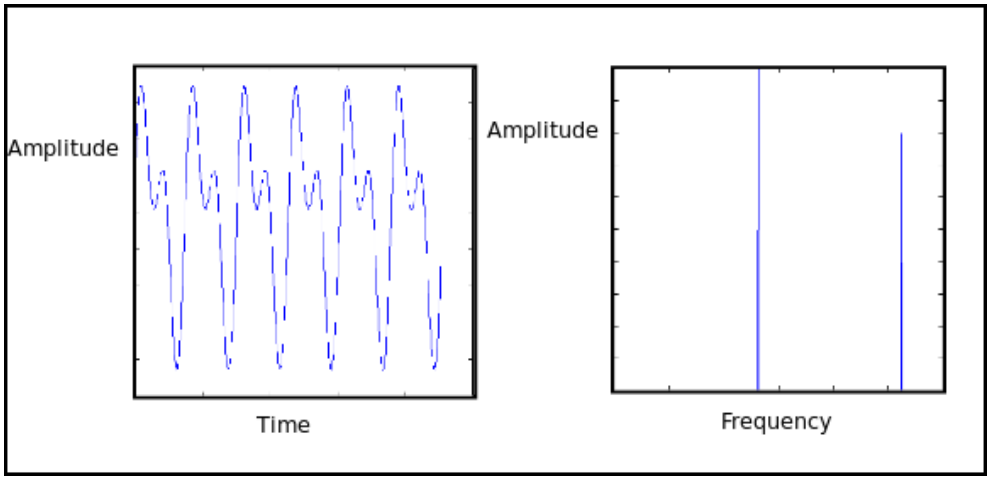


Fig. 2. Signal and frequency decomposition

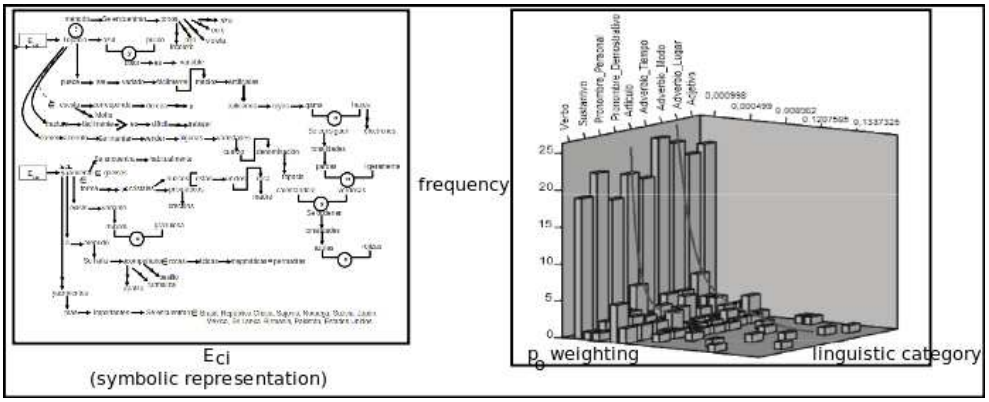


Fig. 3. Original text and knowledge structure model

Figure 4 shows a sample wavelet decomposition. It is a signature decomposition using a Daubechies wavelet, a wavelet specially suited for this type of image. Figure 5 shows a MLW decomposition of a generic text. There, C_i, and $C_{j,k}$ stand for abstract knowledge and F_m represents filters. This Figure will be described further in the final section.

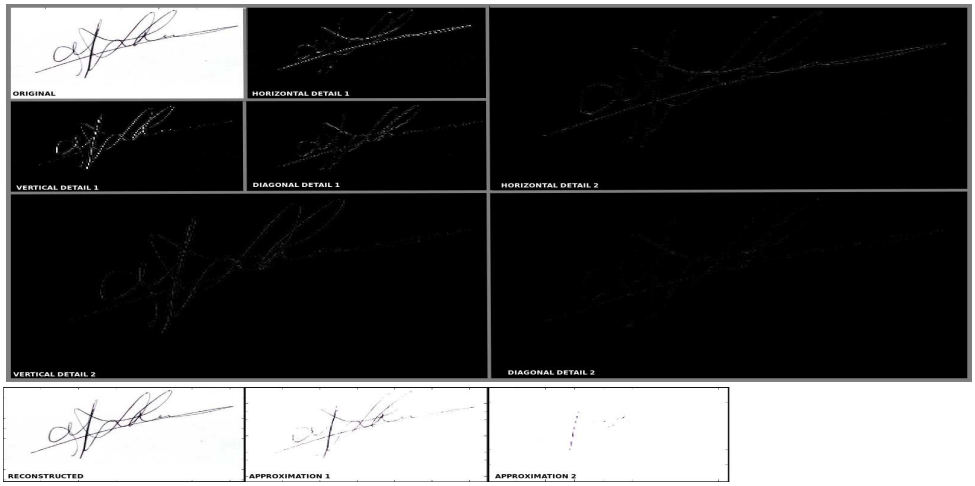


Fig. 4. Traditional wavelet decomposition

2. Technical overview

2.1 Wavelets

Wavelets are mathematical tools that are used to decompose/transform data into different components (coefficients) that describe different levels of detail (Lahm, 2002). Thus, they can extract the main features of a signal while simultaneously and independently analyzing details.

These tools have been applied to several problems, including the challenges of linguistic information retrieval. For example, wavelets have been used to build a Fuzzy Wavelet Neural Network (FWNN) for decision making over multiple criteria (Chen, 2008). In that analysis, custom built linguistic labels were used to represent information about events and situations and were processed with the FWNN.

Wavelets are sometimes used to replace linguistic analysis. For example, Tolba (Tolba, 2005) used consonant and vowel segmentation to develop automatic speech recognition for Arabic speech without linguistic information. Segmentation was performed with a combination of wavelet transformation and spectral analysis.

Hui and Wanglu combined the Linguistic Cloud Model (LCM) with wavelets to produce Advanced Synthetic Aperture Radar (ASAR) image target detection (Hui, 2008). This approach first solves image segmentation, avoids noise and recovers errors. Then, it uses LCM to solve the uncertainty of pixels. Representation using LCM bridges the gap between qualitative knowledge and quantitative knowledge, and it is thus used to map linguistic terms with contextually specific meaning to numeric processing.

2.2 Comparison between MLW and traditional wavelets

To demonstrate the concept of MLW and its relationship to its traditional counterpart, this table summarizes the main characteristics that unite or distinguish them:

characteristic	morphosyntactic	traditional wavelet
goal	content description and classification with granularities	scaled decomposition
scaling	concept abstraction and ontology classification	reduced and representative signal
Uses	Extract the main concept of a signal	extract the main features of a signal
	summarization	compression
	manage spelling and some grammatical errors	De-noising
	Complement knowledge	Reconstruct portions of a corrupted signal
Types of wavelets	Depends on the specific sequence of filters	depends on the functions used as mother function
	auto-fitting	Must be manually detected according to results

Table 2. Characteristics of traditional wavelets and MLW

2.3 Linguistic cloud model and MLW

LCM models linguistic knowledge (Li, 2000) using a set of predefined, customized fuzzy linguistic variables. These variables are generated in accordance with two rules:

1. *The atom generation rule* specifies the manner in which a linguistic "atom" may be generated. An atom is a variable that cannot be sliced into smaller parts.
2. *The semantic rule* specifies the procedure by which composite linguistic terms are computed from linguistic atoms. In addition, there are connecting operators ("and" "or", etc.), modifiers ("very" "quite", etc.) and negatives that are treated as soft operators that modify an operand's (atom's) meaning to produce linguistic "terms".

The MSW and the LCM share a common goal. However, the MSW replaces the manual procedure used to obtain linguistic atoms with automated processing that determines an atom's linguistic category (e.g., noun or verb) (López De Luise, 2007d, 2008c). The result is not an atom or a term but is a structure named E_{ci} (an acronym from the Spanish, Estructura de Composición Interna). The E_{ci} is used to model the morphosyntactic configuration within sentences (López De Luise, 2007; Hisgen, 2010). Thus, the core processing is based on E_{ci} structures instead of linguistic variables. An E_{ci} is a plastic representation that can evolve to reflect more detailed information regarding the represented portion of text. While atoms cannot be sliced, any E_{ci} can be partitioned as required during the learning process. Further differences between the LCM and the MSW are shown in Table 3.

2.4 Morphosyntactics as a goal

Most morphological and syntactical processing is intended for information retrieval, while alignment supports automatic translation. Those approaches are mainly descriptive and are defined by cross-classifying different varieties of features (Harley, 1994) such as number and person. When morphological operations are an autonomous subpart of the derivation, they acquire a status beyond descriptive convenience. They become linguistic primitives, manipulated by the rules of word formation.

	LCM	**MSW**
Basis	Atom	E_{ci}
Characteristics of the unit of processing	Cannot be sliced	Can be sliced
	Represents the meaning of a word	Represents morphosyntactic characteristics of a sentence
	Fixed	Can evolve
	Hand made	Automatically extracted
Rules	Atom generation rule	Morphosyntactic rules
	Semantic rules	Clustering filters
Semantic	Directly manipulated by term definitions	Indirectly manipulated by clustering filtering and morphosyntactic context

Table 3. Comparison Between the LCM and the MSW

In many approaches, they are manipulated as an undifferentiated bundle divided only into nominal and verbal atoms. The following section describes elements of the morphosyntactic approach.

2.4.1 Detecting language tendencies

Language tendencies denote cultural characteristics, which are represented as dialects and regional practices. Noyer (Noyer, 1992) described a hierarchical tree organization defined by applying manually predefined morphological feature filters to manage morphological contrasts. They used this organization as an indicator of linguistic tendencies in language usage. Extensions of this approach attempt to derive the geometry of morphological features[2] (Harley, 1994, 1998), with the goal of classifying features into subgroups based on an universal geometry while accounting for universals in feature distribution and realization. In MLW, the structure of the information is organized in a general oriented graph (E_{ci} structure) for only the smallest unit of processing (a sentence), and a hierarchy is defined by a chained sequence of clustering filters (Hisgen, 2010). Language tendencies are therefore visible in the configuration of a current graph.

2.4.2 Sentence generation

Morphosyntax has also been used to implement a language sentence generator. In an earlier study (Martínez López, 2007), Spanish adverbial phrases were analyzed to extract the reusable structures and discard the remainder, with the goal of using the reusable subset to generate new phrases. Interestingly, the shortest, simplest structures presented the most productive patterns and represented 45% of the corpus.

Another study (López De Luise, 2007) suggested translating Spanish text, represented by sets of E_{ci}, into a graphic representing the main structure of the content. This structure was

[2] This is a well-known method that is used to model phonological features (Clements, 1985) (Sagey, 1986)

tested with 44 subjects (López De Luise, 2005). The results showed that this treatment, even without directly managing semantics, could communicate the original content. Volunteers were able to reconstruct the original text content successfully in 100% of the cases. As MLW is based on E_{ci} structure, it follows that it:

- represents keywording well.
- performs well independent of an individual's knowledge on a specific subject.
- performs well independent of an individual's knowledge of informatics.

2.4.3 Language comprehension detection

As language is an expression of mind and its processes, it becomes also the expression of meaning (or lack of meaning) in general. This fact is also true when the subject is the language itself. A recent study focused on the most frequently recurrent morphosyntactic uses in a group of students who study Spanish as a foreign language (González Negrón, 2011) revealed a peculiar distribution of nouns and personal pronouns. These parts of speech were present at a higher frequency than in the speech of native speakers, probably to guarantee the reader comprehension of the text. Other findings included preposition repetition and a significant number of misplaced prepositions. Thus, morphosyntactic statistics detect deficient language understanding. A similar study was performed in (Konopka, 2008) with Mexican subjects living in Chicago (USA). In the case of MLW, the E_{ci} and E_{ce} structures will shape irregular language usage and make detection of incorrect language practices easy.

2.4.4 Semantics detection

Morphosyntactics can be used to detect certain types of semantics in a text. An analysis of vowel formant structure and vowel space dispersion revealed overall spectral reduction for certain talkers. These findings suggest an interaction between semantic and indexing factors in vowel reduction processes (Cloppera, 2008).

Two morphosyntactic experimental studies of numeral quantifiers in English (*more than* k, *at least* k, *at most* k, and *fewer than* k) (Koster-Moeller, 2008) showed that Generalized Quantifier Theory (GQT)[3] must be extant to manage morphosyntactic differences between denotationally equivalent quantifiers. The formal semantic is focused on the correct set of entailment patterns of expressions but is not concerned with deep comprehension or real-time verification. However, certain systematic distinctions occur during real-time comprehension. The degree of compromise implicit in a semantic theory depends on the types of semantic primitives it assumes, and this also influences its ability to treat these phenomena. In (López De Luise, 2008b), sentences were processed to automatically obtain specific semantic interpretations. The shape of the statistics performed over the E_{ci}'s internal weighting value (named p_o) is strongly biased by the semantics behind sentence content.

[3] Generalized Quantifier Theory is a logical semantic theory that studies the interpretation of noun phrases and determinants. The formal theory of generalized quantifiers already existed as a part of mathematical logic (Mostowski, 1957), and it was implicit in Montague Grammar (Montague, 1974). It has been fully developed by Barwise & Cooper (1981) and Keenan & Stavi (Barwise, 1981) as a framework for investigating universal constraints on quantification and inferential patterns concerning quantifiers.

2.4.5 Improvement of translation quality/performance

Automatic translation has an important evolution. Translation quality depends on proper pairing or alignment of sources and on appropriate targeting of languages. This sensible processing be improved using morphosyntactic tools.

Hwang used morphosyntactics intensively for three kinds of language (Hwang, 2005). The pairs were matched on the basis of morphosyntactical similarities or differences. They investigated the effects of morphosyntactical information such as base form, part-of-speech, and the relative positional information of a word in a statistical machine translation framework. They built word and class-based language models by manipulating morphological and relative positional information.

They used the language pairs Japanese-Korean (languages with same word order and high inflection/agglutination[4]), English-Korean (a highly inflecting and agglutinating language with partial free word order and an inflecting language with rigid word order), and Chinese-Korean, (a highly inflecting and agglutinating language with partially free word order and a non-inflectional language with rigid word order).

According to the language pairing and the direction of translation, different combinations of morphosyntactic information most strongly improve translation quality. In all cases, however, using morphosyntactic information in the target language optimized translation efficacy. Language models based on morphosyntactic information effectively improved performance. E_{ci} is an important part of the MLW, and it has inbuilt morphophonemic descriptors that contribute significantly to this task.

2.4.6 Speech recognition

Speech recognition requires real-time speech detection. This is problematic when modeling languages that are highly inflectional but can be achieved by decomposing words into stems and endings and storing these word subunits (morphemes) separately in the vocabulary. An enhanced morpheme-based language model has been designed for the inflectional Dravidian language Tamil (Saraswathi, 2007). This enhanced, morpheme-based language model was trained on the decomposed corpus. The results were compared with word-based bi-gram and trigram language models, a distance-based language model, a dependency-based language model and a class-based language model. The proposed model improves the performance of the Tamil speech recognition system relative to the word-based language models. The MLW approach is based on a similar decomposition into stems and endings, but it includes additional morphosyntactical features that are processed with the same importance as full words (for more information, see the last sections). Thus, we expect that this approach will be suitable for processing highly inflectional languages.

[4] This term was introduced by Wilhelm von Humboldt in 1836 to classify languages from a morphological point of view. An agglutinative language is a language that uses agglutination extensively: most words are formed by joining morphemes together. A morpheme is the smallest component of a word or other linguistic unit that has semantic meaning.

3. Morphosyntactic linguistic wavelet approach

3.1 A sequential approach to wavelets

Because language is complex, soft decomposition into a set of base functions (as in traditional wavelets) is a multi-step process with several components.

Developing numeric wavelets usually includes the following steps:

1. Take the original signal sample
2. Apply filtering (decomposition using the mother wavelet)
3. Analyze coefficients defined by the basis function
4. If the granularity and details are inadequate for the current problem, repeat from step 2
5. Take the resulting coefficients as a current representation of the signal

Language requires additional steps, which are described in more detail in the following section. In brief, these steps are:

1. Take the original text sample
2. Compress and translate text into an oriented graph (E_{ci}) preserving most morphosyntactic properties
3. Apply filtering using the most suitable approach
4. If abstraction granularity and details are insufficient for the current problem
   4.1 Insert a new filter, E_{ce}, in the knowledge organization
   4.2 Repeat from step 3
5. Take the resulting sequence of filtering as a current representation of the knowledge about and ontology of the text
6. Take the resulting E_{ci} as the internal representation of the new text event

A short description of the MLW steps is presented below, with an example in the Use case.

3.2 Details of the MLW process

Further details of the MLW process are provided in this section, with the considerations relevant to each step included.

3.2.1 Take the original text sample

Text can be extracted from Spanish dialogs, Web pages, documents, speech transcriptions, and other documents. The case study in the section 4 uses dialogs, transcriptions, and other documents. Several references mentioned in this chapter were based on Web pages.

3.2.2 Compress and translate text into an oriented graph (Called E_{ci}) preserving most morphosyntactic properties

Original text is processed using predefined and static tables. The main components of this step are as follows:

- Filter useless morphemes[5] using reference tables.

[5] Syntagm (linguistics) is any sequenced combination of morphologic elements that is considered a unit, has stability and is commonly accepted by native speakers.

- Extract morphosyntactic descriptors (numeric values automatically extracted, such as the number of vowels) for each word processed. Words were previously represented by Porter's Stemming, but this tool does not have enough classification power for use as a sole instrument. Morphosyntactic descriptors are required to process text with sufficient confidence levels (López De Luise, 2007d).
- Collapse syntagmas into a condensed internal representation (usually, selected morphemes[6]). The resulting representation is called EBH (Estructura Básica Homogénea, uniform basic structure). EBHs are linked with specific connectors.
- Calculate and set the morphosyntactic weighting p_o for E_{ci}.

More details of each of these steps are outside of the scope of this chapter (but see (López De Luise, 2008c) and (López De Luise, 2008)).

3.2.3 Apply filtering using the most suitable approach

Since knowledge management depends on previous language experiences, filtering is dynamic process that adapts itself to current cognitive capabilities. Furthermore, as shown in the Case Study section, filtering is a very sensitive step in the MLW transformation.

Filtering is a process composed of several filters. The current paper includes the following three clustering algorithms: Simple K-means, Farthest First and Expectation Maximization (Witten, 2005). They are applied sequentially for each new E_{ce}. When an E_{ce} is "mature", the filter no longer changes.

The distance used to evaluate clustering is based on the similarity between the descriptor values and the internal morphosyntactic metric, p_o, that weights EBH (representing morphemes). It has been shown that clusters generated with p_o represent consistent word agglomerations (López De Luise, 2008, 2008b). Although this chapter does not use fuzzy clustering algorithms, it is important to note that such filters require a specific adaptation for distance using the categorical metrics defined in (López De Luise, 2007e).

3.2.4 If "Abstraction" granularity and details are inadequate for the current problem

Granularity is determined by the ability to discriminate the topic and by the degree of detail required to represent the E_{ci}. In the MLW context it is the logic distance between the current E_{ci} and the E_{ce} partitions[7] (see Figure 5). This distance depends on the desired learning approach. In the example included herein (Section 4), it is the number of elements in the E_{ci} that fall within each E_{ce} partition. The distribution of EBHs determines whether a new E_{ce} is a necessary. When the EBHs are too irregular, a new E_{ce} is built per step 3.2.4.1. Otherwise the new E_{ci} is added to the partition that is the best match.

3.2.4.1 Insert a new filter, E_{ce}, in the knowledge organization

The current E_{ce} is cleaned so that it keeps all the E_{ci}s that best match its partitions, and a new E_{ce} that includes all the E_{ci}s that are not well represented is created and linked.

[6] A meaningful linguistic unit that cannot be divided into smaller meaningful parts.
[7] Partition in this context is a cluster obtained after the filtering process.

3.2.4.2 Repeat from step 3.2.3

3.2.5 Take the resulting sequence of filtering as a current representation of the knowledge about and ontology of the text

The learned E_{ci}'s ontology is distributed along the chain of E_{ce}s.

3.2.6 Take the resulting E_{ci} as the internal representation of the new text event

The specific acquired, concrete knowledge is now condensed in the E_{ci}. This provides a good representation of the original text and its keywording (López De Luise, 2005).

Real texts include contradictions and ambiguities. As previously shown (López De Luise, 2007b), they are processed and handled despite potentially inadequate contextual information. The algorithm does not include detailed clause analysis of or encode linguistic knowledge about the context because these components complicate the process and make it less automatic.

Furthermore, using the p_o metric can distinguish the following Writing Profiles: general document, Web forum, Web index and blogs. This metric is therefore independent of document size and mentioned text styles (López De Luise, 2007c). Consequently, it is useful to define the quality of the text that is being learned and to decide whether to accept it as a source of knowledge.

3.3 Gelernter's perspective on reasoning

Section 3.2.3. defines that the clustering algorithms must be used first hard clusterings and afterwards fuzzy. It is not a trivial restriction. Its goal is to organize learning across a range from specific concrete data to abstract and fuzzy information. The filters are therefore organized as a sequence from simple k-means clustering to fuzzy clustering. This approach is compatible with Gelernter's belief that thinking is not a static algorithm that applies to every situation. Thinking requires a set of diverse algorithms that are not limited to reasoning. Some of these algorithms are sharp and deep, allowing clear manipulation of concrete objects, but there are other algorithms with different properties.

David Gelernter Theory (Gelernter, 2010) states that thinking is not the same as reasoning. When your mind wanders, you are still thinking. Your mind is still at work. This free association is an important part of human thought. No computer will be able to think like a man unless it can perform free association.

People have three common misconceptions:

3.3.1 The belief that "thinking" is the same as "reasoning"

There are several activities in the mind that are not reasoning. The brain keeps working even when the mind is wandering.

3.3.2 The belief that reality and thoughts are different and separated things

Reality is conceptualized as external while the mental landscape created by thoughts is seen as internal and mental. According to Gelernter, both are essentially the same although the attentional focus varies.

3.3.3 The separation of the thinker and the thought

Thinking is not a PowerPoint presentation in which the thinker watches the stream of his thoughts. When a person is dreaming or hallucinating, the thinker and his thought-stream are not separate. They are blended together. The thinker inhabits his thoughts.

Gelernter describes thinking as a spectrum of many methods that alternate depending on the current attentional focus. When the focus is high, the method is analytic and sharp. When the brain is not sharply focused, emotions are more involved and objects become fuzzy. That description is analogous to the filtering restriction: define sharp clustering first and leave fuzzy clustering approaches for the final steps. As Gelernter writes, "No computer will be creative unless it can simulate all the nuances of human emotion."

4. Case study

This section presents a sample case to illustrate the MLW procedure. The database is a set of ten Web pages with the topic "Orchids". From more than 4200 original symbols and morphemes in the original pages, 3292 words were extracted; 67 of them were automatically selected for the example. This section shows the sequential MLW decomposition. Table 4 shows the filtering results for the first six E_{ci}s.

4.1 Build $E_{ci}1$

Because the algorithm has no initial information about the text, we start with a transition state and set the d parameter to 20%. This parameter assesses the difference in the number of elements between the most and least populated partitions.

4.2 Apply filters to $E_{ci}1$

The K-means clustering, in the following KM, is used as the first filter with settings N= 5 clusters, seed 10. Diff=16%<d. Keep KM as the filter.

4.3 Apply filters to $E_{ci}2$

Filter using KM with the same settings, and the current Diff=11%<d. Keep KM as the filter.

4.4 Apply filters to $E_{ci}3$

Filter using KM with the same settings, and the current Diff=10%<d. Keep KM as the filter and exit the transition state.

4.5 Apply filters to $E_{ci}4$

Filter using KM with d=10% for steady state. This process will indicate whether to change the filter or build a new E_{ce}. Clustering settings are the same, and the current Diff=20%>d. Change to Farthest First (FF) as the filter.

4.6 Apply filters to $E_{ci}5$

Filter using FF with clustering settings N=5 clusters, seed 1. The current Diff=45%>d. Change to Expectation Maximization (EM) as the next filter.

4.7 Apply filters to $E_{ci}6$

Filter using EM. The clustering settings are min stdDev :1.0E-6, num clusters: -1 (automatic), seed: 100. The current Diff=13%>d, Log likelihood: -18.04546. Split E_{ce} and filter the more cohesive[8] subset of E_{ci}s using EM. Log likelihood: -17.99898, diff=8.

Cluster	$E_{ci}1$	$E_{ci}2$	$E_{ci}3$	$E_{ci}4$	$E_{ci}5$	$E_{ci}6$	$E_{ci}6^*$
0	1 (17%)	3 (33%)	3 (20%)	3 (20%)	2 (6%)	13 (29%)	8 (18%)
1	1 (17%)	1 (11%)	5 (33%)	5 (33%)	4 (11%)	7 (16%)	8 (18%)
2	1 (17%)	2 (22%)	2 (13%)	2 (13%)	9 (26%)	11 (24%)	12 (26%)
3	2 (33%)	2 (22%)	3 (20%)	3 (20%)	18 (51%)	14 (31%)	8 (18%)
4	1 (17%)	1 (11%)	2 (13%)	2 (13%)	2 (6%)		9 (20%)
diff	16,00%	11,00%	10,00%	20,00%	45,00%	13,00%	8,00%

*This is the result of EM to define the splitting of $E_{ce}1$.

Table 4. Filtering results for each E_{ci}

4.8 Build $E_{ce}2$ to $E_{ci}6$

Keep all the individuals as $E_{ci}1$, and put in $E_{ce}2$ the individuals in cluster 1 (one of the three less cohesive, with lower p_o). This procedure is shown in Figure 6.

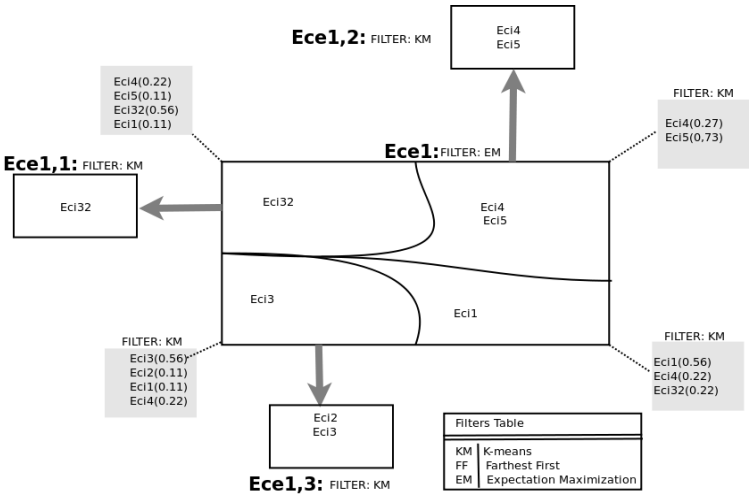


Fig. 6. $E_{ce}1$ after cleaning up the less cohesive E_{ci}s

[8] Cohesiveness is defined according to MLW as distance and sequence of filters. In this case it is implemented using EM forcing 5 clusters, and selecting the four clusters with more elements.

4.9 Apply filters to $E_{ci}7$

Detect the $E_{ce}1$ partition that best suits $E_{ci}7$ using cohesiveness criteria. The result shows that the partition that holds $E_{ci}6$ is the best. $E_{ci}7$ now hangs from this partition as indicated in Figure 7.

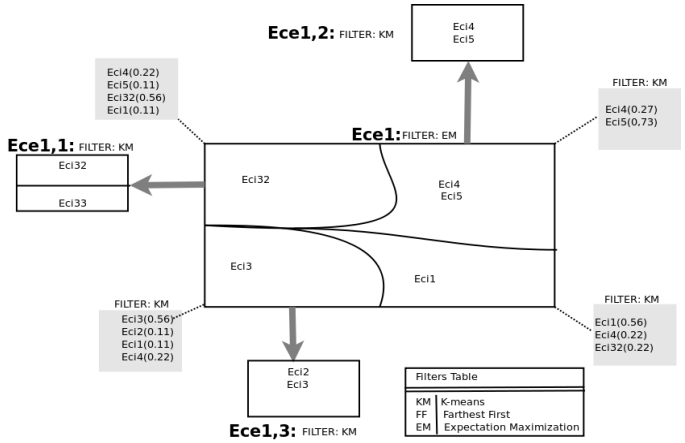


Fig. 7. $E_{ce}1,1$ hanging from $E_{ce}1$

4.10 Apply filters to $E_{ci}8$

Detect the $E_{ce}1$ partition that best suits $E_{ci}8$ using the same cohesiveness criteria. The partition that holds $E_{ci}5$ is the best. $E_{ce}1,1$ now contains $E_{ci}4$, $E_{ci}5$ and $E_{ci}6$. Filter $E_{ce}1,1$ using KM with clustering settings of N= 5 clusters, seed 10. The value of Diff=20%>d. Change to Farthest First (FF) as the next filter.

Now the E_{ce} sequence is as indicated in Figure 8.

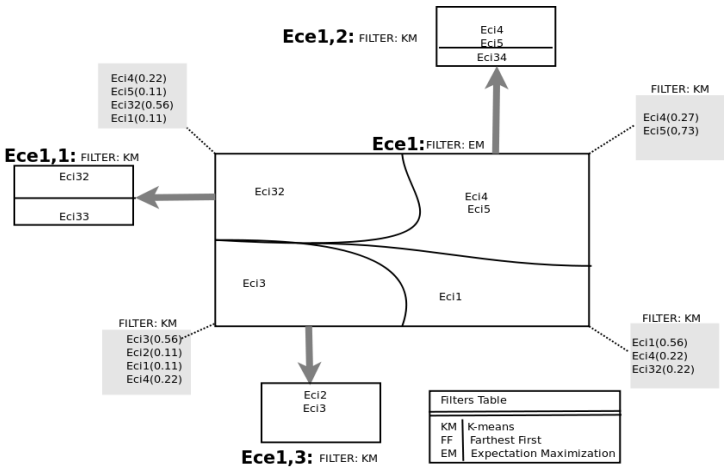


Fig. 8. $E_{ce}1$ and $E_{ce}1,1$ after learning $E_{ci}8$

4.11 The representation in MLW

We do not expect E_{ce} content to be understood from the human point of view, but it should be considered a tool to condense and potentially regenerate knowledge from textual sources. This is a first step in the study of this type of tool that uses mathematical and statistical extraction of knowledge to automatically decompose text and represent it in a self-organizational approach.

For instance, the following sentence from the dataset,

> "Dactylorhiza incarnata es orquídea de especies Europeas"
> (Dactylorhiza incarnata is an European orchid species)

corresponds to the EBH number 04, and can be found (after MLW) as the sequence $E_{ce}1$-$E_{ce}1,2$-$E_{ci}4$.

If there is an interest in understanding the topic, the main entry of the set of E_{ci}s in the cluster can be used as a brief description. To regenerate the concepts saved in the structure for human understanding, it is only necessary to use the symbolic representation of the E_{ci} (López De Luise, 2007).

5. Conclusion

MLW is a new approach that attempts to model natural language automatically, without the use of dictionaries, special languages, tagging, external information, adaptation for new changes in the languages, or other supports. It differs from traditional wavelets in that it depends on previous usage, but it does not require human activities to produce definitions or provide specific adaptations to regional settings. In addition, it compresses the original text into the final E_{ci}. However, the long-term results require further testing, both to further evaluate MLW and to evaluate the correspondence between human ontology and conceptualization and the E_{ce}s sequence .

This approach can be completed with the use of a p_o weighting to filter the results of any query or browsing activity according to quality and to detect additional source types automatically.

It will also be important to test the use of categorical metrics for fuzzy filters and to evaluate MLW with alternate distances, filter sequences and cohesiveness parameters.

6. References

(Altmann,2004) E.G. Altmann, J.B. Pierrehumbert & A.E. Motter. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE* 4(11): e7678. ISSN 1932-6203.

(Brillouin, 2004) L. Brillouin. La science et la théorie de l'information. Masson, Paris. *Open Library*. ISBN 10-2876470365.

(Chen, 2008) K. Chen, J. Li. Research on Fuzzy MCDM Method based on Wavelet Neural Network Model. *Information Sci. And Eng*. ISISE'08. 2008. ISBN: 978-0-7695-3494-7.

(Clements, 1985) G.N. Clements. The Geometry of Phonological Features. *Phonology Yearbook* 2. pp. 225 - 252. ISBN 9780521332323 . USA.

(Cloppera, 2008) C. G. Cloppera & J. B. Pierrehumbert. Effects of semantic predictability and regional dialect on vowel space reduction. *Journal of the Acoustical Society of America*, 124, 1682-1688. ISSN: 0001-4966. USA.

(Ferrer, 2003) R. Ferrer & R.V. Sole. Least effort and the origins of scaling in human language. *Proc. of the National Academy of Sciences* of the United States of America 100 (3): 788–791.*ISSN* 0027-8424. USA.

(Gelernter, 2010) D. Gelernter. Dream-logic, the internet and artificial thought. *EDGE*. Available in: www.edge.org

(González Negrón, 2011) N. González Negrón. Usos morfosintácticos en una muestra de exámenes de estudiantes que cursan el español como idioma extranjero. *ELENET*. N. 1. ISBN: 2-9524532-0-9. Spain.

(Harley, 1994) H. Harley. Hug a tree: deriving the morphosyntactic feature hierarchy. *MIT Working Papers in Linguistics* 21, 289-320. ISBN: 9780262561211. USA.

(Harley, 1998) H. Harley & E. Ritter. Meaning in Morphology: motivating a feature-geometric analysis of person and number. *Ms. University of Calgary & University of Pennsylvania*.

(Hisgen, 2010) D. Hisgen & D. López De Luise. Dialog Structure Automatic Modeling. *MICAI*. *ISBN* 978-3-642-16772-0. Mexico.

(Hui, 2008) H. Hui & P. Wanglu. ASAR Image target recognition based on the combined wavelet transformation. *ISPRS Congress*. Beijing, Proceedings of Commission VII. ISBN:0-7803-9051-2. South Korea.

(Hwang, 2005) Y. Hwang, T. Watanabe & Y. Sasaki. Empirical Study of Utilizing Morph-Syntactic Information in SMT. *2nd IJCNLP*. ISBN 3-540-29172-5. Korea.

(Kampen, 2005) J. Van Kampen. Morph-syntactic development and the effects on the lexicon (A comparison between normal hearing children and children with a temporary hearing deficiency. Poster. *ELA2005*. *ISBN* 9780387345871. France.

(Koster-Moeller, 2008) J. Koster-Moeller, J. Varvoutis & M. Hackl. Verification Procedures for Modified Numeral Quantifiers. *Proc. of the 27th WCCFL*. ISBN 978-1-57473-428-7. USA.

(Konopka, 2008) K. Konopka. Vowels in Contact: Mexican Heritage English in Chicago. Salsa XVI- *Texas Linguistic Forum*. 52: 94–103. *ISSN* 1615-3014. Germany.

(Lahm, 2002) Z. Lahm. Wavelets: A tutorial. University of Otago. In: Dep. Of Computer Science (2011). Available from www.cs.otago.ac.nz.

(Li, 1992) W. Li. Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. *IEEE Trans. on Information Theory*. 38 (6): 1842–1845. ISSN: 0018-9448. USA.

(Li, 2000) D. Li, K. Di, D. Li & X. Shi. Mining Association Rules With Linguistic Cloud Models. *Journal of Software*,Vol.11, No. 2, pp.143-158.

(López De Luise, 2005) D. López De Luise. A Morphosyntactical Complementary Structure for Searching and Browsing. *CISSE 2005*. Pp. 283 – 290.*ISBN* 1-4020-5262-6. USA.

(López De Luise, 2007) D. López De Luise. Una representación alternativa para textos (Alternate representation for [Spanish] texts). *J. Ciencia y Tecnología*. Vol 6. ISSN 1850 0870. Argentina.

(López De Luise, 2007b) D. López De Luise. Ambiguity and Contradiction From a Morpho-Syntactic Prototype Perspective. *CISSE*. Bridgeport. *ISBN* 978-1-4020-8740-0. USA.

(López De Luise, 2007c) D. López De Luise. A Metric for Automatic Word categorization. *SCSS*. Bridgeport. ISBN 978-1-4020-8740-0. USA.

(López De Luise, 2007d) D. López De Luise & J. Ale. Induction Trees for automatic Word Classification. *CACIC*.

(López De Luise, 2007e) D. López De Luise. Aplicación de Métricas Categóricas en Sistemas Difusos. *IEEE LATIN AMERICA*. ISSN: 1548-0992. Brasil.

(López De Luise, 2008) D. López De Luise, M. Soffer. Modelización automática de textos en castellano. *ANDESCON*. ISBN 978-603-45345-0-6. Peru.

(López De Luise, 2008b) D. López De Luise & M. Soffer. Automatic Text processing for Spanish Texts. *CERMA 2008*. ISBN: 978-0-7695-3320. Mexico.

(López De Luise, 2008c) D. López De Luise. Mejoras en la usabilidad de la Web a través de una estructura complementaria. PhD thesis. Universidad Nacional de La Plata. Argentine.

(Martínez López, 2007) J. A. Martínez López. Patrones e índice de frecuencia en algunas locuciones adverbiales. *Forma funcion*, Bogotá. v 20. pp 59-78. ISSN 0120-338X. Colombia.

(Montague, 1974) R. Montague. *Formal Philosophy. Yale University Press*. ISBN: 0300015275. USA.

(Barwise, 1981) J. Barwise & R. Cooper. Generalized Quantifiers and Natural Language. *Linguistics and Philosophy* 4, pp. 159-219. ISBN 978-94-007-2268-2. USA.

(Mostowski, 1957) A. Mostowski. A generalization of Quantifiers. *Fundamenta Mathematicae*. vol 44. pp. 12 - 36. ISSN : 0016-2736. Polish.

(Noyer, 1992) R. Noyer. Features, Positions and Affixes in Autonomous Morphological Structure. MIT PhD dissertation. Cambridge MITWPL. USA.

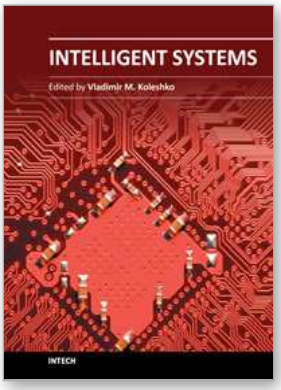
(Sagey, 1986) E. Sagey. The representation of Features and Relations in Non-Linear Phonology. PhD dissertation. MIT. MITWPL. USA.

(Saraswathi, 2007) S. Saraswathi & T.V. Geetha. Comparison of Performance of Enhanced Morpheme-Based Language Model with Different Word-based Language Models for Improving the Performance of Tamil Speech Recognition System. *ACM Trans. Asian Lang. Information*. V. 6, N. 3, Article 9. ISBN: 978-1-4503-0475-7. USA.

(Tolba, 2005) M. F. Tolba, T. Nazmy, A. A. Abdelhamid & M. E. Gadallah. A novel method for Arabic consonant/vowel segmentation using wavelet transform. *IJICIS*, Vol. 5, No. 1. ISBN: 978-960-474-064-2. USA.

(Witten, 2005) I.H. Witten & E. Frank. Data Mining - *Practical Machine Learning Tools And Techniques*, 2Nd Edition. Elsevier. *ISBN*: 978-0-12-374856-0. New Zeland.

(Wolfram,2011) Zipf's Law. (2011), Wolfram Research, Inc. In: *Wolfram MathWorld*, 2011, Available from http://mathworld.wolfram.com/ZipfsLaw.html



Intelligent Systems

Edited by Prof. Vladimir M. Koleshko

ISBN 978-953-51-0054-6
Hard cover, 366 pages
Publisher InTech
Published online 02, March, 2012
Published in print edition March, 2012

This book is dedicated to intelligent systems of broad-spectrum application, such as personal and social biosafety or use of intelligent sensory micro-nanosystems such as "e-nose", "e-tongue" and "e-eye". In addition to that, effective acquiring information, knowledge management and improved knowledge transfer in any media, as well as modeling its information content using meta-and hyper heuristics and semantic reasoning all benefit from the systems covered in this book. Intelligent systems can also be applied in education and generating the intelligent distributed eLearning architecture, as well as in a large number of technical fields, such as industrial design, manufacturing and utilization, e.g., in precision agriculture, cartography, electric power distribution systems, intelligent building management systems, drilling operations etc. Furthermore, decision making using fuzzy logic models, computational recognition of comprehension uncertainty and the joint synthesis of goals and means of intelligent behavior biosystems, as well as diagnostic and human support in the healthcare environment have also been made easier.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821