Developing a large scale FrameNet for Italian: the IFrameNet experience

Roberto Basili° **Silvia Brambilla**§ **Danilo Croce**° **Fabio Tamburini**§

° Dept. of Enterprise Engineering
University of Rome Tor Vergata
{basili,croce}@info.uniroma2.it

§ Dept. of Classic Philology and Italian Studies
University of Bologna
fabio.tamburini@unibo.it,
silvia.brambilla@studio.unibo.it

Abstract

English. This paper presents work in progress for the development of IFrameNet, a large-scale, computationally oriented, lexical resource based on Fillmore's frame semantics for Italian. For the development of IFrameNet linguistic analysis, corpus-processing and machine learning techniques are combined in order to support the semi-automatic development and annotation of the resource.

Italiano. *Questo articolo presenta un work in progress per lo sviluppo di IFrameNet, una risorsa lessicale ad ampia copertura, computazionalmente orientata, basata sulle teorie di Semantica dei Frame proposte da Fillmore. Per lo sviluppo di IFrameNet sono combinate analisi linguistica,* corpus-processing *e tecniche di* machine learning *al fine di semi-automatizzare lo sviluppo della risorsa e il processo di annotazione.*

1 Introduction

Firstly developed at the University of Berkeley (California) in 1997, FrameNet adopts theories from Frame Semantics (Fillmore 1976, 1982, 1985) to NLP and explains words' meanings according to the semantic frames they evoke. It illustrates semantic *frames* (i.e. schematizations of prototypical events, relations or entities in the reality), through the involved participants (called *frame elements*, FEs) and the evoking words (or, better, the *lexical units*, LUs). Moreover, FrameNet aims to give a valence representation of the lexical units and underline the relations between frames and between frame elements (Baker et al. 1998).

The initial American project has since been extended to other languages: French, Chinese, Brazilian Portuguese, German, Spanish, Japanese, Swedish and Korean.

All these projects are based on the idea that most of the Frames are the same among languages and that, thanks to this, it is possible to adopt Berkeley's Frames and FEs and their relations, with few changes, once all the language-specific information has been cut away (Tonelli *et al.* 2009, Tonelli 2010).

With regard to Italian, over the past ten years several research projects have been carried out at different universities and Research Centres. In particular, the ILC-CNR in Pisa (e.g. Lenci et al. 2008; Johnson and Lenci 2011), FBK in Trento (e.g. Tonelli *et al.* 2009, Tonelli 2010) and the University of Rome, Tor Vergata (e.g. Pennacchiotti *et al.* 2008, Basili *et al.* 2009) proposed automatic or semiautomatic methods to develop an Italian FrameNet. However, as of today, a resource even remotely equivalent to Berkeley's FrameNet (BFN) is still missing.

As a lexical resource of this kind is useful in many computational applications (such as Human-Robot interaction), a new effort is currently being jointly made at the universities of Bologna and Roma, Tor Vergata. The IFrameNet project aims to develop a large-coverage FrameNet-like resource for Italian, relying on robust and scalable methods, in which the automatic corpus processing is consistently integrated with manual lexical analysis. It builds upon the achievements of previous projects that automatically harvested FrameNet LUs exploiting both distributional and WordNet based models (Pennacchiotti *et al.* 2008). Since the LUs induction is a noisy process, the data thus obtained need to be manually refined and validated.

The aim is also to provide Sample Sentences for LUs with the highest corpus frequency. On the one side, they will be derived from already existing resources such as the HuRIC corpus (Bastianelli 2014) or the EvalIta2011 FLaIT task data: FBK set (Tonelli, Pianta 2008) and ILC set (Lenci *et al.* 2012). On the other side, candidate sentences will also be extracted through semi-automatic distributional analysis of a large corpus - i.e. CORIS (Rossini Favretti *et al.* 2002) - and refined through linguistic analysis and manual validation of data thus obtained.

2 The development of the large scale IFrameNet resource

The need for a large-scale resource cannot be satisfied without resorting to a semi-automatic process for the gathering of linguistic evidence, selection of lexical examples as well as the annotation of the targeted texts. This work is thus at the cross roads of linguistic theoretical investigation, corpus analysis and natural language processing.

On the one hand, the matching between LUs and frames is always granted through manual linguistic validation applied to the data in the development stage. For every Frame the correctness of the inducted LUs is analysed and the 'missing' LUs, that is the BFN LUs' translations, which are absent in the inducted LU's list, are detected.

On the other hand, most choices rely on large sets of corpus examples, as made available by CORIS. Finally, the scaling to large sets of textual examples is supported by automatically searching candidate items through semantic pre-filtering over the corpus: frame phenomena are here used as queries while intelligent retrieval and ranking methods are applied to the corpus material to minimize the manual effort involved.

In the following section, we will sketch the main stages of the process that integrate the above paradigms.

2.1 Integrating corpus processing and lexical analysis for populating IFrameNet

The beneficial contribution of the interaction between corpus processing techniques and lexical analysis for the semi-automatic expansion of the FrameNet resource has been discussed since (Pennacchiotti *et al.* 2008), where LU induction is presented as the task of assigning a generic lexical unit not yet present in the FrameNet database (the so-called unknown LU) to the correct frame(s). The number of possible classes (i.e. frames) and the problem of multiple assignment make it a challenging task. This task is discussed in (Pennacchiotti *et al.* 2008, De Cao *et al.* 2008, Croce and Previtali 2010), where different models combine distributional and paradigmatic lexical information (i.e. derived from WordNet) to assign unknown LUs to frames. In particular, distributional models are used to select a list of frames suggested by the corpus' evidence and then the plausible lexical senses of the unknown LU are used to re-rank proposed frames.

In order to rely on comparable representations for LUs and sentences for transferring semantic information from the former to the latter, we exploit Distributional Models (DM) of Lexical Semantics, in line with (Pennacchiotti *et al.* 2008) and (De Cao *et al.* 2008). DMs are intended to acquire semantic relationships between words, mainly by looking at the word usage. The foundation for these models is the Distributional Hypothesis (Harris 1954), i.e. words that are used and occur in the same "contexts" tend to be semantically similar. A context is a set of words appearing in the neighborhood of a target predicate word (e.g. a LU). In this sense, if two predicates share many contexts then they can be considered similar in some way. Although different ways for modeling word semantics exist (Sahlgren 2006; Pado and Lapata 2007; Mikolov *et al.* 2013; Pennington *et al.* 2014), they all derive vector representations for words from more or less complex processing stages of large-scale text collections. This kind of approach is advantageous in that it enables the estimation of semantic relationships in terms of vector similarity. From a linguistic perspective, such vectors allow for some aspects of lexical semantics to be geometrically modelled, and to provide a useful way to represent this information in a machine-readable format. Distributional methods can model different semantic relationships, e.g. topical similarities (if vectors are built considering the occurrence of a word in documents) or paradigmatic similarities (if vectors are built considering the occurrence of a word in the (short) contexts of another word (Sahlgren 2006)). In such models, words like *run* and *walk* are close in the space, while *run* and *read* are likely to be projected in different subspaces. Here, we concentrate on DMs mainly devoted to modelling paradigmatic relationships, as we are more interested in capturing phenomena of quasi synonymy, i.e. semantic similarity that tends to preserve meaning.

2.2 The development cycle

In the following paragraphs, we outline the different stages in the development process. Each stage corresponds to specific computational processes.

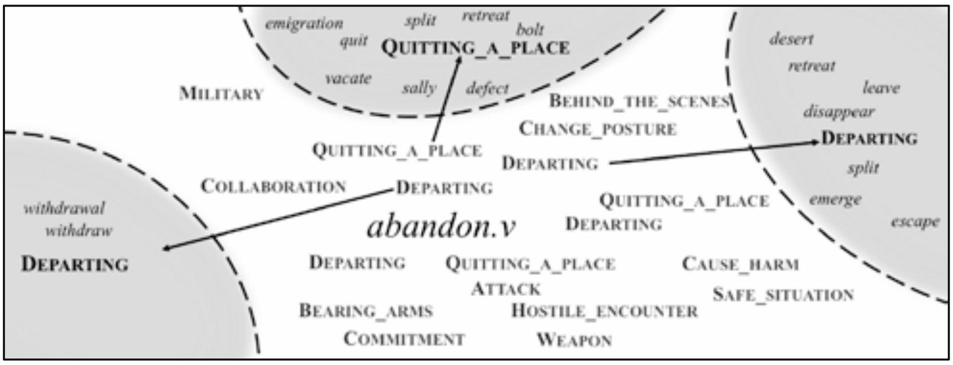

Figure 1: Three lexical clusters for the frames triggered by the verb *abandon*.v: pairs closed in the map correspond to (paradigmatic) semantic similar words and frames

Validation of existing resources. At this stage, the existing resources, dating back to previous work, are analysed and manually pruned of errors such as lexical units wrongly assigned to frames (e.g. *'asta'* or *'colmo'* to the Frame 'BODY_PARTS'), or words never assigned to their correct frame, for instances the LU '*piede*' or '*mano*' for the Frame 'BODY_PARTS'.

All the acquired Italian LUs have been compared, frame by frame, to BFN's ones, using bilingual dictionaries (e.g. Oxford bilingual dictionary) and WordNet in order to verify the correctness of matching between lexical and frames. Over the 15,134 automatically acquired ⟨*LU, frame*⟩ pairs (6,670 nouns and 8,464 verbs and adjectives), 7,377 LUs have been considered correctly assigned (2,506 verb and adjective and 4,871 noun pairs).

In addition, bilingual dictionaries, ItalWordNet and MultiWordNet have been used to manually insert a list of missing lexical entries for each frame. At the end of the process, the resulting validated and refined ⟨*LU,frame*⟩ amount to 7,902 (5,128 nouns and 2,774 verbs and adjectives).

Corpus processing and lexical modeling. At this stage, the LUs made available from manual validation are used to model distributionally the individual frames. Firstly, distributional corpus analysis is applied to map individual LUs into distributional vectors. A distributional model will be acquired from the CORIS corpus by applying the neural method presented in (Mikolov *et al.* 2013). It will enable the acquisition of geometrical representations for words in a high dimensional space where distance reflects the paradigmatic relation among words. This model can also be adopted to build a representation for sentences, as traditionally carried out by Distributional Semantic models, e.g. (Landauer and Dumais 1997) or (Mitchell and Lapata, 2010).

Lexical clustering is important here as specific space regions enclosing the instance vectors of some considered LUs correspond to semantically coherent lexical subsets. This is a priming function for mapping unseen word vectors to frames, as applied in (De Cao *et al.* 2008): the centroids of the possibly multiple clusters generated by the known LUs of a given frame *f* are used to detect all regions expressing *f* and thus predict the predicate *f* over previously unseen words and sentences. Examples of semantically coherent regions evoked by the verb *abandon* for the English Framenet are reported in Fig. 1. Here different lexical clusters for a given frame (i.e. DEPARTING) are depicted while different frames (e.g. DEPARTING, QUITTING_A_PLACE, COLLABORATION) are also evoked by the verb. It should be noted that in the figure distances in the two-dimensional plot correspond to distances between the word embedding vectors, while each lexical cluster is expressed as the centroid of its member vectors.

The distributional information has been acquired for the considered 7,902 LUs from CORIS and used to support the LU mapping and the sentence validation. In fact, given a sentence *s* containing a target LU *l*, a specific geometrical representation for *s* can be derived by linearly combining all vectors representing words *w* surrounding *l* in sentence *s*. This duality property allows the embedding space to represent sentences *s*, lexical units *l* as well as generic words *w*. This enables to model the relevance of a frame *f* for an incoming sentence *s* through the distance *d*(*f,s*) between vectors $\underline{f}$ related to a centroid for a frame *f* and the vector $\underline{s}$ of the sentence *s*. It corresponds to a confidence measure computed for a rule such as:

"*s* is a valid example of the usage of frame *f*"

The open aspects of the above semi-automatic process are the following:

I. How to design a suitable representation (centroid or model) for a frame *f*
II. How to define the vector for a sentence *s*
III. How to compute the distance function *d*(*f,s*)

The current research activity is focusing on the best solution for these issues and part of our experimental activity is devoted to assess these design choices, as discussed in Section 3.

First Lexical Analysis and Validation. A further stage for the resource development focuses on the selection of a significant sample of LUs, chosen on the basis of their high semantic salience and for their high number of occurrences in the corpus (*primary* LUs). By relying on the method described above, we use the distributional representation of words, lexical units and sentences, to gather CORIS sentences *s* where a LU occurs and evaluate its suitability as an example for the evoked *f*. This decision function is based on the geometric distance *d*(*f,s*) that can be computed over a large number of sentences *s*. When this step is carried out in CORIS, the validation of the acquired candidate sentences allows for positive examples of a frame *f* to develop quickly*:* this is used to trigger supervised learning of *f*.

The manually validation in fact confirms the proper correspondence between automatically selected sentences and LUs that evoke a targeted frame *f*. It produces novel seed examples for *f*: these will serve as a training set for a semi-automatic stage of resource expansion.

Semi-automatic resource expansion. The acquired distributional model will support the semi-automatic expansion of the seed set, by selecting the most semantically similar word to the seed set and assigning them to frames by applying the methodologies suggested in (Pennacchiotti *et al.* 2008, De Cao *et al.* 2008, Croce and Previtali 2010). Moreover, the same distributional model will support the assignment process of sentences to frames. We will in fact investigate semi-supervised models based on clustering techniques (Pennacchiotti *et al.* 2008) or other supervised approaches such as Support Vector Machines as in (Croce and Previtali 2010).

Final Validation and Release. The extracted sentences will be ordered by decreasing probability, according to their distributional collocation, and a list of 15 to 20 candidates per LU will be provided. This list will be manually validated. The aim is to provide at least 4 sample sentences for each of the primary LUs.

3 Status of the Project and Perspective Views

Although the general software architecture for the project progress is available, the overall process described above has not been fully accomplished.

Current material covers a set of 554 frames and 7,902 lexical units, of which 2,604 verbs, 5,128 nouns and 170 adjectives. The average number of occurrences for each of these selected words is higher than 9,400, although there are still 508 words not present in CORIS.

All these occurrences correspond to a number of about 70 millions non validated and unsorted sentences. In the rest of the paper, we describe the outcome of the First Lexical Analysis and Validation stage: its aim is to trigger the semi-automatic learning and tagging of the whole corpus, according to the methods suggested in section 2.2.

3.1 Empirical Investigation: First Lexical Analysis and Validation

The stage *First Lexical Analysis and Validation* has been currently accomplished. The three research questions posed above: (I) the modelling of a frame *f*, (II) the sentence representation and (III) the definition of a distance function able to model similarity between sentences.

About the problem (I) two approaches are possible. We can model a frame via clustering its lexical units and applying the method described in (Pennacchiotti *et al.* 2008, De Cao *et al.* 2008). On the contrary, we can adopt a supervised technique. A frame *f* is represented as the target class of instances corresponding to $\langle s,l \rangle$ pairs, where *s* is an input sentence and *l* is a lexical unit: a statistical classifier is trained to map $\langle s,l \rangle$ into a confidence value and its output *h*(*s,l,f*) corresponds to the system's confidence that the sentence

"*f* is the frame evoked by *l* in *s*"

is true. Notice that the pair $\langle s,l \rangle$ can be expressed as an instance by combining the embedding vector $\underline{l}$ of its lexical unit *l* with a vector $\underline{s}$ for *s*.

As a solution for the problem (II) we define $\underline{s}$ as the linear combination of vectors $\underline{w}$, for each word *w* in *s,* i.e. $\underline{s} = \sum_{w \in s} \underline{w}$.

The above formulation allows to define the classification task as follows:

Given a sentence *s* including a word *l* as a potential frame evoking LU, *Find* the frame *f* that characterizes *l* in *s*.

The solution of the above problem over a $\langle s,l \rangle$ pair would also be a useful solution for the problem (III), as the confidence *h*(*s,l,f*) in the classification of a sentence *s* in a frame *f* for *l* can be retained as the inverse of the target distance function *d*(*f,l*) local to the sentence.

The major problem with the above formulation is that the training of the statistical classifier is not possible without the availability of useful examples of different frame *f*. The idea is thus to develop ways to derive from CORIS the proper candidates *s* for *f* through the knowledge of some of its LUs. In the bootstrapping stage, we define as virtual examples the pairs $\langle l, \{l\} \rangle$ that are retained as positive examples for the frame *f*, for every *l* that is a known lexical unit for *f*. In our approach, an example is thus obtained by modelling the sentence *s* as a singleton {*l*}, i.e. the lexical unit *l*.

A statistical classifier considers every known LU as an individual (positive) example and can be applied to every LU in our initial resource (i.e. 7,902 for the 554 frames).

In synthesis, the method works as follows. First, for every lemma *w* in the corpus, an *n*-dimensional embedding vector $\underline{w}$ is derived, according to (Mikolov *et al.* 2013). As a side effect, for every LU *l* of each known frame *f*, the lexical embedding vector $\underline{l}$ is used to build the example ($\underline{l}$, $\underline{l}$) for the LU sentence pair: $\langle l, \{l\} \rangle$.

A multiclass-statistical categorizer is trained for every frame *f* for which at least 5 examples (i.e. 5 different LUs) where available.

When applied to an incoming sentence *s* including a LU *l,* the classifier outcome *h*(*l,s, f*) is said to accept the frame *f* if:

- *f* belongs to the set of frames evoked by *l*
- *f* = argmax$_{f'}$ { *h*(*l,s,f'*) }

For every sentence *s* including a frame evoking lexical unit *l*, the above function suggests one candidate frame among the possibly multiple ones. When the scoring function *h* is negative everywhere (e.g. with the SVM formulation of a classification task), the sentence is rejected and is not considered a valid example for future iterations.

The application of this method to the CORIS corpus has been carried out applying a multi-classifier SVM with linear kernel to the 2*n*-dimensional vectors of each pair $\langle l, \{l\} \rangle$. Starting from the lexicon validated in the first stages, the SVM has been able to label over 2 million sentences.

3.2 Empirical Investigation: Current Results

In order to evaluate the proposed supervised classification method for the stage "*First Lexical Analysis and Validation*" we run and experimental evaluation over a set of 326[1] frames, the ones with more than 5 lexical units in the initial lexicon. In this way, we selected 1,095 different LUs, represented as an embedding vector in the wordspace. On average, we have 12 LU per frame, and every individual lexical entry *l* appears in about 1.88 frames. The baseline of a classification task that maps a sentence *s* including a lexical unit into its own frame is about 35%, as for the ambiguity characterizing most frequent entries.

We asked three annotators to evaluate individual triples $\langle l, s, f \rangle$ validating the system proposal. Four main cases where possible:

- MISSING FRAME. The sentence *s* is not manifesting any of the frames *f* evoked by the lexical unit *l*, but corresponds to a frame not yet present in the lexicon for *l*. In this case the algorithm cannot provide the suitable frame, as it cannot generate a novel frame.
- NOT APPLICABLE. The sentence *s* does not contain an occurrence of the lexical unit *l* in one of its proper senses: this case is typical for phraseological uses of a verb such as *morire di freddo*, *andare di fretta,* ... that do not directly correspond to lexical predicates and thus cannot be treated through the lexical embedding vectors.
- CORRECT/INCORRECT, when the outcome *argmax*$_{f'}$ { *h*(*l,s,f'*) } is correct (or incorrect) as the frame evoked by *l* in *s* is exactly (or not) *f*.

According to the above method annotators validated 667 sentences for 113 frames and 212 different *verbal* lexical units. The analysis resulted into a *precision* (i.e. the number of correct candidate frames emitted by the algorithm w.r.t. the number of valid cases, that is all but the MISSING FRAME or NOT APPLICABLE cases) is 75,2%, well beyond the 35% baseline. The method could be applied onto the 74,5% of the sentences, including CORRECT cases and MISSING FRAME cases. We neglected in this coverage score the NOT APPLICABLE cases that amount to 44 sentences, i.e. about 6,4%.

Examples of the correct assignment of the algorithm on quite ambiguous verbs, such as *finire* (i.e. *to end*, in frames ACTIVITY_FINISH, CAUSE_TO_END and KILLING) or *rivelare* (i.e. *to reveal*, in frames REVEAL_SECRET, OMEN, EVIDENCE) are the following:

La vicenda avrebbe potuto [*finire*]$_{\text{ACTIVITY\_FINISH}}$ *lì , ma il prefetto di Nuoro fece presentare ...*

In prova si è [*rivelato*]$_{\text{EVIDENCE}}$ *ad altissimo livello sia sull' asciutto sia sul ...*

[1] By keeping the frames that include at least 4 lexical units the number of targeted frames grows to 371.

An example of Missing Frame is BEAT_OPPONENT for the verb *battere* in

... impegnato a fornire quante più informazioni possibili, anche per [*battere*]$_{\text{BEAT\_OPPONENT}}$ *la concorrenza dei siti Ipsoa e il ...*

as the lexicon of the verb *battere* only includes the frames CAUSE_HARM, CORPORAL_PUNISHMENT and EXPERIENCE_BODI-LY_HARM.

The experiments only run over verbal lexical units will be extended soon to nouns and adjectives. However, the encouraging precision reached by the method allows for direct use it in an iterative active learning schema, where the more ambiguous sentences found and annotated within a specific training stage are used to train the system at the next stage. We expect this to speed up the lexicon development process and to allow bootstrapping with fewer resources. The lexicon will be made available for crowdsourcing further annotations and delivered incrementally in the next few months.

References

Baker C. F., Fillmore, C. J., Lowe, J. B.. (1998). The Berkeley FrameNet project. In: COLING '98 Proceedings of COLING '98, 1. Canada, 86-90.

Basili R., De Cao D., Croce D., Coppola B., Moschitti A. (2009). Cross-Language Frame Semantics Transfer in Parallel Corpora. In: Proceedings of the CICLing 2009, Best Paper Award. Mexico

Bastianelli, E., Castellucci, G., Croce, D., Iocchi, L., Basili, R., & Nardi, D. (2014). HuRIC: a Human Robot Interaction Corpus. In Proceeings of *LREC* 2014, 4519-4526.

Croce, D. and Previtali, D. (2010). Manifold learning for the semi-supervised induction of framenet predicates: an empirical investigation. In Proceedings of GEMS '10, pages 7–16, Stroudsburg, PA, USA.

De Cao D., Croce D., Pennacchiotti M., Basili R. (2008). Combining word sense and usage for modeling frame semantics. In Proceedings of STEP 2008, Italy

De Cao D., Croce D., Basili R. (2010). Extensive Evaluation of a FrameNet-WordNet mapping resource. In: Proceedings of the LREC 2010, Malta.

Fillmore, C.J. (1985). Frames and the semantics of understanding. Quaderni di Semantica, VI(2), 222-254.

Fillmore, Charles J. (1976). Frame semantics and the nature of language, Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, vol. 280, pp. 20-32

Fillmore, C. J. (1982). Frame semantics. Linguistics in the morning calm, pp. 111-137.

Harris, Z. (1954). Distributional structure. In Jerrold J. Katz and Jerry A. Fodor, editors, The Philosophy of Linguistics, New York. Oxford University Press.

Johnson, M. And Lenci, A. (2011). Verbs of visual perception in Italian FrameNet, Advances in Frame Semantics, 3(1), 9–45

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104.

Lenci, A, Johnson, M, Lapesa, G. (2010). Building an Italian FrameNet through Semi-automatic Corpus Analysis. Proceedings of LREC 2010. Malta.

Lenci, A., Montemagni, S., Venturi, G, Cutrullà, M. G. (2012). Enriching the ISST-TANL Corpus with Semantic Frames in Proceedings of LREC 2012, Istanbul, Turkey

Mikolov, T., Kai Chen, Greg Corrado, and Jeffrey Dean. (2013). Efficient estimation of word representations in vector space. CoRR abs/1301.3781. http://arxiv.org/abs/1301.3781.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. Cognitive Science, 34(8):1388–1429.

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

Pennacchiotti M., De Cao D., Basili R., Croce D., Roth M. (2008). Automatic induction of FrameNet lexical units. In: Proceedings of the EMNLP 2008, Hawaii

Pennington, J., Socher, R. and Manning, C. (2014). GloVe: Global Vectors for Word Representation, In Proceedings of EMNLP 2014, 1532-1543.

Rossini Favretti R., Tamburini F., De Santis C. (2002). CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In A Rainbow of Corpora: Corpus Linguistics and the Languages of the World, Lincom-Europa, Munich, 27-38.

Sahlgren, M.. (2006). The Word-Space Model. Ph.D. thesis, Stockholm University.

Tonelli, S. and Pianta, E. (2008). Frame information transfer from English to Italian. In Proceedings of LREC, Marrekech, Morocco

Tonelli, S, Pighin, D, Giuliano, C, Pianta, E. (2009). Semi-automatic Development of FrameNet for Italian. In Proceedings of the FrameNet Workshop and Masterclass, Milano, Italy. Milan, Italy

Tonelli, S. and Pighin, D. (2009). 'New Features for FrameNet - WordNet Mapping', in Proceedings of CoNLL 2009, Boulder, Colorado, 219–227

Tonelli, S. (2010). "Semi-automatic techniques for extending the FrameNet lexical database to new languages", Università Ca' Foscari, Venezia

Venturi G., Lenci A., Montemagni S., Vecchi E., Sagri M., Tiscornia D., Agnoloni T. (2009). Towards a FrameNet Resource for the Legal Domain. In Proceedings of LOAIT 2009. Barcelona, Spain