Linking an Abstract Corpus Grammar to a Lexical Semantic Network

Jean-Philippe Prost

Linking an Abstract Corpus Grammar to a Lexical Semantic Network

Jean-Philippe Prost
Laboratoire Parole et Langage (LPL), Aix-Marseille Université, France
Jean-Philippe.Prost@univ-amu.fr
Research report, February 2021

Abstract

In this paper we address the problem of the combined representation of heterogeneous sources of knowledge within a unique and homogeneous data structure. The goal is ultimately to enable the holistic processing of linguistic and world knowledge, where all the dimensions may interact seamlessly. We focus here on bridging the gap between Syntax and Semantics. We propose to link an abstract grammar to an existing lexical network through the adoption of the same underlying graph structure. The resulting structure may be seen as a multi-layer linguistic network. The solution we introduce for the abstract grammar layer relies on a graph-theoretic interpretation of Property Grammar. The typed structure we propose supersedes both phrase structure and dependency structure, which are covered with specific relation types – Constituency and Dependency respectively. We present a procedure to derive the grammar from an annotated corpus, and we illustrate the procedure with the French Treebank.

Keywords: Knowledge graph, linguistic network, Property Grammar, JeuxDeMots, Syntax, Semantics

1. Introduction

For decades the common trend in NLP has been to model natural language along different and separate dimensions (lexicon, syntax, semantics ...), typically through pipeline architectures, for the sake of *Divide and Conquer* problem solving strategies. A key problem with this model is that keeping the dimensions nearly hermetically separate indeed prevents the integration of different dimensions, e.g. Syntax and Semantics, within the same process. Yet such an integration could often help solve cases of ambiguity, whether lexical, semantic or syntactic, or combinations thereof. We can also expect an integrated representation of linguistic knowledge to be helpful in situations such as the comprehension of non-canonical language, whether noisy, erroneous, fragmented, ..., which often narrows down to disambiguating likely interpretations.

The challenge for such an integration to happen at the data level is to find a way to represent the knowledge in all the dimensions at the same level of abstraction.

In this paper we introduce a graph-theoretic interpretation of Property Grammar (Blache, 2001), which enables the representation of an abstract grammar to be consistent with the data structure of a lexico-semantic network. The result is a single graph, where syntactic relations and nodes coexist along with semantic ones. We present a procedure to derive the grammar from a treebank annotated with both phrase structures and dependency structures.

2. Literature Review

Language-related data and Knowledge Bases The structuring of language knowledge in large Knowledge Bases (KBs) and resources is mostly organised along the different steps of the traditional pipeline architecture. The resources referenced in LLOD[1], for example, are concerned with different dimensions of language knowledge: lexicon and dictionaries (e.g. WordNet, WordNet-RDF, BabelNet, Wiktionary) for word level and lexical semantics, domain-related terminologies, ontologies and knowledge bases about world knowledge (e.g. YAGO, DBPedia), annotated corpora for syntax, and a few others for metadata and other kinds of resources. As far as we know no Knowledge Base is referenced which would account for abstract grammar knowledge about language.

Although none of those resources adopts a holistic perspective on language, the overall intention of the LLOD initiative goes towards an integration of all dimensions. At this stage, it solely aims to address the problem of format compatibility among resources, and to serve as a central repository of linked resources in order to ease their interoperability. The alignment of all aspects of the linguistic resources involved remain a challenge to be addressed. The next step is a full and deep integration across all the linguistic dimensions within a single homogeneous resource, in order to enable the holistic representation and processing of language knowledge.

SAR-Graphs introduce a different kind of resource inclusive of syntactic data while linking to different lexico-semantic resources such as WordNet, BabelNet, and YAGO. SAR-Graphs are described as *graphs of Semantically-Associated Relations* (Krause et al., 2015a). They address the question of the integration of different aspects of linguistic knowledge within a homogeneous structure referred to as a *language network* (Uszkoreit and Xu, 2013; Krause et al., 2015b). They integrate lexical semantics with semantic relations about facts and events extracted from KBs such as Freebase (Bollacker et al., 2008)[2] through the use of syntactic patterns learned from machine-parsed input text. For every high-level semantic relation (i.e. modelling world knowledge, such as facts and events) found in a given KB a SAR-Graph merges all the machine-parsed dependency structures, which match the syntactic patterns extracted (separately) for that relation. The syntactic extraction patterns are learned with the DARE relation extraction system (Xu et al., 2006; Xu et al., 2007) and

[1] http://linguistic-lod.org/

[2] Now included in the Google Knowledge Graph.

kept separate from the main resource. Therefore the SAR-Graphs still do not include any level of abstract grammar knowledge. The grammar is actually contained in DARE, through the use of the principle-based MINIPAR dependency parser (Lin, 1994).

Complex Syntactic Networks The past decade has seen the emergence of the use of complex networks (Newman, 2010) for studying the syntax of human languages, through Complex Syntactic Networks. As of today, they have only been used for representing dependency relationships among words, or word forms (Čech et al., 2016): the words (lemmas) are nodes, and the dependency relations are edges between nodes. The term *dependency* must be taken in the sense of the Dependency Grammar formalism (Tesnière, 1959; Hudson, 2006). The construction of a syntactic network requires the prior dependency-based syntactic analysis of corpora, either through an automated parsing process or a dependency treebank, or both. Although nothing prevents the use of other grammar formalism (e.g. phrase structure grammars) for syntactic network analysis, the literature shows no evidence of it (Čech et al., 2016).

Semantic networks WordNet is a lexical network based on synsets which can be roughly considered as concepts. EuroWordnet (Vossen, 1998), a multilingual version of WordNet and WOLF (Sagot and Fier, 2008), a French version of WordNet, were built by automated crossing of WordNet and other lexical resources along with some manual checking. Navigli and Ponzetto (2010) constructed automatically BabelNet, a large multilingual lexical network, from term co-occurrences in the Wikipedia encyclopedia. HowNet (Dong and Dong, 2006) is a hand-crafted lexical network based on concepts linking both English and Chinese. The Réseau Lexical du Français (RLF, *French Lexical Network* (Polguère, 2014) is a resource based on the notion of *lexical function* as defined by Igor Mel'čuk. The resource concerns about 10000 terms and is mainly manually populated with data.
JDMRezo is an open lexico-semantic network for French (Lafourcade, 2007). It has been built and validated for years through an ecosystem of semi-automatic processes. Crowdsourcing is central, with a collection of Games With A Purpose (GWAPs)[3]. Automatic inference mechanisms also contribute to acquire new knowledge, some relying on external resources, and some not. The network grows constantly. As of today[4] it is made up of 14+ million nodes including 4.2+ million terms, 310+ relations and 150+ relation types. It is by far the largest open resource of its kind in the world. The nodes are terms, concepts and symbolic information. The relations are lexical, morphological, pragmatic, logical, ontological, ...

Model-Theoretic Syntax (MTS) MTS provides a great deal of features for representing non-canonical language and other linguistic phenomena such as lexical openness, grammatical fragments, or a graded notion of grammaticality, for which frameworks for Generative-Enumerative Syntax are of no use (Pullum and Scholz, 2001). These phenomena correspond to situations where no strong model exists for a given utterance. From a computational point of view their parsing requires the relaxation of violated constraints through some form of non-classical reasoning. Prost (2008) addresses questions related to graded grammaticality (also known as *gradience*[5]) as a Constraint Optimisation Problem and experiments with different ways of grading the grammaticality of non-canonical utterances within the Property Grammar (PG) framework for MTS (Blache, 2001). Later works (Prost, 2009; **?**) show how to rely on that framework in the context of grammar error detection.

3. The grammar network

Within frameworks for Model-Theoretic Syntax is Property Grammar (PG) (Blache, 2001). In PG the grammar of a natural language may be represented as a graph (V, E), where the vertices in V model morpho-syntactic categories, and the edges in E model unary or binary syntactic relations between vertices. PG pre-defines the semantics of 8 relation types (Constituency, Dependency, Agreement, Linearity, Obligation, Uniqueness, Requirement, Exclusion) called *Properties*. A Model-Theoretic axiomatisation was provided by Duchier et al. (2009).
For instance, the graph-theoretic interpretation of the property of linear precedence NP:D $\prec$ N, which states that in a Noun Phrase (NP) the Determiner (D) precedes the Noun (N) in read direction, corresponds to the typed relation $\prec$ illustrated in Figure 1. Note that in this representation the classical – phrase and dependency – structures simply correspond to specific types of relation, respectively *Constituency* and *Dependency*. In the case of the phrase structure the right hand side of any rewrite rule is represented by a node of type *aggregate*; this aggregate node is then related, through a *rewrites* relation, to a node modelling the left hand side of the corresponding rule.
The graph illustrated in Figure 1 represents the typed relations that compose the rewrite rule (NP $\rightarrow$ D N)$(w1)$ weighted with $(w1)$ and the dependency $D \rightsquigarrow N$. The PG properties correspond to different types.

3.1. Grammar Derivation from a Treebank

Prost (2014) described a procedure for deriving a partial PG grammar from the CFG grammar encoded in a treebank. Every PG property type except Dependency and Obligation is associated with a derivation rule according to its semantics. Recall that a PG property type corresponds to a relation type in our graph-theoretic current interpretation. For instance, the Constituency relationship associates any morpho-syntactic category found in the left hand side of rewrite rules with the set of all the categories that appear in the right hand sides of the concerned rules. For example, let us consider that we observe on corpus the rules (r1)(NP $\rightarrow$ D N) and (r2)(NP $\rightarrow$ D A N). From (r1) and (r2) we can derive that the Constituency relationship for the NP category is the set $\{D, N, A\}$.
Table 1 gives the semantics of the relation types according to PG.

[3] `http://jeuxdemots.org/`

[4] 26 February 2020, according to the values checked on `http://www.jeuxdemots.org/jdm-about.php`.

[5] For a linguistic account of gradience in syntax see (Aarts, 2007).

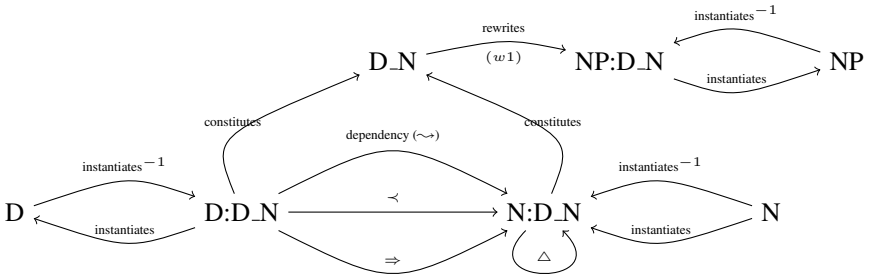

Figure 1: The syntactic relations that compose the WCFG rewrite rule (NP $\rightarrow$ D N)$(w1)$ and the dependency $D \rightsquigarrow N$

Obligation	$A : \triangle B$	at least one daughter of A is of category B
Constituency	$A : S?$	the category of every daughter must be in S
Uniqueness	$A : B!$	at most one daughter of A is of category B
Linearity	$A : B \prec C$	a daughter of category B precedes a daughter of category C in read direction
Requirement	$A : B \Rightarrow C$	the existence of a daughter of category B requires a a daughter of category C
Exclusion	$A : B \nRightarrow C$	daughters of categories B and C may not co-occur under the same A

Table 1: Semantics of the usual PG property types

Derivation rules Let C be the label of a morpho-syntactic category. R_C denotes the set of all the context-free rules, which take C as a left hand side. We define RHS as the function, which to every C associates the set $\mathrm{RHS}(R_C)$ of all the categories present in the right hand side of a rule in R_C.
Table 3.1. is a reminder of all the derivation rules introduced in (Prost, 2014), which correspond to all the PG property types except Obligation, Agreement and Dependency. These three are dealt with separately, since they can not be derived easily. The special case of Obligation is discussed later.

The dependency structure Obligation and Dependency are two property types that are not automatically derived from a constituency treebank using the rules in (Prost, 2014). Meanwhile, constituency trees may be automatically converted into Dependency trees (Candito et al., 2010). The latest version of the French Treebank (FTB) (Seddah et al., 2013) provides both constituency structures and the corresponding dependency structures that were obtained with a slightly modified version of the conversion procedure by Candito et al. (2010). As a consequence the FTB now comes with both constituency and dependency trees for the same underlying corpus, which allows us to easily extract the missing relations for our grammar network.

3.2. Modelling syntactic gradience

We already discussed that as an MTS framework Property Grammar comes with handy features for modelling gradience and non-canonical language. Prost et al. (2016) have shown that given a constituency treebank the derived PG corpus grammar can be compiled into a Weighted Context-Free Grammar (WCFG), where the corpus rules are completed with theoretical ones – though unobserved. Provided the corpus PG grammar every additional theoretical rule $C : R*$ of left hand side C can be characterised into P^+_{R*}

Obligation rule – *not implemented* The Obligation property $C : \triangle H_C$ is specified by the set H_C of the disjunctions $\psi = \bigvee e$ of the distinct categories e, such that $\forall r \in \mathrm{RHS}(R_C) \exists e \in r$.

Constituency rule The Constituency property $C : E_C?$ is specified for the category C by the set E_C of unique categories e, such that $\exists r \in \mathrm{RHS}(R_C)$ with $e \in r$

Uniqueness rule The Uniqueness property $C : U_C!$ is specified for the category C by the set U_C of all the unique categories, which never co-occur with themselves within the right hand side of a rule.

Linearity rule The set of Linearity properties $C : a \prec b$ for the category C is specifie by the set L_C of the consistent ordered pairs (a, b), where (a, b) is consistent iff $\exists r \in R_C$ such that a and b co-occur in the right hand side of r, and $\neg\exists r' \in R_C$ such that $(b, a) \in r'$.

Requirement rule The semantic of the Requirement property $C : x \Rightarrow y$ differs from the classical implication, in that unlike the implication $(C : x \Rightarrow y) \not\equiv (\neg x \vee y)$. Therefore the set Z_C of the Requirement properties is specified by the set of co-occurrences (a, b) for the category C, minus those for which $\exists r \in \mathrm{RHS}(R_C)$ such that $a \in r$ and $b \notin r$.

Exclusion rule The set of Exclusion properties $C : x \nRightarrow y$ is specified by the set X_C of the unordered pairs of categories (a, b) such that a and b never co-occur within the same right hand side of a rule.

Figure 2: Rules for deriving relation types from a constituency corpus

and P^-_{R*}, the sets of respectively all the satisfied and all the violated properties for C.

For example, we could complete the grammar illustrated in Figure 1 with the theoretical rule (r3*)(NP $\rightarrow$ D N N). For building the graph for (r3*) we need to characterise it, that is, to check for every possible relation whether it holds true or false. If we assume that all the properties deemed true for the category NP are those represented in the graph from Figure 1, then (r3*) is characterised by $P^+_{r3*} = \{\triangle N, D \prec N, D \Rightarrow N, D \rightsquigarrow N, \{D, N\}?\}$ and $P^-_{r3*} = \{N!\}$. In our graph-theoretic representation the truth values are represented by weights. An arbitrarily positive value associated with an edge stands for a relation that holds true, while a negative value stands for one that holds false.

Other grammaticality- and acceptability-related scores can also be computed on the basis of those numbers of satisfied and violated properties (Prost, 2008; Prost et al., 2016), which can be used to weight relations. In Figure 1 the relation `rewrites` is weighted with the score $(w1)$, meant to represent the score of grammaticality associated with the corresponding rewrite rule.

3.3. Hooking up a corpus grammar to the lexical network JDMRezo

A lexical network such as JDMRezo relates every term to one or more Part-Of-Speech (POS) nodes, i.e., nodes that are typed `n_pos`. The `n_pos` type is a generalisation of the notion of morpho-syntactic category and the features that may come along, e.g. number, gender, tense, etc. The domain of the type `n_pos` theoretically includes all the possible combinations of values. For instance, the network includes all the following nodes: `Nom` for Noun, `Nom:PL` for 'Noun plural' and `Nom:Fem+PL` for 'Noun feminine plural'. Therefore, a noun feminine plural is expected to be related to all three `n_pos` nodes `Nom`, `Nom:PL` and `Nom:Fem+PL`. Yet in practice since the values have not been added to the network in a systematic way but whenever needed, some values go missing.

As far as annotation models are concerned, JDMRezo adopted one of its own, hence does not meet any existing standard. Therefore when using a constituency treebank a lookup table is required, which matches the annotation models. The latest version of the French Treebank (FTB) (Seddah et al., 2013) meets the Penn Treebank annotation scheme, which makes it OLiA-compatible (Chiarcos and Sukhareva, 2015). On the lexical level JDMRezo is linked to BabelNet through specific relation types.

4. Conclusion

In this paper we introduced a graph-theoretic architecture for representing an abstract grammar in a way that makes it consistent with the data structure of a lexico-semantic network. This architecture is based on Property Grammar (PG) (Blache, 2001), a framework for Model-Theoretic Syntax (MTS) (Duchier et al., 2009). The underlying architecture is a labelled and typed oriented graph, where the PG property types are modelled as relation types that are compatible with the lexical-semantic relation types in use in network such as JDMRezo (Lafourcade, 2007). Based on previous works from Prost et al. (2016), we also presented a procedure for deriving a PG grammar from a constituency/dependency treebank, such as the latest release of the French Treebank (Seddah et al., 2013). This architecture goes towards a holistic perspective on the computational modelling of natural language.

Further works include the full implementation of our model for a graph-theoretic abstract grammar, and its integration with other linguistic resources. Then foremost, graph algorithms must be experimented with in order to check whether some might prove efficient for NLP tasks such as semantic parsing. What if the semantic parsing problem could be modelled as a graph traversal problem, such as the shortest path or the business salesman problem? If so, how do existing algorithms perform?

The ambition is to make the resource grow and improve with time, and to develop it as an ever-growing holistic (linguistic) knowledge base. All kinds of questions are expected to arise along the way, regarding its development, its maintenance, its processing, its evaluation, etc. Crowdsourcing through GWAPs might be an option for several of these questions, following the works around JeuxDeMots (Lafourcade et al., 2015) to maintain, validate and populate JDMRezo, or around Zombilingo (Fort et al., 2014) to annotate dependency structures.

The model for syntax itself can also be improved in various ways. The grammar can likely be lexicalised through the memorisation of partial parses and syntactic patterns. The patterns in use in the SAR-Graphs, for instance, might be compatible with our structure. The SAR-Graphs could also be of interest for linking world knowledge, such as the facts and events that are stored in Freebase-like knowledge bases.

5. Acknowledgements

6. Bibliographical References

Aarts, B. (2007). *Syntactic gradience: the nature of grammatical indeterminacy*. Oxford University Press.

Blache, P. (2001). *Les Grammaires de Propriétés: des contraintes pour le traitement automatique des langues naturelles*. Hermès Sciences.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Candito, M., Crabbé, B., and Denis, P. (2010). Statistical french dependency parsing: treebank conversion and first results.

Čech, R., Mačutek, J., and Liu, H., (2016). *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, volume 99 of *Understanding Complex Systems*, chapter Syntactic Complex Networks and their Applications, pages 167–186. Springer Berlin Heidelberg, Berlin, Heidelberg, jan.

Chiarcos, C. and Sukhareva, M. (2015). Olia–ontologies of linguistic annotation. *Semantic Web*, 6(4):379–386.

Dong, Z. and Dong, Q. (2006). *HowNet and the Computation of Meaning*. WorldScientific, London.

Duchier, D., Prost, J.-P., and Dao, T.-B.-H. (2009). A Model-Theoretic Framework for Grammaticality Judgements. In *Proceedings of Formal Grammar (FG'09)*, volume 5591 of *LNCS*. FOLLI, Springer.

Fort, K., Guillaume, B., and Chastant, H. (2014). Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 2–6.

Hudson, R. (2006). *Language Networks: The New Word Grammar*. Oxford University Press.

Krause, S., Hennig, L., Gabryszak, A., Xu, F., and Uszkoreit, H. (2015a). Sar-graphs: A Linked Linguistic Knowledge Resource Connecting Facts with Language. *ACL-IJCNLP 2015*.

Krause, S., Hennig, L., Gabryszak, A., Xu, F., and Uszkoreit, H. (2015b). Sar-graphs: A linked linguistic knowledge resource connecting facts with language. *ACL-IJCNLP 2015*, page 30.

Lafourcade, M., Joubert, A., and Le Brun, N. (2015). *Games with a Purpose (GWAPS)*. Wiley Online Library.

Lafourcade, M. (2007). Making people play for Lexical Acquisition. In *Proc. SNLP 2007*, pages 13–15, Pattaya Thailande, December. 8 p. 7th Symposium on Natural Language Processing.

Lin, D. (1994). Principar: an efficient, broad-coverage, principle-based parser. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 482–488. Association for Computational Linguistics.

Navigli, R. and Ponzetto, S. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

Newman, M. (2010). *Networks: An Introduction*. OUP Oxford.

Polguère, A. (2014). From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, 27(4):396–418.

Prost, J.-P., Coletta, R., and Lecoutre, C. (2016). Compilation de grammaire de propriétés pour l'analyse syntaxique par optimisation de contraintes. In *Actes de TALN 2016, 23ème conférence sur le Traitement Automatique des Langues Naturelles*, pages 396–402, Paris, France, Juillet. Association pour le Traitement Automatique des Langues.

Prost, J.-P. (2008). *Modelling Syntactic Gradience with Loose Constraint-based Parsing*. Ph.D. thesis, Macquarie University, Australia and Université de Provence, France (cotutelle).

Prost, J.-P. (2009). Grammar error detection with best approximated parse. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 172–175. Association for Computational Linguistics.

Prost, J.-P. (2014). Jugement exact de grammaticalité d'arbre syntaxique probable. In *Proceedings of TALN 2014 (Volume 1: Long Papers)*, pages 352–362, Marseille, France. Association pour le Traitement Automatique des Langues.

Pullum, G. K. and Scholz, B. (2001). On the Distinction Between Model-Theoretic and Generative-Enumerative Syntactic Frameworks. In *Logical Aspects of Computational Linguistics: 4th International Conference (LACL)*, Lecture Notes in Artificial Intelligence, pages 17–43, Berlin.

Sagot, B. and Fier, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *Proceedings of TALN 2008*, Avignon, France.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J., Farkas, R., Foster, J., Goenaga, I., Gojenola, K., Goldberg, Y., et al. (2013). Overview of the spmrl 2013 shared task: cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*, pages 146â–182. Association for Computational Linguistics.

Tesnière, L. (1959). Éléments de syntaxe structurale.

Uszkoreit, H. and Xu, F. (2013). From strings to things sar-graphs: A new type of resource for connecting knowledge and language. In *Proceedings of the 2013th International Conference on NLP & DBpedia-Volume 1064*, pages 109–117. CEUR-WS. org.

Vossen, P. (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Xu, F., Uszkoreit, H., and Li, H. (2006). Automatic event and relation detection with seeds of varying complexity. In *Proceedings of the AAAI workshop event extraction and synthesis*, pages 12–17.

Xu, F., Uszkoreit, H., and Li, H. (2007). A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In *ACL*, volume 7, pages 584–591.