Extracting an expectation-based lexicon for UD treebanks

Matteo Gay and Cristiano Chesi

IUSS, P.zza Vittoria 15, Pavia, 27100, Italy

Abstract

In this paper, we present a refinement of the deterministic lexicon extraction algorithm discussed in [1]. This new version is able to infer categorial expectations from any University Dependency (UD) treebank [2] and recast non-trivial backward dependencies in terms of movement operations [3,4]. The improvements with respect to a preliminary version of the algorithm are related to a more granular definition of the functional categories and a significant reduction in lexical ambiguity. These results might constitute a substantial baseline for (very) large language models assessment in terms of descriptive adequacy [5].

Keywords

universal dependencies, lexicon, minimalist grammars, parsing, lexical and structural ambiguity

1. Introduction

The advent of large language models (LLMs), especially those based on pre-trained transformers (GPT-like, [6]) demonstrated the impressive capability of next-token prediction task in training. This prediction is largely based on expectations of various kinds (morphosyntactic, semantic, and inferential) that, in virtue of a well-tuned attention mask mechanism can generalize categorial constraints across word sequences, apparently, without significant distance or directionality limitations among tokens, though the number of tokens accepted in input at each iteration remains fixed. This represents a clear difference with respect to recurrent networks training (e.g., LSTM, [7]) which benefits from a more cognitively plausible working memory mechanism that tries to preserve a sound incremental processing also modeling "backward dependencies" (i.e. when β depends on α, but β linearly precedes α). It has been demonstrated however that these new generation attention-based transformers are computationally very intensive [8] and their performance on various linguistic datasets demonstrates the implausibly of their linguistic generalizations [9]. In a recent study, Wilcox, Futrell & Levy [10] demonstrate the impressive ability of certain recurrent networks (JRNN and GRNN) to model islands constraints in a way that is comparable (or even better, sometimes) to that of larger LMs such as GTP-3. These results, apparently, not only challenge the poverty of stimulus hypothesis [5], but also suggest that, as far as linguistic competence is concerned, models with a smaller number of parameters (and a different architecture) might be more efficient in learning a human-like linguistic competence. What is then the minimum model size we might afford to maintain this high level of meta-linguistic competence?

In this paper, to address the "minimum model size problem", we stress the relevance of a cognitively plausible expectation mechanism (model architecture) that benefits from explicit modeling of (at least) certain categorial constraints (minimum number of parameters needed).

Expectation-based Minimalist Grammars (e-MGs, [11]), for instance, simply rely on the notion of satisfied local selection (through categorial features) to identify a successful local dependency or the requirement to be satisfied through a non-local one. The advantages are (i) the complete transparency of this model compared to a cognitively plausible grammatical theory [12] and (ii) the computational complexity cost in recognition which grows polinomially to the length of the sentence (unless specific parameterization to extend empirical coverage is adopted, [1]). A possible disadvantage of this approach is related to the fact that a fully explicit lexicon is required to run the recognition or the generation algorithm. Based on [13], this work extends the proposed algorithm to retrieve a large-scale lexicon from annotated datasets such as UD treebanks. In Section 2, we will introduce the critical aspects of this procedure, essentially focusing on the directionality of each dependency and on the inherent lexical ambiguity obtained by applying a naïf lexicon extraction procedure. In Section 3, we propose a refinement of the extraction approach, and we will calculate the efficiency of this new method in terms of reduction of lexical and syntactic ambiguity and, overall, lexicon size. Section 4 will conclude this paper by suggesting further improvements to be evaluated in the future.

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

matteo.gay@iusspavia.it (M. Gay); cristiano.chesi@iusspavia.it (C. Chesi)

0009-0003-6066-4036 (M. Gay); 0000-0003-1935-1348 (C. Chesi)

2. Categorial selection in UD

Categorial morphosyntactic selection is the standard constraint used in (e-)MGs to deal both with local (“[the dog]”) and non-local (“[what] did John [eat $_{\text{what}}$]?”) dependencies through feature-matching [11,14]; in Minimalist Grammars, both kinds of dependencies are expressed by destructive feature checking operations targeting specific features (*select* and *base* for local dependencies; *licensee* and *licensor* for non-local ones), while only local dependencies are feature-destructive in e-MGs. In both formalisms, we dub “Merge” the syntactic operation establishing a local dependency as informally expressed in **Errore. L'origine riferimento non è stata trovata.** (“X” is a categorial morphosyntactic feature expressed by β and selected/expected by α, that is $\alpha_{=\text{expect expected}}\beta$):

(1) Merge ($\alpha_{=X}$, ${}_{X}\beta$) = [$\alpha_{\cancel{=X}}$ [${}_{\cancel{X}}\beta$]]

When a non-local dependency must be established (i.e., a link between words/tokens spanning over other tokens/words, as in most wh- dependencies in languages like English or Italian), e-MGs postulate that an unsatisfied local Merge operation forces the unsatisfied categorial expectation storage into a memory buffer. This keeps the unexpected item’s features in the computation and forces its re-Merge as soon as an appropriate categorial expectation is processed, as informally expressed in (2):

(2) i. MERGE ($\alpha_{=X}$, ${}_{X\,Y}\beta$) = [$\alpha_{\cancel{=X}}$ [${}_{\cancel{X}\,Y}\beta$]]
ii. MOVE ([${}_{\cancel{X}\,Y}\beta$]) ⇒ memory([${}_{Y}(\beta)$])
iii. MERGE ($\gamma_{=Y}$, memory([${}_{Y}(\beta)$]) =
[$\alpha_{=X}$ [${}_{X\,Y}\beta$] … $\gamma_{\cancel{=Y}}$ [${}_{\cancel{Y}}(\beta)$]]

The advantage of this approach is that all non-local dependencies (of the “movement” type) are (i) forward, (ii) increase monotonically the complexity of the computation and (iii) recast syntactic ambiguity (attachment) at the lexical level. One relatively simple algorithm to extract a general-purpose lexicon from annotated UD treebanks [15] is described in [1] and can be summarized as follows.

Assuming that α is a distinct token in UD:

(i) for each distinct UPOS associated with at least one occurrence of α in UD, add a lexical entry α and associate the UPOS string to the category to be expected to license a Merge operation (*expected* feature);
(ii) for each node dependent on α, add its UPOS feature as an expectation of α (*expect* feature);
(iii) repeat (ii), and duplicate the α lexical entry, for any occurrence of α that introduces different dependents (i.e., if one occurrence of α has just one dependent X and another occurrence of α has also one extra dependent Y, then add both $\alpha_{=X}$ and $\alpha_{=X\,=Y}$ to the lexicon).

The procedures guarantee that lexical ambiguity (POS ambiguity, i.e., ${}_{X}\alpha$ and ${}_{Y}\alpha$ are both in the lexicon, α is the same token, and X and Y are two distinct UPOSs) is preserved (step i.) and that syntactic ambiguity (node attachment) is represented directly in the lexicon (step iii.). The advantage of this approach is that it makes transparent the level of morphosyntactic ambiguities a parser should deal with, distinguishing between local and non-local dependencies as well as backward and forward dependencies: each time a dependent is resolved within an adjacent item, the dependency is considered “trivial” and can be readily solved by the recognition (or the generation) algorithm discussed in [11]. If the dependency is local and “forward” (i.e. α → β, where α immediately precedes β in the text), then MERGE ($\alpha_{=X}$, ${}_{X}\beta$) = [$\alpha_{\cancel{=X}}$ [${}_{\cancel{X}}\beta$]]; if the local dependency is “backward” (i.e. β → α, where α immediately follows β in the text), then MOVE trivially applies (as in (2)), deriving in two steps [${}_{X}\beta$ [$\alpha_{\cancel{=X}}$ [${}_{\cancel{X}}(\beta)$]].

A preliminary extraction experiment following this procedure was implemented on four treebanks adopting the UPOS tagset, two for “head initial” languages [16], UD English GUM for English [17], and UD Italian ISDT [18], two for “head-final” languages, namely the UD Turkish PENN treebank [19], and the UD Japanese GSD treebank [20,21]. Although, as noticed by an anonymous reviewer, there is not yet consensus on the criteria to assign UPOS tags and this might surely affect our extraction results, UD treebanks still represent the most reliable repositories to run an automatic explorative study as the one we conducted here. The results of this extraction experiment are reported in Table 1:

Table 1
Ambiguity ratio and dependency locality in the extracted lexica English (EN), Italian (IT), Turkish (TR), and Japanese (JP)

	EN	*IT*	*TR*	*JP*
Tokens	126530	294403	166514	168333
Lexicon size (Types)	13757	27021	33036	20140
Ambiguity ratio	0.38	0.28	0.34	0.33
Lexical	*0.33*	*0.22*	*0.20*	*0.25*
Morpho.	*0.17*	*0.03*	*0.08*	*< 0.01*
Depend.	*0.50*	*0.75*	*0.72*	*0.75*
Backward depend. ratio	0.69	0.61	0.83	0.48
Locality ratio	*0.54*	*0.60*	*0.69*	*0.32*

The estimated levels of ambiguity (Lexical, i.e., POS-related; Morphological, i.e., related to a morphosyntactic featural specification such as agreement features; dependency-based, i.e., variance in terms of number and kind of dependents), as well as the amount of "backward dependencies" (i.e., those triggering movement in e-MGs), are indicated with a specification of the locality of this last kind of dependencies. We concluded that a critical issue was related to the number of “non-trivial” (i.e., non-local) backward dependencies that, especially in English, constitute a robust 15% of the lexical entries. The exploration is however incomplete at least for the following three reasons: (i) UPOS tags are relatively poor and underspecified, especially for those functional items that behave in a very different syntactic way (e.g. articles and quantifiers); (ii) the directionality of the dependencies assumed in UD is the reverse of the one presumed in other generative approaches: e.g. the determiner selects the noun in (e-

)MGs, that is, NOUN depends on DET and not the way around (this is critical, since, as predicted by e-MGs, an argument is not licensed in its thematic position unless properly determined); (iii) empty elements and multiple dependencies are absent from UD. These issues have been preliminarily addressed in [13] and expanded in the following Section.

3. A better extraction algorithm

Following the critics addressed in [22], we first inverted the original directionality of the UD dependencies between lexical and functional items, then we decorated the items extracted with specific categories and added them to the lexicon according to the following procedure in five steps:

(i) When, in UD, a functional category (e.g., a preposition, annotated as ADP) is locally dependent on a lexical one (e.g., a VERB, as in Figure 1), the first (ADP) becomes the expected feature of the functional category and the second the expectation (i.e., ADV *per* =VERB in e-MG format).

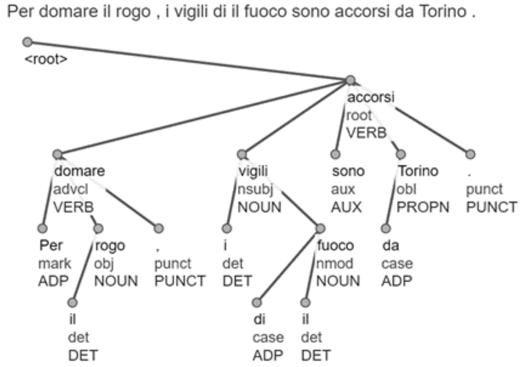

Figure 1: Example of annotation extracted from ISDT [18]

When in UD a functional category is not locally dependent on a lexical one (e.g. "del fuoco", that is "di il fuoco", "of the fire"), we preserve linear order and transform the dependencies as a left-right cascade of expectations (i.e. ADP *di* =DET, DET *il* =NOUN, NOUN *fuoco*); this way, many instances of spurious non-local dependencies are all reduced to local dependencies.

(iii) Because of the finiteness of functional categories [23–25], their universal ordering [26,27] and relative (monotonic) optionality [1] (if C depends on B which depends on A, C can directly follows A), we refine each UPOS category and modify the directionality and nature of the dependency accordingly (including phonetically empty lexical items such as null complementizers);

(iv) Phonetically empty pronominal elements (pro, PROs, [28]) are associated, and merged as empty items, directly with the verbal inflection through parameterization (in Italian, finite inflection can license an empty subject, but not in English);

(v) Unselected adjuncts (e.g., locative modifications, relative clauses) are not supposed to be "expected", so, no select feature is added to the lexical categories they modify (i.e. an optional modification is not included in the lexicon, but predicted by the e-MG implementation of Merge: in case of a relative clause modification, for instance, a raising analysis is implemented, triggered by the relative complementizer/pronoun that behaves as a DP/PP within the relative clause).

The lexicon extracted from the sentence in Figure 1 is then the following one (the format used is expected α =expect; agreement features and morphemic decomposition are ignored for the sake of compactness):

(3) e-MG extracted Lexicon = { ADP.MODE *per* =VERB,
VERB *domare* =DET, DET.DEF *il* =NOUN, NOUN *rogo*,
DET.DEF *i* =NOUN, NOUN *vigili*, ADP.ARG *di* =DET,
NOUN *fuoco*, AUX.BE *sono* =VERB.UNACC,
VERB.UNACC *accorsi* =DET,
ADP.LOCATIVE.FROM *da* =DET, DET *Torino*}

3.1. Results

The application of the algorithm to ISDT treebank produces a significant reduction of the number of non-local dependencies (-58%) as well as the overall number of non-local critical dependencies (inducing structural ambiguities, -36%) compared to the first version discussed in [1].

The results of our experiment are reported in Table 2 (sentences including token with UPOS 'INTJ', 'X', 'SYM', 'NUM' are removed):

Table 2
Ambiguity ratio and dependency locality using the revised extraction procedure in Italian (ISDT treebank)

	ISDT (original)	ISDT (revised)
Sentences processed	10616	=
Tokens/Types	193062/ 22709	=
Type/Token Ratio	0,118	=
Backward Dependencies (BD)	107382	45157
Local BD	58317	13669
Locality ratio in BD	0.543	0.303
Final Lexicon	24809	=

4. Discussion

These preliminary results are encouraging and demonstrate, on the one hand, that structural ambiguity can be reasonably reduced by assuming properly crafted linguistic generalizations, on the other hand, valency-based theories [29] are less accurate and induce an unwanted level of ambiguity that can be readily avoided.

More work needs to be done at least in the following direction: lexical items are considered as atomic entities. As demonstrated by the efficacy of word embeddings based on sub-words units (notably the byte-pair encoding, [30,31]) this step is more than necessary (and doable in the current e-MG framework, see [32]).

Back to the original consideration on how these results can serve as a baseline for LLM assessment, there is at least one possible road we consider: if explicit categorial expectations are implicitly represented in LLM, then the number of features needed in the lexicon should be considered as the minimal dimensionality needed for word embedding: as far as structural constraints are considered, this should represent the lowest number of parameters (i.e. levels of abstraction/representation) of the lexical item. Although no lexical semantic consideration is addressed here (even though the categorial approach is, in principle, tenable also from the semantic compositional perspective), specific generalization should be obtained both with e-MGs and with LLM of comparable size in terms of parameters (i.e., roughly corresponding to the number of categories in the expect(ed) required by the e-MGs lexicon). The phenomena to be tested should include islands violations (e.g. "*what did John read the book that was talking about?"), thematic selection constraints (e.g. "*John put", "*John eats a sandwich to Mary"), etc. (on the line, for instance, of [33] or [9]).

If more parameters (where "more" corresponds to at least one order of magnitude) are needed to perform similarly to e-MGs on these constraints, we might conclude that LLMs still need some "optimization" since they do not qualify (yet) as efficiently, and descriptively adequate.

Acknowledgements

This project is partially founded by BEXT (New Behavioral EXperimental approaches to complexity perception and stress assessment in linguistic and cognitive research) MIUR PRO3 2021-2023 and T-GRA2L (Testing GRAdeness and GRAmmaticality in Linguistics) PRIN 2022 (code: 202223PL4N) Projects.

References

[1] C. Chesi, Parameters of cross-linguistic variation in expectation-based Minimalist Grammars (e-MGs), IJCoL. 9 (2023) 21.

[2] J. Nivre, Ž. Agić, L. Ahrenberg, L. Antonsen, M.J. Aranzabe, M. Asahara, L. Ateyah, M. Attia, A. Atutxa, L. Augustinus, others, Universal Dependencies 2.1, (2017).

[3] E. Stabler, Two Models of Minimalist, Incremental Syntactic Analysis, Top Cogn Sci. 5 (2013) 611–633. https://doi.org/10.1111/tops.12031.

[4] N. Chomsky, The minimalist program, MIT press, Cambridge, MA, 1995.

[5] N. Chomsky, Aspects of the Theory of Syntax, MIT press, 1965.

[6] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, arXiv:2005.14165 [Cs]. (2020). http://arxiv.org/abs/2005.14165 (accessed April 21, 2021).

[7] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation. 9 (1997) 1735–1780.

[8] J. Vanian, K. Leswing, ChatGPT and generative AI are booming, but the costs can be extraordinary, CNBC News. (2023).

[9] C. Chesi, F. Vespignani, R. Zamparelli, Modelli generativi e sintassi generativa, Sistemi Intelligenti. 2 (2023).

[10] E.G. Wilcox, R. Futrell, R. Levy, Using Computational Models to Test Syntactic Learnability, Linguistic Inquiry. (2023) 1–44. https://doi.org/10.1162/ling_a_00491.

[11] C. Chesi, Expectation-based Minimalist Grammars, arXiv:2109.13871 [Cs]. (2021). http://arxiv.org/abs/2109.13871 (accessed November 2, 2021).

[12] N. Chomsky, Three factors in language design, Linguistic Inquiry. 36 (2005) 1–22.

[13] M. Gay, From Universal Dependencies to Expectation-Based Minimalist Grammars: automatic lexicon extraction and ambiguity issues., Bachelor, IUSS, 2023.

[14] E. Stabler, Computational Perspectives on Minimalism, in: C. Boeckx (Ed.), The Oxford Handbook of Linguistic Minimalism, Oxford University Press, 2011. https://doi.org/10.1093/oxfordhb/9780199549368.013.0027.

[15] M.-C. de Marneffe, C.D. Manning, J. Nivre, D. Zeman, Universal Dependencies, Computational Linguistics. 47 (2021) 255–308. https://doi.org/10.1162/coli_a_00402.

[16] M.C. Baker, The atoms of language, 1st ed, Basic Books, New York, 2001.

[17] A. Zeldes, The GUM Corpus: Creating Multilayer Resources in the Classroom, Language Resources and Evaluation. 51 (2017) 581–612. http://dx.doi.org/10.1007/s10579-016-9343-x.

[18] C. Bosco, F. Dell'Orletta, S. Montemagni, The Evalita 2014 Dependency Parsing Task, in: Proceedings of the First Italian Conference on Computational Linguistics CLiC-It 2014 and of the Fourth International Workshop EVALITA 2014 9-11 December 2014, Pisa, pisa university press, 2014. https://doi.org/10.12871/clicit201421.

[19] K. Oflazer, B. Say, D.Z. Hakkani-Tür, G. Tür, Building a Turkish Treebank, in: A. Abeillé (Ed.), Treebanks, Springer Netherlands, Dordrecht, 2003: pp. 261–277. https://doi.org/10.1007/978-94-010-0201-1_15.

[20] T. Tanaka, Y. Miyao, M. Asahara, S. Uematsu, H. Kanayama, S. Mori, Y. Matsumoto, Universal Dependencies for Japanese, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA),

Portorož, Slovenia, 2016: pp. 1651–1658. https://aclanthology.org/L16-1261.

[21] M. Asahara, H. Kanayama, T. Tanaka, Y. Miyao, S. Uematsu, S. Mori, Y. Matsumoto, M. Omura, Y. Murawaki, Universal Dependencies Version 2 for Japanese, in: N.C. (Conference chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018.

[22] T. Osborne, K. Gerdes, The status of function words in dependency grammar: A critique of Universal Dependencies (UD), Glossa: A Journal of General Linguistics. 4 (2019). https://doi.org/10.5334/gjgl.537.

[23] G. Cinque, ed., Functional structure in DP and IP, Oxford University Press, Oxford ; New York, 2002.

[24] L. Rizzi, ed., The structure of CP and IP, Oxford University Press, Oxford ; New York, 2004.

[25] A. Belletti, Structures and Beyond: The Cartography of Syntactic Structures, Volume 3, Oxford University Press, 2004.

[26] G. Cinque, Deriving Greenberg's Universal 20 and Its Exceptions, Linguistic Inquiry. 36 (2005) 315–332. https://doi.org/10.1162/0024389054396917.

[27] G. Cinque, Adverbs and functional heads: A cross-linguistic perspective, Oxford University Press, Oxford (UK), 1999.

[28] L. Rizzi, Null objects in Italian and the theory of pro, Linguistic Inquiry. 17 (1986) 501–557.

[29] L. Tesnière, Elements of structural syntax, John Benjamins Publishing Company, Amsterdam ; Philadelphia, 2015.

[30] P. Gage, A new algorithm for data compression, C Users Journal. 12 (1994) 23–38.

[31] R. Sennrich, B. Haddow, A. Birch, Neural Machine Translation of Rare Words with Subword Units, (2016). http://arxiv.org/abs/1508.07909 (accessed July 24, 2023).

[32] G.M. Kobele, Minimalist Grammars and Decomposition, in: G. Kleanthes K., E. Leivada (Eds.), The Cambridge Handbook of Minimalism, Cambridge University Press, Cambridge (UK), 2023.

[33] A. Warstadt, A. Singh, S.R. Bowman, Neural Network Acceptability Judgments, arXiv Preprint arXiv:1805.12471. (2018).