Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1000–1007
Marseille, 11–16 May 2020

ThemePro: A Toolkit for the Analysis of Thematic Progression

Mónica Domínguez[1], Juan Soler-Company[1], Leo Wanner[2,1]
University Pompeu Fabra[1], ICREA[2]
Roc Boronat, 138, Barcelona, Spain
{monica.dominguez, juan.soler, leo.wanner}@upf.edu

Abstract

This paper introduces ThemePro, a toolkit for the automatic analysis of thematic progression. Thematic progression is relevant to natural language processing (NLP) applications dealing, among others, with discourse structure, argumentation structure, natural language generation, summarization and topic detection. A web platform demonstrates the potential of this toolkit and provides a visualization of the results including syntactic trees, hierarchical thematicity over propositions and thematic progression over whole texts.

Keywords: information structure, thematic progression, theme, rheme, natural language processing, NLP, discourse structure

1. Introduction

As a rule, coherent written or spoken discourse follows a story line, which implements an interlinked theme-rheme (also referred to as *topic–focus*) sequence, such that the reader/listener can follow the development of the narration easily.[1] The linkage follows specific, theoretically well-studied, theme-rheme patterns referred to as *thematic progression* (Daneš, 1974).

Thematic progression determines the organization of the discourse and is, therefore, of great relevance to a number of natural language processing (NLP) applications. However, while the clausal theme–rheme structure, also known as Information Structure, has been used in natural language generation (Wanner et al., 2003; Ballesteros et al., 2015), text summarization (Bouayad-Agha et al., 2012) and even prosody generation (Domínguez et al., 2017), the automatic detection of thematic progression patterns in discourse, let alone the potential of thematic progression patterns in downstream applications, has not been explored yet. Among the few who have tackled the topic of thematic progression under a somewhat broader angle (and under different headings) that includes, e.g., anaphoric links, are, for instance, Kruijff-Korbayová and Kruijff (1996) and Hajičová and Mírovský (2018).

In this paper, we present an implementation that automatically labels thematic progression in English and demonstrate its functionalities through a web interface. The rest of the paper is structured as follows: Section 2 briefly lists the main trends in theoretical studies on thematic progression and introduces the fundamental concepts and terminology. The methodology deployed in the implementation and functionalities of our system are explained in Section 3. Conclusions and future work are outlined in Section 4.

[1]We follow the common interpretation that the theme–rheme/topic–focus dichotomy is a "communicative" segmentation of the meaning of an utterance into "what the utterance is about" (aka *theme* or *topic*) and "what is being uttered about it" (aka *rheme*, *focus*) (Mathesius, 1929; Daneš, 1970; Hajicová, 1986).

2. Fundamentals and Theoretical Background

Thematic progression cannot be studied without a clear underlying definition of thematicity, i.e., theme–rheme dichotomy. Since several interpretations of this dichotomy exist in the literature, we introduce first in this section the notion of thematicity we work with, and then present the thematic progression patterns that our tool recognizes. Examples are used to illustrate fundamental concepts and hopefully to make these concepts accessible to an audience who might not be acquainted with the linguistic terminology around Information Structure.

2.1. Thematicity

Thematicity is traditionally considered to be defined by *theme* (what is being talked about in a proposition) and *rheme* (what is being said about the theme); cf. (Mathesius, 1929; Firbas, 1964; Halliday, 1967). Instead of "theme–rheme" also the terms "topic–focus" are often used; cf., (Daneš, 1970; Hajicová, 1986; Lambrecht, 1994; Sgall, 2000) among others.

Mel'čuk (2001) revises the traditional theme–rheme (T–R) dichotomy and introduces a third element, the *specifier* (SP), which sets up the context of a statement. Consider, for illustration a simple example in (1):

(1) [After nine months]SP [a little boy]T [was born]R.

Mel'čuk emphasizes the hierarchical nature of thematicity: any of the three elements of a thematicity span can include further thematicity spans. The concept of hierarchical thematicity departs from the idea that thematicity is defined over propositions. A sentence may contain one or more propositions. If it contains several propositions, it is either *paratactic* (such that the propositions are at the same level, labeled as {*some text here*}P2{*and some more text here*}P3, e.g., coordination, causal sentences) or *hypotactic* (such that one proposition contains another one, labeled as {*some text here*{*that is embedded in this case*}P1.1}P1, e.g., relative clauses) as shown in example (2).[2] For the

[2]Given that a sentence may consist of several propositions and each of them possess its own thematicity pattern, we use indices

sake of simplicity, P1 propositions are not explicitly labeled, as it is implied that P1 always spans across the whole sentence. Thematicity spans and propositions are identified with square and curly brackets respectively. The label of the span is placed when the bracket is closed. Any span may also include further embedded spans; for instance, a specifier may be subdivided into theme and rheme. Consequently, both proposition and thematicity spans may contain different levels of thematicity.

(2) [After nine months]SP1 [a little boy {[who]T1 [was named Johny]R1 }P1.1]T1 [was born]R1.

In our implementation, we follow the interpretation of thematicity as defined by Mel'čuk. The conventions for annotating thematicity, explained in broad terms above, are explained in more detail in (Bohnet et al., 2013). In what follows, we will exemplify this annotation schema for the reader to have a grasp of the fundamentals of thematicity annotation in connection with thematic progression.

An example of thematicity annotation in a longer text is provided as example (3). The fragment is the beginning of the tale number 1 from the evaluation corpus described in Section 3.2.2.[3] The title and first paragraph of this tale have been included in the fragment annotated with thematicity; cf. example (3). Sentences have been numbered from 1 to 5 (S1 to S5) to make a clearer reference to these units in the text.

On top of the standard annotation of thematicity in example (3), themes are highlighted in bold. The part of the rheme that is the origin of a thematic progression thread is underlined to make a clearer connection to the simplified annotation of this fragment illustrated in example (4), which will be used to explain thematic progression in the next section. We would like to draw the readers' attention to the fact that even though themes usually coincide with the subject of a clause, such a heuristics cannot be established as a rule of thumb. The last sentence in example (3) shows a theme that does not coincide with the syntactic subject of the sentence, *for a bear*.

(3) S1 [[How]R1-1 [**Johny, the fearless Bear**]T1 , [was born]R1-2]HT .

S2 {[Today there are very few bears in the Pyrenees]R1}P2 {[but]SP1, [before]SP2, [**large numbers**]T1 [lived amongst the mountains of the range {[**that**]T1 [extends all the way from the Bay of Biscay to the Mediterranean Sea]R1}P3.1]R1}P3.

S3 [**Shepherds, charcoal burners, woodcutters and hunters**]T1 [also lived there]R1 .

S4 {[**The bears**]T1 [moved from one valley to another...]R1}P2 {[one day]SP1 [**they**]T1 [would be in the Basque Country and another in Bearn]R1}P3 {[or]SP1 [**they**]T1 [would travel from Catalonia to Aragon]R1}P4.

S5 [It was not unusual on these journeys]SP1 [**for a bear**]T1 [to get lost or be left behind]R1 .

In this short fragment, we can find examples of several thematicity spans such as: (i) split rhemes (see S1, where the rheme is split in two parts, namely R1-1 and R1-2, and the theme is located between them); (ii) different types of specifiers (SP1, SP2), e.g., S2 includes a conjunction labeled as SP1 and a temporal adverb as SP2; (iii) hypotactic propositions as P3.1 in S2; and (iv) paratactic propositions, e.g., S4 includes three paratactic propositions (P2, P3, P4).[4]

2.2. Thematic Progression

The first theoretical studies on thematic progression are attributed to Daneš (1970; 1974), who argues that thematic progression (TP) establishes the connection between sentences based on how themes are related to previous themes, rhemes or to a general theme, known as *hypertheme*. Since then, several other works on the topic have been carried out. Firbas (1971) elaborated on the idea of what he calls *communicative dynamism* – an idea also reflected by Chomsky (1982) in his statements that the order of constituents establishes the degrees of communicative dynamism and salience. Polanyi et al. (2003) studied a basic thematic progression from theme to theme and from rheme to theme. Further investigations have been carried out by Kruijff-Korbayová and Kruijff (1996) with a different name, namely "topic-focus chains", and a slightly different set of concepts, like "isotopic chains" and relationships with referential identity (co-reference) and referential non-identity. More recent works that target discourse coherence through "topic-focus" and anaphoric links include, e.g., (Hajičová and Mírovský, 2018).

In our work, we follow Daneš' definition of thematic progression (1974). Daneš establishes four types of TP (see Fig. 1 for a visual representation of this typology):

- Simple progression: the themes of adjacent sentences involve a thematization of the previous rheme.
- Progression with continuous themes: the themes of a sentence sequence are repeated, but the rhemes are different.
- Progression with derived themes from a hypertheme: the hypertheme may be a common hyperonym, a whole paragraph, a title, etc.
- A combination of the above: a rheme may be split in two concepts, and the following themes refer, at turns, to those distinctive parts of the rheme.

There are two types of thematic progression, within and across sentences. Both of them are relevant to the analysis of thematic progression, but it must be underlined that thematic progression within a sentence is more closely related to the syntactic structure.

to identify all the elements 'SP1': specifier 1, 'T1': theme 1, 'R1': rheme 1, etc. Note, however, that only one theme and one rheme are admitted in each proposition, even if they may be split.

[3]This tale as well as all the annotated texts used in this paper are available at `https://github.com/joanSolCom/ThemePro`.

[4]P1 is understood to cover the whole sentence, that is the reason why the first paratactic proposition starts in P2.

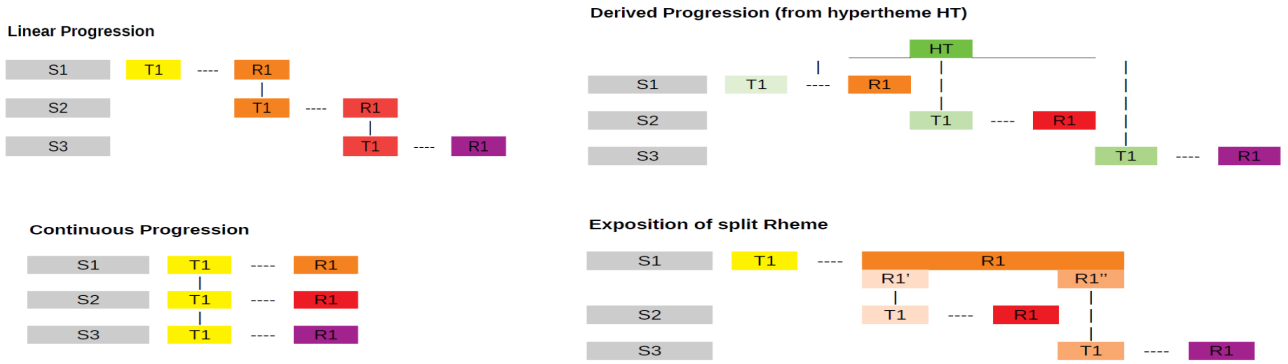

Figure 1: Types of Thematic Progression

Example (4) provides a simplified annotation scheme derived from example (3), where only themes are annotated. We specify what proposition they belong to (e.g., *large numbers* is the theme of P3, labeled as T1-P3). Sentences have been numbered excluding the title, and thematicity span numbers coincide with sentence numbers, that is, T2 is the theme in the second sentence from the example (S2). The title has not been numbered because it coincides with the hypertheme (HT), and thus has its own label (TT).

(4) TT [How [**Johny, the fearless Bear**]T1 , was born.]HT

S1 Today there are *very few bears* in the Pyrenees but, before, [**large numbers**]T1-P3 lived amongst the mountains of *the range* [**that**]T1-P3.1 extends all the way from the Bay of Biscay to the Mediterranean Sea.

S2 [**Shepherds, charcoal burners, woodcutters and hunters**]T2 also lived there.

S3 [**The bears**]T3-P2 moved from one valley to another... one day [**they**]T3-P3 would be in the Basque Country and another in Bearn or [**they**]T3-P4 would travel from Catalonia to Aragon.

S4 It was not unusual on these journeys [**for a bear**]T4 to get lost or be left behind.

The first sentence (S1) consists of two paratactic propositions (namely P2 and P3) and a hypotactic proposition in P3, labeled as P3.1; cf. example (3). S1 starts with a rhematic proposition (P2), which does not contain a theme.[5]. The theme of P3 (T1-P3), *large numbers*, is a derived progression of part of the rheme in P2, namely, from *very few bears*, T1-P3 progresses to *large numbers (of bears)*. Then, T1-P3.1 coincides with the relative pronoun *that* referring to part of the rheme in P3 *the range (of mountains)*.
Regarding the type of thematic progression, we could summarize the thematic progression pattern in S1 as derived from part of the rheme (DR) both between P2 and P3 and between P3 and P3.1. A schematic representation of the-matic progression in S1 is illustrated in Figure 2. In terms of TP typology, there are two progressions of type "derived from rheme" (DR).

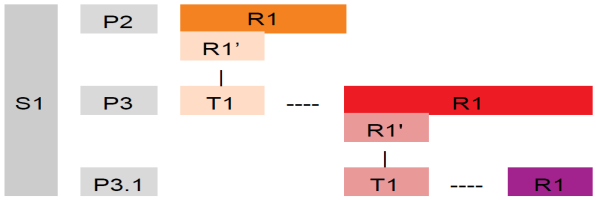

Figure 2: Thematic progression scheme in sentence 1 from example (4).

In S2, a new theme (NT) is introduced. However, the theme in S3, *The bears* refers back to T1-P3. In other words, there is a continuous thematic progression between non-adjacent sentences. We call this typology "continuous progression with gap" (CG) following (Hawes, 2015).[6]
S3 includes three paratactic propositions, namely P2, P3 and P4. Their thematic progression is continuous (C), that is, T3-P3 (*they*) progresses from T3-P2 (*The bears*), and T3-P4 (also *they*) progresses from the previous theme T3-P3. Finally, the theme in sentence 4, *a bear*, is a derived progression from the last theme (DT) in S3, which referred to *the bears*.
All in all, the thematic progression pattern of the whole paragraph with respect to TP typology follows the scheme below. There are two themes derived from rhemes (DR), a new theme (NT), a continuous gap (CG), two continuous progressions (C) and a final theme derived from the previous theme (DT).

DR -¿ DR -¿ NT -¿ CG -¿ C -¿ C -¿ DT

In order to establish the thematic progression for a given text automatically by our system *ThemePro*, we identify and label the thematicity spans in each sentence using a thematicity parser (cf. Section 3.2) and then, we establish the thematic progression pattern of the identified themes following the methodology outlined in Section 3.1.

[5]By definition, existential propositions are all-rhematic, i.e., do not have a theme.

[6]Hawes (2015) argues that Daneš ignored thematic progressions which are not adjacent, and proposed the idea of breaks and constraints to facilitate the empirical analysis of thematic progression in texts written by learners of English as a second language.

3. ThemePro: Description of Functionalities

This section explains the methodology applied in the implementation of ThemePro in Section 3.1 and the functionalities of the main two components of our framework, namely, the thematicity parser in Subsection 3.2 and the visualization capabilities of ThemePro in Subsection 3.3.

3.1. Methodology

The process for the determination of thematic progression patterns in a given text consists of the following steps:

1. The SpaCy parser (Honnibal and Montani, 2017) is used to derive the universal dependency (UD) syntactic structures of the sentences of the input text.
2. The resulting syntactic structures are converted into CoNLL structures,[7] which includes six columns for identifier (id), word, lemma, part of speech (POS), dependency function and dependency relation.
3. The thematicity structure is predicted for each sentence using a thematicity parser. This dimension is introduced as the seventh column in the CoNLL file.
4. The thematic progression is computed selecting a set of contiguous spans and a hypertheme to establish the TP of a given theme labeled as such by the thematicity parser. Theme spans may include open class words or pronouns (personal pronouns, demonstrative pronouns, relative pronouns, etc.). Two different strategies are devised to deal with each of these categories, namely:
   (a) Open class words: a word2vec representation (using Glove word embeddings (Pennington et al., 2014) [8]) is derived for each word of the span, and its centroid is computed. The cosine similarity between the current span and the candidate reference spans (hypertheme, previous theme and previous rheme, if exists) are computed and the distance is displayed in the graph representation of the thematic progression. In the future, we plan the use of distance thresholds to prune themes and to only link closely-related elements.
   (b) Pronouns (personal and demonstrative pronouns) are treated with co-reference resolution (using the SpaCy extension `neuralcoref`.[9]) Only candidate spans are included in the set of possible co-referent words. The referents of relative pronouns are derived from the syntactic tree obtained with SpaCy.

This methodology is currently tested on English, yet the required changes to cover other languages are few. To adapt our methods to a new language, a UD-based model and word embeddings for this new language would need to be used. Since plenty of UD-based parsing and word embedding models are available, such an adaptation would be straightforward.

The web service is able to show an overall display of long texts that can give a first impression of the thematic progression pattern. As future work, ThemePro will be extended to label the type of thematic progression and export the annotated text in a machine-readable format.

3.2. Thematicity Parser

Based upon Firbas (1971), recent work on the automatic identification of theme and rheme has also been carried out by Pala and Svoboda (2014). The main concern with respect to this work is that their theme–rheme tagger has been developed and fine-tuned based on a very limited corpus of 120 sentences in Czech. Unfortunately, the code is not available to compare our tool with it. The only available tool for comparison was (Bohnet et al., 2013)'s parser, which has been used as a baseline to evaluate our parser in ThemePro.

A rule-based approach has been implemented, which has proved to successfully cope with the deficiencies spotted in (Bohnet et al., 2013)'s parser as we will explain in Section 3.2.2. We expect that our implementation is scalable to several registers and languages, and will certainly contribute to saving an important amount of annotation time and effort. In the following sections, we sketch the implementation and compare our thematicity parser with the baseline by Bohnet et al. (2013).

3.2.1. Implementation

The functions that have been developed are based on a preliminary implementation presented in (Domínguez et al., 2018) for application in thematicity-based prosody generation. While the preliminary implementation searched through the CoNLL input structure, the present implementation includes a more advanced and efficient search technique using tree data structures. Hence, a library parses the CoNLL and converts it into a tree data structure. Based on this data structure, a strategy is defined to traverse the dependency tree and find out what type of syntactic structure it contains, for instance, coordination, subordination, simple sentence, etc. or a combination of syntactic structures. This implies a big progress in comparison with the implementation in (Domínguez et al., 2018), which was only able to annotate simple and coordinated sentences.

Once the sentence typology is detected, heuristics to detect propositions and thematicity structure are applied, based on the correspondence between thematicity and morphology and syntax. Then, writing functions are called to annotate each span, as explained in Section 2.1. In the guidelines by (Bohnet et al., 2013), hypotactical propositions were meant to be annotated with different numbers (i.e., {{}P3}P2). We will see that the parser did actually ignore propositions. To improve on this functionality, our implementation annotates propositions with several digits separated with dots, depending on the level of embeddedness in the parent proposition (i.e., {{}P2.1}P2).

A CoNLL file is generated by this module with a final col-

[7] Details about the CoNLL format are provided in `http://universaldependencies.org/docs/format.html`

[8] Contextual embeddings were tested, but made the demo much slower and less usable.

[9] `https://github.com/huggingface/neuralcoref`

umn for thematicity annotation. The web service does not show the CoNLL, but it is available for download upon request.

3.2.2. Evaluation

Given the central role of the thematicity parser in ThemePro, we evaluated its performance using Bohnet's parser (2013) as a baseline. As evaluation corpus, three short tales in English have been chosen from a multilingual web resource.[10] Table 2 shows the overall characteristics of the evaluation corpus. The texts are children tales and include several parts of direct speech. In terms of grammatical complexity, syntax is rather simple in these tales. All texts have been manually annotated by three expert linguists, who have reached a consented gold standard after several rounds of discussion. These texts were not used to fine-tune the parser for evaluation run time.

	Total number of samples
Words	1,312
Sentences	92
Paragraphs	9
Paratactical Propositions	43
Hypotactical Propositions	22
Direct speech	33

Table 1: Description of evaluation corpus

For evaluation, we have used the same metrics as in (Bohnet et al., 2013), that is, accuracy score (AS), unlabeled bracket score (UBS) and labeled bracket score (LBS) and we have, likewise, considered all words in the corpus to compute these metrics. The accuracy score (AS) includes all words in the text, that is, if the output of the parser matches exactly the gold standard, the AS would be 1. The unlabeled bracket score (UBS) considers only those words which have a bracket (either open or closed bracket). In other words, UBS assesses how accurate the parser is in detecting thematicity spans. Finally, the labeled bracket score (LBS) takes into account those words that get a thematicity label, i.e., one word per thematicity span. Thus, LBS assesses the parser's accuracy to detect thematicity as such.

Some deficiencies were observed in the output of Bohnet et al. (2013)'s parser, namely, it does not annotate propositions and it does not output number labels. In order to make a fair comparison with our parser, we have discarded proposition labels and number labels from our annotation when running the evaluation. Otherwise, the labeled bracket score (LBS) would have been zero for the baseline.

	AS	UBS	LBS
Baseline	0.62	0.40	0.44
ThemePro's Parser	**0.74**	**0.53**	**0.54**

Table 2: Evaluation results: accuracy score (AS), unlabeled bracket score (UBS) and labeled bracket score (LBS)

[10] http://amarauna.org/nabar/en/

Results show that ThemePro's parser outperforms the baseline in all scores, obtaining between 0.10 and 0.12 points more. We would like to highlight the fact that the baseline parser's output has some deficiencies as we mentioned above. One major inconvenience of this output is that brackets are often inconsistently opened and closed, whereas ThemePro's parser makes sure that bracket pairs are balanced. Another important issue is regarding propositions spans, which are not labeled (as already mentioned) in (Bohnet et al., 2013)'s parser, whereas ThemePro's parser includes a fine-grained annotation distinguishing paratactic and hypotactic propositions.

These deficiencies that have been detected and corrected in ThemePro are crucial for the prediction of thematic progression. On the one hand, the fact that brackets are consistently opened and closed guarantee that the module that computes centroids using word embeddings is selecting the right set of words. An inconsistency in the brackets would be equivalent to unmatched tags in the xml format or unbalanced curly braces in json. If the basic building blocks of thematic progression (namely, theme/rheme spans) are incorrectly detected, the thematic progression cannot be computed.

On the other hand, the consequence of not labeling propositions for thematic progression is that the only available unit for segmentation would be the sentence. This implies that complex sentences containing several theme spans, which usually involve a varied range of thematic progressions, would be neglected.

All in all, results of the evaluation of the parser in ThemePro are more than what is shown by the quantitative analysis of scores: major functionalities to detect thematic progression would be seriously affected if the baseline parser would have been used, thus leading to unfortunate malfunctioning of the whole pipeline.

3.3. Visualization Capabilities in ThemePro

ThemePro is a web interface where any text (in English for now) can be fed as input via the main menu (see Figure 3). The backend described in Subsection 3.1 is automatically run upon clicking on the "Submit" button. This main menu may be hidden or may stay in the upper part of the screen by pressing the "Hide/Show" button on the upper left corner at any time.

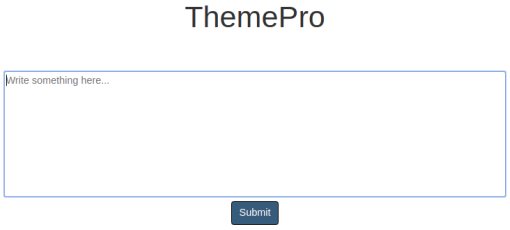

Figure 3: ThemePro's main menu.

Once the backend has terminated all processes, four tabs appear on the screen to visualize syntactic trees, thematicity, co-reference chains and thematic progression. Clicking on each tab, the output of each module is displayed in a user-friendly design.

Syntactic trees are shown sentence by sentence (see Figure 4). If users wish to visualize another sentence, they can click on the left/right arrows provided on screen. Each POS tag is assigned a different color. Figure 4 shows the syntactic tree for the title of tale 1 also used in example (3): *How Johny, the fearless bear, was born*.

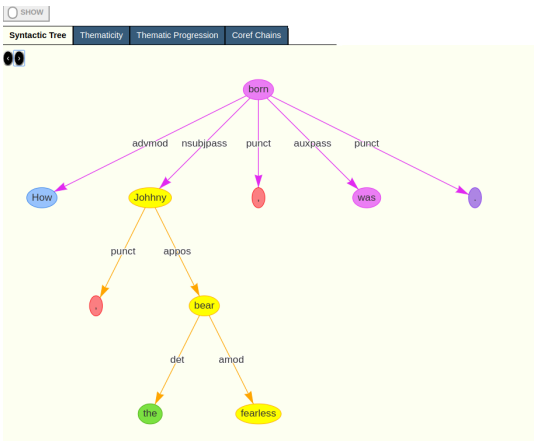

Figure 4: Syntactic tree visualization.

The output of the thematicity parser is displayed in ordered sentences (see Figure 5). Propositions and thematicity at different levels are highlighted in colors.

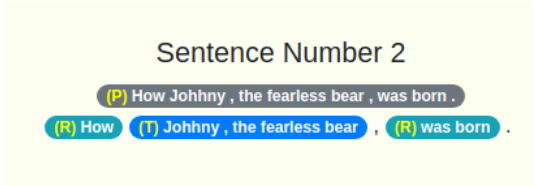

Figure 5: Thematicity visualization.

Given that co-reference resolution is used to predict one part of thematic progression (as already mentioned in the Methodology Section), the visualization also includes co-reference chains (see Figure 6).

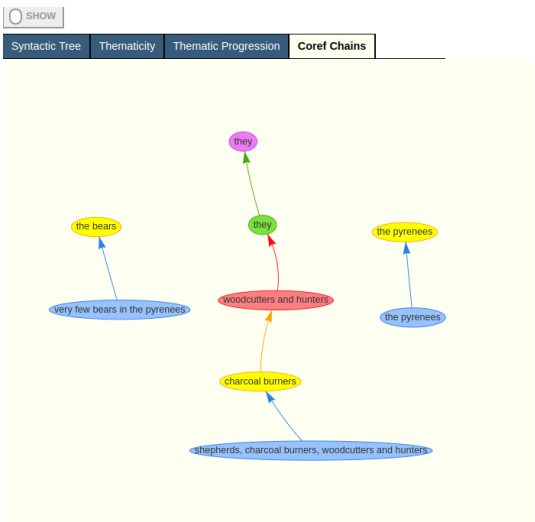

Figure 6: Co-reference chains visualization.

The web demonstration displays thematic progression as a graph representation (see Figure 7). We designed the visualization as a graph to be able to demonstrate longer and more complex texts than the small paragraph provided in example (3). The graph in Figure 7, for instance, represents tale 1 from our evaluation corpus. The graph shows how themes are related to different units in the text, namely, previous themes, the first sentence (established as hypertheme) or previous rhemes. Thus, new themes are separated from themes which are related through a thematic progression pattern.
Arrows point to the span each theme is related to. If spans include content words, a distance score is shown in the arrow. If the relation was found through co-reference resolution, only the arrow (with no score) is visible. The text included in each span is shown in each node's tooltip.

4. Conclusions

This paper presents ThemePro, an operational toolkit for the determination of thematic progression in English texts. The toolkit is demonstrated as a web interface.[11]
From the methodological point of view, the main contribution consists in the adaptation of Daneš's typologies of thematic progression, which includes the annotation of non-adjacent sentences. We have made this decision after trying to annotate different types of texts. These types, which are proposed below, adapt better to an empirical analysis setup and include:

- Simple progression (S): a theme in S_n is a progression from the rheme in S_{n-1}.
- Continuous progression (C): a theme in S_n is a progression from the immediately previous theme in S_{n-1}.
- Continuous gap (CG): a theme in S_n is a progression from the sentence before the immediately previous theme, that is, from S_{n-2}.
- Derived progression: a theme from S_n is derived from the hypertheme (DHT) or from the previous theme (DT) or rheme (DR) in S_{n-1}.

All in all, ThemePro contributes in several aspects to the state of the art, namely: (i) a formal description of hierarchical thematicity is used which has been previously tested in other areas of NLP; (ii) a methodology for automatic analysis of thematic progression is introduced; and (iii) a visually-friendly platform demonstrates thematic progression patterns in long texts as a graph representation. In a nutshell, ThemePro pivots the transition from theoretical to applied work on thematic progression. It is a steady first step that allows further insights in the field of Information Structure using the text rather than the sentence as a unit. Moreover, the presented framework provides a starting point for the development of other applications that may use thematic progression as a high-level linguistic feature

[11] All code, evaluation corpus and web interface together with a demonstration video is accessible under a GNU License v.3 on the Github repository `https://github.com/joanSolCom/ThemePro`.

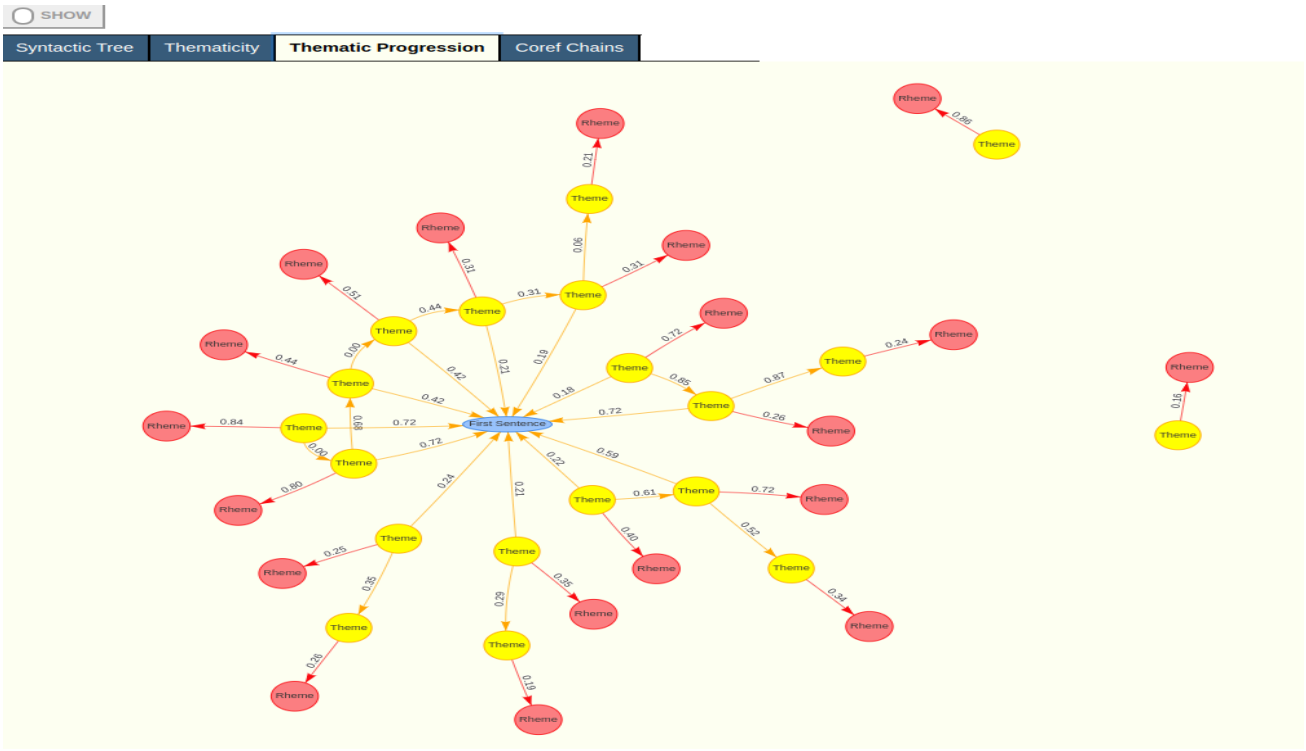

Figure 7: Thematic Progression visualization.

to analyze discourse structure. Thematic progression has a great potential for exploitation in NLP and this tool is meant to foster interest for such potential within the research community.

In our ongoing work, ThemePro is aimed for the analysis of how thematic progression patterns affect speech prosody in monologue discourse, so that more expressive voices can be obtained using text-to-speech (TTS) applications to avoid monotony when reading longer stretches of texts. We are also using the toolkit for the analysis of speaker dominance in chats. More precisely, we explore to what extent thematic progression can help distinguish the argumentation patterns of speakers in the context of harassing behaviour in social media and thus contribute to the identification of speakers who dominate (and thus guide) the interaction.

The work presented in this paper involves some aspects that still remain unexplored. Firstly, the evaluation of the thematicity parser needs to be extended to different registers and languages, and a methodology needs to be devised for the evaluation of thematic progression as such. This requires gathering a reasonable amount of annotated data and establishing metrics, which serve to assess our methodology. Another limitation is the automatic annotation and visualization of TP typologies. We are currently working on this aspect, and we hope to implement an extension of ThemePro in the recent future.

5. Acknowledgements

This work is part of the WELCOME project, which has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under the Grant Agreement number 870930.

6. References

Ballesteros, M., Bohnet, B., Mille, S., and Wanner, L. (2015). Data-driven sentence generation with non-isomorphic trees. In *Proceedings of the Annual Conference of the North American Association for Computational Linguistics – Human Language Technologies (NAACL – HLT)*.

Bohnet, B., Burga, A., and Wanner, L. (2013). Towards the annotation of penn treebank with information structure. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1250–1256, Nagoya, Japan.

Bouayad-Agha, N., Casamayor, G., Mille, S., and Wanner, L. (2012). Perspective-Oriented Generation of Football Match Summaries: Old Tasks, New Challenges. *ACM Transactions on Speech and Language Processing*, 9(2).

Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding*. MIT Press, Cambridge, MA.

Daneš, F. (1970). One instance of Prague School methodology: Functional analysis of utterance and text. *Garvin*, pages 132–141.

Daneš, F. (1974). Functional sentence perspective and the organization of the text. In *Papers on Functional Sentence Persepctive*, pages 106–128. Mouton, The Hague.

Domínguez, M., Farrús, M., and Wanner, L. (2017). A thematicity-based prosody enrichment tool for cts. In *Proceedings of INTERSPEECH: show and tell demonstrations*, pages 3421–2, Stockholm, Sweden.

Domínguez, M., Burga, A., Farrús, M., and Wanner, L. (2018). Towards expressive prosody generation in tts for reading aloud applications. In *Proceedings of Iber-*

SPEECH, pages 40–44, Barcelona, Spain. International Speech Communication Association (ISCA).

Firbas, J. (1964). On defining the theme in functional sentence perspective. In *Travaux linguistiques de Prague*, pages 1267–1280.

Firbas, J. (1971). On the concept of communicative dynamism in the theory of functional sentence perspective. In *SPFFBU*, pages 135–144.

Hajicová, E. (1986). Focussing- A Meeting Point Linguistics and Artificial Intelligence. In *Artificial Intelligence II: Methodology, Systems, Applications - Proceedings of the Second International Conference on Artificial Intelligence: Methodology, Systems, Applications, AIMSA 1986, Varna, Bulgaria, September 16-19, 1986*, pages 311–321.

Hajičová, E. and Mírovský, J. (2018). Discourse Coherence Through the Lens of an Annotated Text Corpus: A Case Study. In *Proceedings the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1637–1642. Miyazaki, Japan.

Halliday, M. (1967). Notes on Transitivity and Theme in English, Parts 1-3. *Journal of Linguistics*, 3(1):37–81.

Hawes, T. (2015). Thematic progression in the writing of students and professionals. *Ampersand*, 2:93 – 100.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Kruijff-Korbayová, I. and Kruijff, G. M. (1996). Identification of topic-focus chains. In *Approaches to Discourse Anaphora: Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC96)*, pages 165–179.

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus and the mental representations of discourse referents*. Cambridge University Press, Cambridge.

Mathesius, V. (1929). Zur Satzperspektive im modernen Englisch. In *Archiv für das Studium der neueren Sprachen und Literaturen*, volume 155, pages 202–210. Erich Schmidt Verlag.

Mel'čuk, I. A. (2001). *Communicative Organization in Natural Language: The semantic-communicative structure of sentences*. Benjamins, Amsterdam, Philadephia.

Pala, K. and Svoboda, O. (2014). An Experiment with ThemeâRheme Identification. In *Text, Speech and Dialogue: 17th International Conference, TSD 2014*, pages 275–284, Brno, Czech Republic. Springer International Publishing.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP*.

Polanyi, L., Van Den Berg, M., and Ahn, D. (2003). Discourse Structure and Sentential Information Structure: An Initial Proposal. *Journal of Logic, Language, and Information*, 12(3):337–350.

Sgall, P. (2000). Functional sentence perspective in written and spoken communication. Studies in English Language.

Wanner, L., Bohnet, B., and Giereth, M. (2003). Deriving the Communicative Structure in Applied NLG. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*.