Segmentation of patent claims for improving their readability

Gabriela Ferraro[1 2], **Hanna Suominen**[1−4], **Jaume Nualart**[1 3]
[1]NICTA / Locked Bag 8001, Canberra ACT 2601, Australia
[2]The Australian National University
[3]University of Canberra
[4]University of Turku
firstname.lastname@nicta.com.au

Abstract

Good readability of text is important to ensure efficiency in communication and eliminate risks of misunderstanding. Patent claims are an example of text whose readability is often poor. In this paper, we aim to improve claim readability by a clearer presentation of its content. Our approach consist in segmenting the original claim content at two levels. First, an entire claim is segmented to the components of preamble, transitional phrase and body, using a rule-based approach. Second, a conditional random field is trained to segment the components into clauses. An alternative approach would have been to modify the claim content which is, however, prone to also changing the meaning of this legal text. For both segmentation levels, we report results from statistical evaluation of segmentation performance. In addition, a qualitative error analysis was performed to understand the problems underlying the clause segmentation task. Our accuracy in detecting the beginning and end of preamble text is 1.00 and 0.97, respectively. For the transitional phase, these numbers are 0.94 and 1.00 and for the body text, 1.00 and 1.00. Our precision and recall in the clause segmentation are 0.77 and 0.76, respectively. The results give evidence for the feasibility of automated claim and clause segmentation, which may help not only inventors, researchers, and other laypeople to understand patents but also patent experts to avoid future legal cost due to litigations.

1 Introduction

Clear language is important to ensure efficiency in communication and eliminate risks of misunderstanding. With written text, this clarity is measured by readability. In the last years, we have witnessed an increasing amount work towards *improving text readability*. In general, these efforts focus on making general text easier to understand to non-native speakers and people with special needs, poor literacy, aphasia, dyslexia, or other language deficits.

In this paper, we address making *technical text* more readable to *laypeople*, defined as those without professional or specialised knowledge in a given field. Technical documentation as scientific papers or legal contracts are two genres of written text that are difficult to understand (Alberts et al., 2011). An extreme example that takes the worst from both these worlds is the *claim section of patent documents*: it defines the boundaries of the legal protection of the invention by describing complex technical issues and using specific legal jargon (Pressman, 2006). Moreover, due to international conventions, each patent claim must be written into a single sentence. This leads to very long sentences with complex syntactic structures that are hard to read and comprehend not only for laypeople but also for technical people who are not trained to read patent claims.

As an example of other efforts with similar goals to improve the readability of technical text to laypeople, we mention the CLEF eHealth shared tasks in 2013 and 2014 (Suominen et al., 2013). However, instead of inventors, researchers, and other claim readers, they target patients and their next-of-kins by developing and evaluating technologies to improve the readability of clinical reports and help them in finding further information related to their condition in the Internet.

Some proposals have also been made in order to improve claim readability, for example, by applying simplification, paraphrasing, and summarisation methods (see Section 2). However, these approaches *modify the claim content*. This increases

Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014, pages 66–73,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

the risk of changing also the meaning, which is not desirable in the context of patent claims and other legal documents.

In this paper, we propose an alternative method that focuses on *clarifying the presentation* of the claim content rather than its modification. Since readability strongly affects text comprehension (Inui et al., 2003), the aim of this study is to make the content of the patent claims more legible and consequently make them easier to comprehend.

As the first steps towards this improved presentation of the patent claims, we propose to *segment the original text*. Our approach is data driven and we perform the segmentation at two levels: first, an *entire claim* is segmented *into three components* (i.e., preamble, transition, and body text) and second, the components are further segmented *into clauses*. At the first level, we use a *rule-based* method and at the second level, we apply a *conditional random field*.

We evaluate segmentation performance *statistically* at both levels and in addition, we analyse errors in clause segmentation *qualitatively*; because our *performance at the first level is almost perfect* (i.e., for detecting the beginning and end of the preamble, the accuracy percentages are 100 and 97 and these numbers are 94 and 100 for the transition and 100 and 100 for the body text), we focus on the errors at the second level alone. In comparison, we have the precision of 77 per cent and recall of 76 per cent in clause segmentation. Even though this performance *at the second level* is not perfect, it is *significantly better than* the respective percentages of 41 and 29 (0.2 and 0.3) for *a baseline* based on both punctuation and keywords (punctuation only).

The rest of the paper is organised as follows: Section 2 describes as background information of this study includes an explanation about what patent claims are, how to read them, and what kind of related work exists on claim readability. Section 3 outlines our materials and methods. Section 4 presents the experiments results and discussion. Finally, conclusions and ideas for future work are presented in Section 5.

2 Background

2.1 Patent claims

Patent documents have a predefined document structure that consists of *several sections*, such as the title, abstract, background of the invention, description of the drawings, and claims. As already mentioned, the claims can be seen as the most important section as they define the scope of legal protection of the invention. In most modern patent laws, patent applications must have at least one claim (Pressman, 2006).

> [Toolholder]$_p$, [comprising]$_t$ [a holder body with an insert site at its forward end comprising a bottom surface and at least one side wall where there projects a pin from said bottom surface upon which there is located an insert having a central bore, a clamping wedge for wedging engagement between a support surface of the holder and an adjacent edge surface of said insert and an actuating screw received in said wedge whilst threadably engaged in a bore of said holder, said support surface and said edge surface are at least partially converging downwards said wedge clamp having distantly provided protrusions for abutment against the top face and the edge surface of said insert, characterised in that the wedge consists of a pair of distantly provided first protrusions for abutment against a top face of the insert, and a pair of distantly provided second protrusions for abutment against an adjacent edge surface]$_b$.

Figure 1: An example patent claim. We have used brackets to illustrate claim components and the sub-scripts *p*, *t*, and *b* correspond to the preamble, transition, and body text, respectively.

The claims are written into a *single sentence* because of international conventions. Figure 1 provides an example claim.

Furthermore, a claim should be composed by, at least, the following parts,

1. *Preamble* is an introduction, which describes the class of the invention.

2. *Transition* is a phrase or linking word that relates the preamble with the rest of the claim. The expressions *comprising*, *containing*, *including*, *consisting of*, *wherein* and *characterise in that* are the most common transitions.

3. *Body text* describes the invention and recites its limitations.

We have also included an illustration of these claim components in Figure 1.

Because a claim is a single sentence, special *punctuation conventions* have been developed and are being used by patent writers. Modern claims follow a format where the preamble is separated

Table 1: Per claim demographics

		Training set	Test set
# tokens	mean	60	66
	min	7	8
	max	440	502
# boundaries	mean	5	5
	min	1	1
	max	53	41

from the transition by a comma, the transition from the body text by a colon, and each invention element in the body text by a semicolon (Radack, 1995). Other specifications regarding punctuation are the following text elaboration and element combination conventions:

- A claim should contain a period only in the end.
- A comma should be used in all natural pauses.
- The serial comma[1] should be used to separate the elements of a list.
- Dashes, quotes, parentheses, and abbreviations should be avoided.

Because a claim takes the form of a single sentence, long sentences are common. Meanwhile, in the general discourse (e.g., news articles) sentences are composed of twenty to thirty words, claim sentences with over a hundred words are very frequent (see, e.g., Table 1 related to materials used in this paper). As a consequence, claims usually contain several *subordinate and coordinate clauses*, as they enable the aforementioned elaboration and the combination of elements of equal importance, respectively.

As claims are difficult to read and interpret, several books and tutorials suggest how claims should be *read* (Radack, 1995; Pressman, 2006). The first step towards reading a claim is to identify its components (i.e., preamble, transition, and body text). Another suggestion is to identify and highlight the different elements of the invention spelled out in the body text of the claims.

The clear punctuation marks and lexical markers enable the claim component segmentation, as suggested above. Moreover, the predominance of intra-sentential syntactic structures (e.g., subordinate and coordinate constructions) favours segmenting patent claims into clauses. These clauses can then be presented as a sequence of segments which is likely to improve claim readability.

2.2 Related work

So far, not many studies have addressed the problem of improving the readability of patents claims. In particular, to the best of our knowledge, there is no research that specifically targets the problem of presenting the claims in a more readable layout. Consequently, we focus on efforts devoted to claim readability in general with an emphasis on text segmentation.

We begin by discussing three studies that address *text simplification in patent claims*. Note that these approaches modify the claim content which may also change their meaning. This is riskier in the context of patent documents and other legal text than our approach of clarifying the presentation. Moreover, in order achieve a reasonable performance, the methods of these studies require sophisticated tools for discourse analysis and syntactic parsing. Usually these tools also need to be tailored to the genre of claim text.

First, a *parsing methodology* to simplify sentences in *US patent documents* has been proposed (Sheremetyeva, 2003). The resulting analysis structure is a syntactic dependency tree and the simplified sentences are generated based on the intermediate chunking structure of the parser. However, neither the tools used to simplify sentences nor the resulting improvement in readability has been formally measured.

Second, simplification of *Japanese claim sentences* has been addressed through a rule-based method (Shinmori et al., 2003). It identifies the *discourse structure* of a claim using cue phrases and lexico-syntactic patterns. Then it *paraphrases* each discourse segment.

Third, a claim simplification method to *paraphrase* and *summarise* text has been introduced (Bouayad-Agha et al., 2009). It is *multilingual* and consists of claim segmentation, coreference resolution, and discourse tree derivation. In claim segmentation, a rule-based system is compared to machine learning with the conclusion of

[1] The serial comma (also known as the Oxford comma) is the comma used mediately before a coordination conjunction (e.g., *CDs, DVDs, and magnetic tapes* where the last comma indicates that *DVDs* and *magnetic tapes* are not mixed). http://oxforddictionaries.com (accessed 28 Feb, 2014)

the former approach outperforming the latter. The machine learning approach is, however, very similar to the clause segmentation task described in this paper. They differ in the features used to characterized the clause boundaries and in evaluation. For the evaluation, these authors use the cosine similarity to calculate a 1:1 term overlap between the automated solution and gold standard set whereas we assess whether a token is an accurate segment boundary or not.

We continue by discussing a *complementary method* to our approach of improving the readability of claims through their clearer presentation without modifying the text itself. This work by Shinmori et al. (2012) is inspired by the fact that claims must be understood in the light of the definitions provided in the description section of the patents. It aims to enrich the content by *aligning claim phrases with relevant text from the description section*. For the evaluation, the authors have inspected 38 patent documents. The automated method generates 35 alignments for these documents (i.e., twenty correct and fifteen false) and misses only six. It would be interesting to test if this alignment method and the claim segmentation proposed in this paper complement each other.

We end by noting that the task of segmenting claim phrases is similar to the task of *detecting phrase boundaries* by Sang and Déjean (2001) in the sense that the segments we want to identify are intra-sentential. However, the peculiar syntactic style of claims makes the phrase detection strategies not applicable (see Ferraro (2012) for a detailed study on the linguistic idiosyncrasy of patent claims).

3 Materials and methods

In this paper, we performed *statistical experiments* and *qualitative error analyses* related to two segmentation tasks (see Figure 2):

1. Segmenting claims section to the components for preamble, transition, and body text.

2. Segmenting each claim to subordinate and coordinate clauses.

For Task 1, we developed a *rule-based method* using the *General Architecture for Text Engineering* (GATE) (Cunningham et al., 2011). The system had three rules, one for each of the claim parts we were interested in identifying. The rules were written in terms of JAPE grammars.[2] In order to process the rules, the GATE pipeline illustrated in Figure 3 was applied. Because transitions should match with the first instance of a closed set of keywords (we used *comprise*, *consist*, *wherein*, *characterize*, *include*, *have*, and *contain*), our first rule identified a transition and, using its boundary indices, we restricted the application of our further rules. This resulted in the following application order:

transition ⟶ preamble ⟶ body text.

Table 2: Dataset demographics

	# claims	# segments	# words
Training set	811	4397	48939
Development set	10	15	260
Test set	80	491	5517

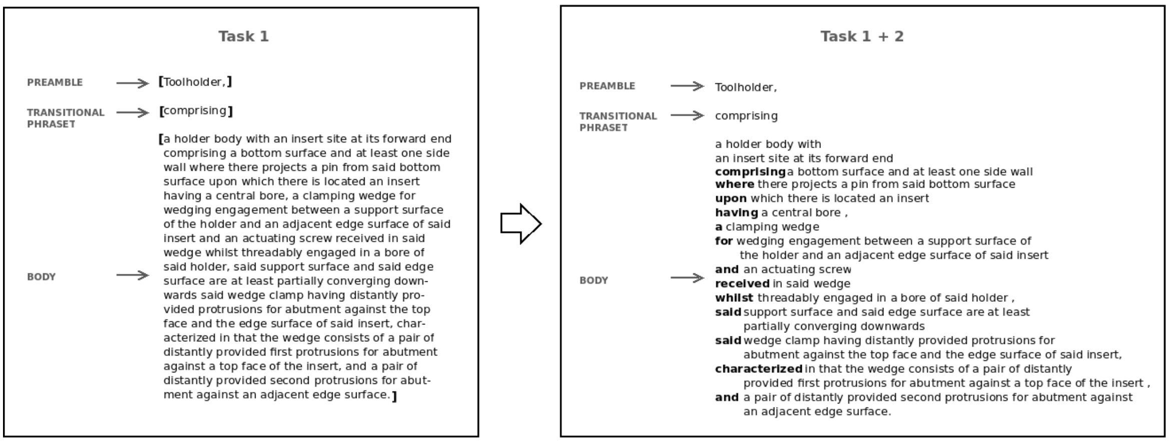
Our two other rules relied on lexico-syntactic patterns and punctuation marks. Note that even though punctuation conventions have been developed for claim writing (see Section 2.1), their following is not mandatory. This led us to experiment these more complex rules. The first task was applied to the complete dataset (training, development, and test sets merged into one single dataset) described in Table 2.

For Task 2, our method was based on *supervised machine learning* (ML). To train this ML classifier, we used a patent claim corpus annotated with clause boundaries. This corpus was provided by the TALN Research Group from Universitat Pompeu Fabra. The aim of the segmentation classifier was to decide whether a claim token is a segment boundary or not, given a context. Thus, every token was seen as a candidate for placing a segment boundary. Following standard ML traditions, we split the dataset in *training*, *development*, and *test sets* (Tables 2 and 1).

The corpus was analysed with a transitional[3] version of *Bohnet's parser* (Bohnet and Kuhn, 2012). It was one of the best parsers in the CoNLL Shared Task 2009 (Hajič et al., 2009).

[2] JAPE, a component of GATE, is a finite state transducer that operates over annotations based on regular expressions.

[3] Patent claim sentences can be very long which implies long-distance dependencies. Therefore, transition-based parsers, which typically have a linear or quadratic complexity (Nivre and Nilsson, 2004; Attardi, 2006), are better suited for parsing patent sentences than graph-based parsers, which usually have a cubic complexity.



[Toolholder,]

[comprising]

[a holder body with an insert site at its forward end comprising a bottom surface and at least one side wall where there projects a pin from said bottom surface upon which there is located an insert having a central bore, a clamping wedge for wedging engagement between a support surface of the holder and an adjacent edge surface of said insert and an actuating screw received in said wedge whilst threadably engaged in a bore of said holder, said support surface and said edge surface are at least partially converging downwards said wedge clamp having distantly provided protrusions for abutment against the top face and the edge surface of said insert, characterized in that the wedge consists of a pair of distantly provided first protrusions for abutment against a top face of the insert, and a pair of distantly provided second protrusions for abutment against an adjacent edge surface.]

Toolholder,

comprising

a holder body with
an insert site at its forward end
comprising a bottom surface and at least one side wall
where there projects a pin from said bottom surface
upon which there is located an insert
having a central bore ,
a clamping wedge
for wedging engagement between a support surface of the holder and an adjacent edge surface of said insert
and an actuating screw
received in said wedge
whilst threadably engaged in a bore of said holder ,
said support surface and said edge surface are at least partially converging downwards
said wedge clamp having distantly provided protrusions for abutment against the top face and the edge surface of said insert,
characterized in that the wedge consists of a pair of distantly provided first protrusions for abutment against a top face of the insert ,
and a pair of distantly provided second protrusions for abutment against an adjacent edge surface.

Figure 2: Example of the claim segmentation experiments

ANNI Tokenizer ⟶ RegEx Sentence Splitter ⟶ OpenNLP
POS Tagger ⟶ Noun Phrase Chunker ⟶ JAPE

Figure 3: GATE pipeline for Task 1

In order to characterise the clause boundaries, the following *features* were used for each token in the corpus:

- lemma of the current token,
- part-of-speech (POS) tag[4] of the current token as well as POS-tags of the two immediately preceding and two immediately subsequent words,
- syntactic head and dependent of the current token, and
- syntactic dependency relation between the current token and its head and dependent tokens.

Moreover, the fifteen most frequent lemmas and five most frequent POS-tags and punctuation marks were used as features we called segmentation keywords (Table 3).

Table 3: The most frequent lemmas and POS-tags in the beginning of a segment.

Rank	Lemmas	Abs. Freq.	Rel. Freq.
1	and	675	0.137
2	wherein	554	0.112
3	for	433	0.088
4	which	174	0.035
5	have	158	0.032
6	to	155	0.031
7	characterize	152	0.031
8	a	149	0.030
9	the	142	0.028
10	say	140	0.028
11	is	64	0.013
12	that	62	0.012
13	form	59	0.012
14	in	58	0.011
15	when	56	0.011
Rank	POS-tag	Abs. Freq.	Rel. Freq.
1	IN	739	0.150
2	CC	686	0.139
3	VBN	511	0.104
4	VBG	510	0.104
5	WRB	409	0.083

For classification we used the *CRF++ toolkit*, an open source implementation of conditional random fields (Lafferty et al., 2001). This framework for building probabilistic graphical models to segment and label sequence data has been successfully applied to solve chunking (Sha and Pereira, 2003), information extraction (Smith, 2006), and other sequential labelling problems. We compared the results obtained by CRF++ with the following baselines:

- *Baseline 1*: each punctuation mark is a segment boundary, and
- *Baseline 2*: each punctuation mark and keyword is a segment boundary.

[4] The POS-tag corresponds to the Peen Tree Bank tag set (Marcus et al., 1993) whereas IN = preposition or conjunction, subordinating; CC = Coordinating Conjunction; VBN = Verb, past participle; VBG = verb, gerund or present participle; WRB = Wh-adverb.

Table 4: Evaluation of claim components

		Correct	Incorrect
Preamble	Beginning	100%	0%
	End	97%	3%
Transition	Beginning	94%	6%
	End	100%	0%
Body text	Beginning	100%	0%
	End	100%	0%

Performance in Task 1 was assessed using the *accuracy*. Due to the lack of a corpus annotated with claims components, we selected twenty claims randomly and performed the annotation ourselves (i.e., one of the authors annotated the claims). The annotator was asked to assess whether the beginning and ending of a claim component was successfully identified.

Performance in Task 2 was evaluated using the *precision*, *recall*, and *F-score* on the test set. We considered that clause segmentation is a precision oriented task, meaning that we emphasised the demand for a high precision at the expense of a possibly more modest recall.

In order to better understand errors in clause segmentation, we analysed errors qualitatively using *content analysis* (Stemler, 2001). This method is commonly used in evaluation of language technologies. Fifty segmentation errors were randomly selected and manually analysed by one of the authors.

4 Results and discussion

4.1 Statistical performance evaluation in Tasks 1 and 2

We achieved a substantial accuracy in Task 1, claim component segmentation (Table 4). That is, the resulting segmentation was almost perfect. This was not surprising since we were processing simple and well defined types of segments. However, there was a small mismatch in the boundary identification for the preamble and the transition segments.

Our ML method clearly outperformed both its baselines in Task 2 (Table 5). It had the precision of 77 per cent and recall of 76 per cent in clause segmentation. The respective percentages were 41 and 29 for the baseline based on both punctuation and keywords. If punctuation was used alone, both the precision and recall were almost zero.

Table 5: Evaluation of claim clauses

	Precision	Recall	F-score
Baseline 1	0.2%	0.3%	2.6%
Baseline 2	41%	29%	34%
CRF++	77%	76%	76%

4.2 Qualitative analysis of errors in Task 2

The most common errors in clause segmentation were due to two reasons: first, ambiguity in coordinating conjunctions (e.g., commas as wll as *and*, *or*, and other particles) and second, consecutive segmentation keywords.

Segmentation errors caused by ambiguous coordinating conjunctions were due to the fact that not all of them were used as segment delimiters. Let us illustrate this with the following automatically segmented claim fragment with two coordinating conjunctions (a segment is a string between square brackets, the integer sub-script indicating the segment number, and the conjunctions in italics):

> ... [said blade advancing member comprises a worm rotatable by detachable handle]$_1$ [*or* key]$_2$ [*and* meshingeorm wheel secured to a shift]$_3$ ...

In this example, the two conjunctions were considered as segment delimiters which resulted in an incorrect segmentation. The correct analysis would have been to maintain the fragment as a single segment since simple noun phrases are not annotated as individual segments in our corpus.

Segmentation errors due to consecutive segmentation keywords resulted in undesirable segments only once in our set of fifty cases. This happened because the classifier segmented every encounter with a segmentation keyword, even when the keywords were consecutive. We illustrate this case with the following example, which contains two consecutive keywords, a verb in past participle (*selected*) and a subordinate conjunction (*for*). Example (a) shows a wrong segmentation, while example (b) shows its correct segmentation.

> ... (a) [said tool to be]$_1$ [selected]$_2$ [for the next working operation]$_3$ ...
>
> ... (b) [said tool to be selected]$_1$ [for working]$_2$ ...

In general, correcting both these error types should be relatively straightforward. First, to solve the problem of ambiguous commas, a possible solution could be to constrain their application as keywords, for example, by combining commas with other context features. Second, segmentation errors caused by consecutive segmentation keywords could be solved, for example, by applying a set of correction rules after the segmentation algorithm (Tjong and Sang, 2001).

5 Conclusion and future work

In this paper we have presented our on-going research on claim readability. We have proposed a method that focuses on presenting the claims in a clearer way rather than modifying their text content. This claim clarity is an important characteristic for inventors, researchers, and other laypeople. It may also be useful for patent experts, because clear clauses may help them to avoid future legal cost due to litigations. Moreover, better capabilities to understand patent documents contribute to democratisation of the invention and, therefore, to human knowledge.

For future work, we plan to conduct a user-centered evaluation study on claim readability. We wish to ask laypeople and patents experts to assess the usability and usefulness of our approach. Furthermore, we plan to consider text highlighting, terminology linking to definitions, and other content enrichment functionalities as ways of improving claim readability.

Acknowledgments

NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. We also express our gratitude to the TALN Research Group from Universitat Pompeu Fabra for their corpus development. Finally, we thank the anonymous reviewers of The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014), held in conjunction with the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014), for their comments and suggestions.

References

D. Alberts, C. Barcelon Yang, D. Fobare-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, and D. DeMarco. 2011. Introduction to patent searching. In M Lupu, J Tait, . Mayer, and A J Trippe, editors, *Current Challenges in Patent Information Retrieval*, pages 3–44, Toulouse, France. Springer.

G. Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 166–170, Stroudsburg, PA, USA. Association for Computational Linguistics.

B. Bohnet and J. Kuhn. 2012. The best of both worlds: a graph-based completion model for transition-based parsers. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 77–87, Stroudsburg, PA, USA. Association for Computational Linguistics.

N. Bouayad-Agha, G. Casamayor, G. Ferraro, S. Mille, V. Vidal, and Leo Wanner. 2009. Improving the comprehension of legal documentation: the case of patent claims. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 78–87, New York, NY, USA. Association for Computing Machinery.

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. 2011. *Text Processing with GATE (Version 6)*. Gateway Press CA, Shefield. UK.

G. Ferraro. 2012. *Towards Deep Content Extraction: The Case of Verbal Relations in Patent Claims. PhD Thesis*. Department of Information and Communication Technologies, Pompeu Fabra Univesity, Barcelona. Catalonia. Spain.

J. Hajič, M. Ciaramita, R. Johansson, D. Kawahara, M. A. Mart, L. Márquez, A. Meyers, J. Nivre, S. Pad, J. Stepanek, et al. 2009. The CoNLL-2009 shared task: syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, page 118, Stroudsburg, PA, USA. Association for Computational Linguistics.

K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. 2003. Text simplification for reading assistance: A project note. In *In Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*, IWP '03, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

D. Lafferty, A. McCallum, and F. C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

J. Nivre and J. Nilsson. 2004. Memory-based dependency parsing. In *Proceedings of the Eight Conference on Computational Natural Language Learning*, CoNLL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

D. Pressman. 2006. *Patent It Yourself*. Nolo, Berkeley, CA.

D. V. Radack. 1995. Reading and understanding patent claims. *JOM*, 47(11):69–69.

E. T. K. Sang and H. Déjean. 2001. Introduction to the CoNLL-2001 shared task: Clause identification. In W. Daelemans and R. Zajac, editors, *Proceedings of the Fith Conference on Computational Natural Language Learning*, volume 7 of *CoNLL '01*, pages 53–57, Toulouse, France.

F. Sha and F. Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1 of *NAACL '03*, pages 134–141, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. Sheremetyeva. 2003. Natural language analysis of patent claims. In *Proceedings of the ACL 2003 Workshop on Patent Processing*, ACL '03, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama. 2003. Patent claim processing for readability: structure analysis and term explanation. In *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*, volume 20 of *PATENT '03*, pages 56–65, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Shinmori, M. Okumura, and Marukawa. 2012. Aligning patent claims with the "detailed description" for readability. *Journal of Natural Language Processing*, 12(3):111–128.

A. Smith. 2006. Using Gazetteers in discriminative information extraction. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL '06, pages 10–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. Stemler. 2001. An overview of content analysis. *Practical Assessment, Research and Evaluation*, 7(17).

H. Suominen, S. Salantera, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. F. Jones, J. Leveling, L. Kelly, L. Goeuriot, Da Martinez, and Gu Zuccon. 2013. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In Pa Forner, H Müller, R Parades, P Rosso, and B Stein, editors, *Information Access Evaluation: Multilinguality, Multimodality, and Visualization. Proceedings of the 4th International Conference of the CLEF Initiative*, volume 8138 of *Lecture Notes in Computer Science*, pages 212–231, Heidelberg, Germany. Springer.

E. F. Tjong and Kim Sang. 2001. Memory-based clause identification. In *Proceedings of the 2001 workshop on Computational Natural Language Learning - Volume 7*, ConLL '01, Stroudsburg, PA, USA. Association for Computational Linguistics.