Dependency-based PropBanking of clinical Finnish

Katri Haverinen,[1,3] **Filip Ginter,**[1] **Timo Viljanen,**[1]
Veronika Laippala[2] **and Tapio Salakoski**[1,3]
[1]Department of Information Technology
[2]Department of French Studies
[3]Turku Centre for Computer Science, TUCS
20014 University of Turku, Finland
first.last@utu.fi

Abstract

In this paper, we present a PropBank of clinical Finnish, an annotated corpus of verbal propositions and arguments. The clinical PropBank is created on top of a previously existing dependency treebank annotated in the Stanford Dependency (SD) scheme and covers 90% of all verb occurrences in the treebank.

We establish that the PropBank scheme is applicable to clinical Finnish as well as compatible with the SD scheme, with an overwhelming proportion of arguments being governed by the verb. This allows argument candidates to be restricted to direct verb dependents, substantially simplifying the PropBank construction.

The clinical Finnish PropBank is freely available at the address http://bionlp.utu.fi.

1 Introduction

Natural language processing (NLP) in the clinical domain has received substantial interest, with applications in decision support, patient managing and profiling, mining trends, and others (see the extensive review by Friedman and Johnson (2006)). While some of these applications, such as document retrieval and trend mining, can rely solely on word-frequency-based methods, others, such as information extraction and summarization require a detailed linguistic analysis capturing some of the sentence semantics. Among the most important steps in this direction is an analysis of verbs and their argument structures.

In this work, we focus on the Finnish language in the clinical domain, analyzing its verbs and their argument structures using the PropBank scheme (Palmer et al., 2005). The choice of this particular scheme is motivated by its practical, application-oriented nature. We build the clinical Finnish PropBank on top of the existing dependency treebank of Haverinen et al. (2009).

The primary outcome of this study is the PropBank of clinical Finnish itself, consisting of the analyses for 157 verbs with 2,382 occurrences and 4,763 arguments, and covering 90% of all verb occurrences in the underlying treebank. This PropBank, together with the treebank, is an important resource for the further development of clinical NLP applications for the Finnish language.

We also establish the applicability of the PropBank scheme to the clinical sublanguage with its many atypical characteristics, and finally, we find that the PropBank scheme is compatible with the Stanford Dependency scheme of de Marneffe and Manning (2008a; 2008b) in which the underlying treebank is annotated.

2 The PropBank scheme

Our annotation work is based on the PropBank semantic annotation scheme of Palmer et al. (2005). For each verb, PropBank defines a number of *framesets*, each frameset corresponding to a coarse-grained sense. A frameset consists of a *roleset* which defines a set of *roles* (*arguments* numbered from Arg0 onwards) and their descriptions, and a set of syntactic *frames*. Any element that occurs together with a given verb sufficiently frequently is taken to be its argument. Arg0 is generally a *prototypical Agent* argument and Arg1 is a *prototypical Patient or Theme* argument. The remaining numbered arguments have no consistent overall meanings: they are defined on a verb-by-verb basis. An illustration of a verb with two framesets is given in Figure 1. In addition to the numbered arguments, a verb occurrence can have a number of modifiers, labeled ArgM, each modifier being categorized as one of 14 subtypes, such as *temporal*, *cause* and *location*.

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 137–141,
Uppsala, Sweden, 15-16 July 2010.

kestää.0: “tolerate”	kestää.1: “last”
Arg0: the one who tolerates	Arg1: the thing that lasts
Arg1: what is being tolerated	Arg2: how long it lasts

Figure 1: The PropBank framesets for *kestää* (translated to English from the original frames file) correspond to two different uses of the verb.

Pitkä yövuoro	Long nightshift
Jouduttu laittamaan	Had to put to
illala bipap:lle,	bipap in the evning,
nyt hapettuu hyvin.	now oxidizes well.
DIUREESI: riittävää	DIURESIS: sufficient
Tajunta: rauhallinen	Consciousness: calm
hrhoja ei enää ole	there are no more hllucinations

Figure 2: Example of clinical Finnish (left column) and its exact translation (right column), with typical features such as spelling errors preserved.

3 Clinical Finnish and the clinical Finnish treebank

This study is based on the clinical Finnish treebank of Haverinen et al. (2009), which consists of 2,081 sentences with 15,335 tokens and 13,457 dependencies. The text of the treebank comprises eight complete patient reports from an intensive care unit in a Finnish hospital. An intensive care patient report describes the condition of the patient and its development in time. The *clinical Finnish* in these reports has many characteristics typical of clinical languages, including frequent misspellings, abbreviations, domain terms, telegraphic style and non-standard syntactic structures (see Figure 2 for an illustration). For a detailed analysis, we refer the reader to the studies by Laippala et al. (2009) and Haverinen et al. (2009).

The treebank of Haverinen et al. is annotated in the Stanford Dependency (SD) scheme of de Marneffe and Manning (2008a; 2008b). This scheme is layered, and the annotation variant of the treebank of Haverinen et. al is the *basic* variant of the scheme, in which the analysis forms a tree.

The SD scheme also defines a *collapsed dependencies with propagation of conjunct dependencies* variant (referred to as the *extended* variant of the SD scheme throughout this paper). It adds on top of the *basic* variant a second layer of dependencies which are not part of the strict, syntactic tree. In particular, the *xsubj* dependency marks external subjects, and dependencies involving the heads of coordinations are explicitly duplicated also for the remaining coordinated elements where appropriate. The *extended* variant of the SD scheme is illustrated in Figure 3.

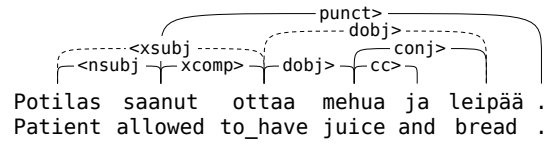

Figure 3: The *extended* SD scheme. The dashed dependencies denote the external subjects and propagated conjunct dependencies that are only part of the *extended* variant of the scheme. The example can be translated as *Patient [has been] allowed to have juice and bread.*

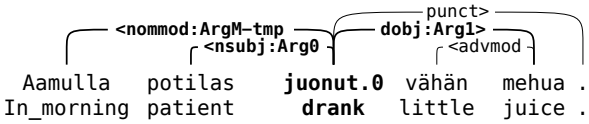

Figure 4: The PropBank annotation scheme on top of the treebank syntactic annotation. The verb *juonut (drank)* is marked with its frameset, in this case the frameset number 0. This frameset specifies that Arg0 marks the agent doing the drinking and Arg1 the liquid being consumed. The ArgM-tmp label specifies that *Aamulla* is a temporal modifier. The example can be translated as *In the morning patient drank a little juice*.

Due to the importance of the additional dependencies for PropBanking (see Section 5 for discussion), we augment the annotation of the underlying treebank to conform to the *extended* variant of the SD scheme by manual annotation, adding a total of 520 dependencies.

The PropBank was originally developed on top of the constituency scheme of the Penn Treebank and requires arguments to correspond to constituents. In a dependency scheme, where there is no explicit notion of constituents, we associate arguments of a verb with dependencies governed by it. The argument can then be understood as the entire subtree headed by the dependent. The annotation is illustrated in Figure 4.

4 PropBanking clinical Finnish

When annotating the clinical Finnish PropBank, we consider all verbs with at least three occurrences in the underlying treebank. In total, we analyze 157 verbs with 192 framesets. Since the treebank does not have gold-standard POS infor-

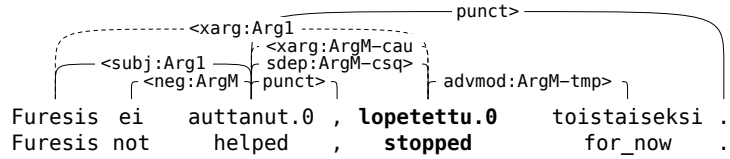

Figure 5: The simplified PropBank annotation strategy. The dashed dependencies labeled with the technical dependency type *xarg* signify arguments and modifiers not in a syntactic relationship to the verb. These arguments and modifiers, as well as those associated with a *conj* or *sdep* dependency (ArgM-csq in this Figure), are only marked in the 100 sentence sample for quantifying unannotated arguments and modifiers. The sentence can be translated as *Furesis did not help, stopped for now*.

mation, we identify all verbs and verbal participles using the FinCG[1] analyzer, which gives a verbal reading to 2,816 tokens. With POS tagging errors taken into account, we estimate the treebank to contain 2,655 occurrences of verbs and verb participles. Of these, 2,382 (90%) correspond to verbs with at least three occurrences and are thus annotated. In total, these verbs have 4,763 arguments and modifiers.

Due to the telegraphic nature of clinical Finnish, omissions of different sentence elements, even main verbs, are very frequent. In order to be able to analyze the syntax of sentences with a missing main verb, Haverinen et al. have added a so called *null verb* to these sentences in the treebank. For instance, the clinical Finnish sentence *Putkesta nestettä (Liquid from the drain)* lacks a main verb, and the insertion of one produces *Putkesta *null* nestettä*. In total, there are 428 null verb occurrences, making the null verb the most common verb in the treebank.

In the clinical PropBank annotation, we treat the null verb in principle as if it was a regular verb, and give it framesets accordingly. For each null verb occurrence, we have determined which regular verb frameset it stands for, and found that, somewhat surprisingly, there were only four common coarse senses of the null verb, roughly corresponding to four framesets of the verbs *olla (to be)*, *tulla (to come)*, *tehdä (to do)* and *laittaa (to put)*. The 26 (6%) null verb occurrences that did not correspond to any of these four framesets were assigned to a "leftover frameset", for which no arguments were marked.

[1] http://www.lingsoft.fi

5 Annotating the arguments on top of the SD scheme

In contrast to the original PropBank, where any syntactic constituent could be marked as an argument, we require arguments to be directly dependent on the verb in the SD scheme (for an illustration, see Figure 5). This restriction is to considerably simplify the annotation process — instead of all possible subtrees, the annotator only needs to look for direct dependents of the verb. In addition, this constraint should naturally also simplify possible automatic identification and classification of the arguments.

In addition to restricting arguments to direct dependents of the verb, coordination dependencies *conj* and *sdep* (implicit coordination of top level independent clauses, see Figure 5) are left outside the annotation scope. This is due to the nature of the clinical language, which places on these dependencies cause-consequence relationships that require strong inference. For instance, sentences such as *Patient restless, given tranquilizers* where there is clearly a causal relationship but no explicit marker such as *thus* or *because*, are common.

Naturally, it is necessary to estimate the effect of these restrictions, which can be justified only if the number of lost arguments is minimal. We have conducted a small-scale experiment on 100 randomly selected sentences with at least one verb that has a frameset assigned. We have provided this portion of the clinical PropBank with a full annotation, including the arguments not governed by the verb and those associated with *conj* and *sdep* dependencies. For an illustration, see Figure 5.

There are in total 326 arguments and modifiers (169 arguments and 157 modifiers) in the 100 sentence sample. Of these, 278 (85%) are governed by the verb in the *basic* SD scheme and are thus in a direct syntactic relationship with the verb. Fur-

ther 19 (6%) arguments and modifiers are governed by the verb in the *extended* SD scheme. Out of the remaining 29 (9%), 23 are in fact modifiers, leaving only 6 numbered arguments not accounted for in the *extended* SD scheme. Thus, 96% (163/169) of arguments and 85% (134/157) of modifiers are directly governed by the verb.

Of the 23 ungoverned modifiers, all are either cause (CAU) or consequence (CSQ)[2]. Of the *sdep* and *conj* dependencies only a small portion (9/68) were associated with an argument or a modifier, all of which were in fact CAU or CSQ modifiers. Both these and the CAU and CSQ modifiers not governed by the verb reflect strongly inferred relationships between clauses.

Based on these figures, we conclude that an overwhelming majority of arguments and modifiers is governed by the verb in the *extended* SD scheme and restricting the annotation to dependents of the verb as well as leaving *sdep* and *conj* outside the annotation scope seems justified. Additionally, we demonstrate the utility of the conjunct dependency propagation and external subject marking in the *extended* SD scheme.

6 Related work

Many efforts have been made to capture meanings and arguments of verbs. For instance, the VerbNet project (Kipper et al., 2000) strives to create a broad on-line verb lexicon, and FrameNet (Ruppenhofer et al., 2005) aims to document the range of valences of each verb in each of its senses. The PropBank project (Palmer et al., 2005) strives for a practical approach to semantic representation, adding a layer of semantic role labels to the Penn Treebank (Marcus et al., 1993).

In addition to the original PropBank by Palmer et al., numerous PropBanks have been developed for languages other than English (e.g. Chinese (Xue and Palmer, 2003) and Arabic (Diab et al., 2008)). Also applications attempting to automatically recover PropBank-style arguments have been proposed. For example, the CoNLL shared task has focused on semantic role labeling four times, twice as a separate task (Carreras and Màrquez, 2004; Carreras and Màrquez, 2005), and twice in conjunction with syntactic parsing (Surdeanu et al., 2008; Hajič et al., 2009).

In semantic analysis of clinical language, Paek et al. (2006) have experimented on PropBank-based machine learning on abstracts of Randomized Controlled Trials (RCTs), and Savova et al. (2009) have presented work on temporal relation discovery from clinical narratives.

7 Conclusion

In this paper, we have presented a PropBank of clinical Finnish, building a new layer of annotation on top of the existing clinical treebank of Haverinen et al. (2009). This PropBank covers all 157 verbs occurring at least three times in the treebank and accounts for 90% of all verb occurrences.

This work has also served as a test case for the PropBank annotation scheme in two senses. First, the scheme has been tested on a highly specialized language, clinical Finnish, and second, its compatibility with the SD syntactic scheme has been examined. On both accounts, we find the PropBank scheme a suitable choice.

In general, the specialized language did not seem to cause problems for the scheme. For instance, the frequent null verbs could be analyzed similarly to regular verbs, with full 94% belonging to one of only four framesets. This is likely due to the very restricted clinical domain of the corpus.

We also find a strong correspondence between the PropBank arguments and the verb dependents in the *extended* SD scheme, with 96% of arguments and 85% of modifiers being directly governed by the verb. The 15% ungoverned modifiers are cause-consequence relationships that require strong inference. This correspondence allowed us to simplify the annotation task by only considering direct verb dependents as argument candidates.

The new version of the treebank, manually anonymized, including the enhanced SD scheme annotation and the PropBank annotation, is freely available at `http://bionlp.utu.fi`.

Acknowledgments

We are grateful to Heljä Lundgren-Laine, Riitta Danielsson-Ojala and prof. Sanna Salanterä for their assistance in the anonymization of the corpus. We would also like to thank Lingsoft Ltd. for making FinTWOL and FinCG available to us. This work was supported by the Academy of Finland.

[2]CSQ is a new modifier subtype added by us, due to the restriction of only annotating direct syntactic dependents, which does not allow the annotation of all causal relationships with the type CAU.

References

Xavier Carreras and Lluís Màrquez. 2004. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 89–97, Boston, Massachusetts, USA, May 6 - May 7. Association for Computational Linguistics.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Marie-Catherine de Marneffe and Christopher Manning. 2008a. Stanford typed dependencies manual. Technical report, Stanford University, September.

Marie-Catherine de Marneffe and Christopher Manning. 2008b. Stanford typed dependencies representation. In *Proceedings of COLING'08, Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8.

Mona Diab, Mansouri Aous, Martha Palmer, Babko-Malaya Olga, Wadji Zaghouani, Ann Bies, and Mohammed Maamouri. 2008. A pilot Arabic PropBank. In *Proceedings of LREC'08*, pages 3467–3472. Association for Computational Linguistics.

Carol Friedman and Stephen Johnson. 2006. Natural language and text processing in biomedicine. In *Biomedical Informatics*, pages 312–343. Springer.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria A. Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Niawen Xue, and Yi Zhang. 2009. The CoNLL-2008 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of CoNLL'09: Shared Task*, pages 1–18. Association for Computational Linguistics.

Katri Haverinen, Filip Ginter, Veronika Laippala, and Tapio Salakoski. 2009. Parsing clinical Finnish: Experiments with rule-based and statistical dependency parsers. In *Proceedings of NODALIDA'09, Odense, Denmark*, pages 65–72.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 691–696. AAAI Press / The MIT Press.

Veronika Laippala, Filip Ginter, Sampo Pyysalo, and Tapio Salakoski. 2009. Towards automatic processing of clinical Finnish: A sublanguage analysis and a rule-based parser. *International Journal of Medical Informatics, Special Issue on Mining of Clinical and Biomedical Text and Data*, 78:7–12.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Hyung Paek, Yacov Kogan, Prem Thomas, Seymor Codish, and Michael Krauthammer. 2006. Shallow semantic parsing of randomized controlled trial reports. In *Proceedings of AMIA'06*, pages 604–608.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1).

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2005. FrameNet II: Extended theory and practice. Technical report, ICSI.

Guergana Savova, Steven Bethard, Will Styler, James Martin, Martha Palmer, James Masanz, and Wayne Ward. 2009. Towards temporal relation discovety from the clinical narrative. In *Proceedings of AMIA'09*, pages 568–572.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing on syntactic and semantic dependencies. In *Proceedings of CoNLL'08*, pages 159–177. Association for Computational Linguistics.

Nianwen Xue and Martha Palmer. 2003. Annotating the propositions in the Penn Chinese Treebank. In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, pages 47–54, Sapporo, Japan. Association for Computational Linguistics.