Sparv: Språkbanken's corpus annotation pipeline infrastructure

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, Anne Schumacher

Språkbanken, Department of Swedish
University of Gothenburg
lars.borin@gu.se, markus.forsberg@gu.se, martin.hammarstedt@gu.se,
dan.rosen@gu.se, roland.schaefer@fu-berlin.de, anne.schumacher@gu.se

Abstract

Sparv is Språkbanken's corpus annotation pipeline infrastructure. The easiest way to use the pipeline is from its web interface with a plain text document. The pipeline uses in-house and external tools on the text to segment it into sentences and paragraphs, tokenise, tag parts-of-speech, look up in dictionaries and analyse compounds. The pipeline can also be run using a web API with XML results, and it is run locally at Språkbanken to prepare the documents in Korp, our corpus search tool. While the most sophisticated support is for modern Swedish, the pipeline supports 15 languages.

1. Introduction

Development of the pipeline Sparv for *corpus import* started at Språkbanken (https://spraakbanken.gu.se) for our corpus search tool Korp (Borin et al., 2012) to enable search queries over attributes such as parts-of-speech, word compounds or lemmas (base forms). The analyses to obtain this rich information are performed by internal and external tools. The pipeline is also accessible independent of Korp as a web service, either through a graphical web interface or the API. This offers an easy way to quickly annotate shorter texts using the same tools that are used to annotate the corpora in Korp.

The main contributions of Sparv and this article are:

- Connects high class annotation tools such as the HunPos tagger and the MaltParser (Section 2.)
- Components can be built on top of it by using the web API (Section 3.2)
- The web interface (Section 3.1) can be used for evaluating the Korp annotations, as well as for educational purposes
- Partial support for historical Swedish (Section 4.)
- Support for multiple languages (Section 5.)
- Free software distributed under the MIT license[1].

2. Analyses and tools

The different types of analyses can be divided into two subcategories: (1) analyses on the word or token level and (2) analyses on the document level. Details on the analyses of type (1) for modern Swedish are described in this section. Document analyses such as text classification and document clustering are not supported by Sparv yet but is an important strand of future work (see Section 6.). It is possible, however, to run partially marked-up material through the Sparv pipeline. Metadata such as information about when and where a text was written or who the author was will thus be preserved and shown in the result.

[1] https://spraakbanken.gu.se/eng/research/infrastructure/sparv/distribution

2.1 Lexical analysis

The lexical analysis in Sparv consists of several steps: tokenisation, lemmatisation, identification of *lemgrams* and word senses and compound analysis (see Section 2.2). A lemgram is a lexical identifier which refers to an inflection table in the SALDO lexicon (Borin et al., 2013a), which provides linkages between lemgrams and word sense identifiers, although the relation is many-to-many. Tokenisation is done using a regular expression tokeniser based on PunktWordTokenizer from NLTK (Loper and Bird, 2002). For lemmatisation, identification of lemgrams and senses, the SALDO morphological lexicon is used. The analysis returns lemgrams and senses as SALDO identifiers which can be looked up in Språkbanken's lexical tool *Karp* (Borin et al., 2013b). SALDO is Språkbanken's lexical pivot resource linked to more than 25 other lexical resources. Thus, the SALDO identifiers link to additional lexical information.

2.2 Compound analysis

The current compound analysis tries to split compounds into two parts which are looked up in the SALDO lexicon.

The development version has a more advanced compound analysis which can split words into any number of components. In this new analysis it is clearly shown which word components belong to the same compound while the previous compound analysis only yielded two sets with unconnected word elements. In the development version a word split into n components $w_1, w_2, \ldots, w_n$ analysed with part-of-speech tags $t_1, t_2, \ldots, t_n$ is ranked using this scoring formula:

$$P(w_1, t_1) \cdot P(w_2, t_2) \cdot \ldots \cdot P(w_n, t_n) \cdot P(t_1, t_2, \ldots, t_n).$$

Here $P(w, t)$ is the relative frequency for a word component occurring with a certain part-of-speech. These values are obtained by looking at a large collection of corpora from Korp. The last factor $P(t_1, t_2, \ldots, t_n)$ is the relative frequency for the part-of-speech tags $t_1, t_2, \ldots, t_n$ occurring in this order in a compound. These values are extracted from the *Leksikalsk database for svensk* resource (Nasjonalbiblioteket, 2011).

2.3 Part-of-speech tagging and syntactic analysis

For part-of-speech tagging, Sparv uses HunPos (Halácsy et al., 2007), a trigram tagger, with a model trained on the SUC 3.0 corpus.

The syntactic analysis of Swedish in Sparv is performed using MaltParser (Nivre et al., 2007), a statistical dependency parser, with a model trained on the Swedish treebank Talbanken (Nivre et al., 2005).

2.4 Named entity recognition

Sparv's development version supports (a yet unevaluated) named entity recognition as an experimental feature. Currently Sparv is using the open source and partially in-house developed HFST-SweNER (Kokkinakis et al., 2014) for the extraction of named entities. Its implementation is based on the Helsinki Finite-State Transducer Technology platform and supports a variety of fine-grained named entity types as well as date/time expressions and numerical expressions.

3. Sparv online

Users can access Sparv online either from the web page interface or the web API. This allows using the pipeline without any installation.

3.1 Web interface

The web page address is https://spraakbanken.gu.se/sparv. Figure 1 shows a screenshot of the test sentence *Katten Gösta slösurfar på jobbet* 'Gösta the cat aimlessly surfs the web at work' typed in by the user which has been processed by the pipeline. The user can view the resulting analysis in the table below, where each word has been annotated with part-of-speech, the lemma, the dictionary lookups to Karp and SALDO, as well as compound analysis. For instance, *slösurfar* 'aimlessly surfs the web' has been given the prefix *slö* 'lazy' and suffix *surfar* 'surfs'. The dependency analysis is presented with a tree above the sentence, drawn using the brat library (Stenetorp et al., 2012). Abbreviations used for parts-of-speech and dependency labels can be hovered to show an expanded description.

3.2 Web API

The pipeline is accessible as web service through our web API. For example, the raw results from the pipeline on previous section's example sentence are obtained by requesting https://ws.spraakbanken.gu.se/ws/sparv/v1/?text=Katten+Gösta+slösurfar+på+jobbet.. The result is an XML document. Here is an illustration of a part of the document, showing the sentence's third word *slösurfar*:

```
<w pos="VB" msd="VB.PRS.AKT"
   lemma="|slösurfa|"
   lex="|slösurfa..vb.1|"
   saldo="|slösurfa..1|"
   prefix="|slö..av.1|"
   suffix="|surfa..vb.1|"
   ref="3" dephead="" deprel="ROOT"
   >slösurfar</w>
```

The same information as in the screenshot in Figure 1 is present: the word has the root dependency relation and no dependency head, the part-of-speech is `VB.PRS.AKT` (verb present active), the prefix *slö* and the suffix *surfa*.

The full documentation of the API is available at https://spraakbanken.gu.se/eng/research/infrastructure/sparv/webservice.

4. Historical Swedish

One of our goals with Sparv is to be able to analyse and annotate any Swedish text independent of when it was written. Older texts, however, pose significant problems to that aim because of different spelling conventions and errors introduced during the digitalisation process (Adesam et al., 2016). Furthermore, most texts that are digitally available today are written in modern Swedish which makes it much easier to train models on this type of language.

At Språkbanken we host a fairly large amount of documents from the 1800s and have therefore made an attempt to improve Sparv's annotations of texts written in that time period. As for modern Swedish, the HunPos-tagger is used for part-of-speech tagging, but we supply an additional morphology table to increase the chance of identifying words not following modern spelling conventions. The table is extracted from the two older Swedish lexicons *Swedberg* and *Dalin*. This extra lexical information improves lemmatisation, sense identification and part-of-speech tagging.

5. Multilinguality

For most resources developed at Språkbanken the focus lies on the analysis of the Swedish language. Nonetheless, we would like our tools to be as flexible as possible and therefore we have incorporated two off-the-shelf multilingual annotation tool kits in the Sparv pipeline: Freeling (Padró and Stanilovsky, 2012) and TreeTagger (Schmid, 1994). Both tools offer tokenisation, lemmatisation and part-of-speech tagging for a large range of languages. Using Freeling and TreeTagger, Sparv can currently provide lexical analysis for 14 different languages (Bulgarian, Dutch, English, Estonian, Finnish, French, German, Italian, Latin, Polish, Portuguese, Russian, Slovak and Spanish) and five more (Catalan, Galician, Norwegian, Romanian and Slovene) are on their way into the pipeline.

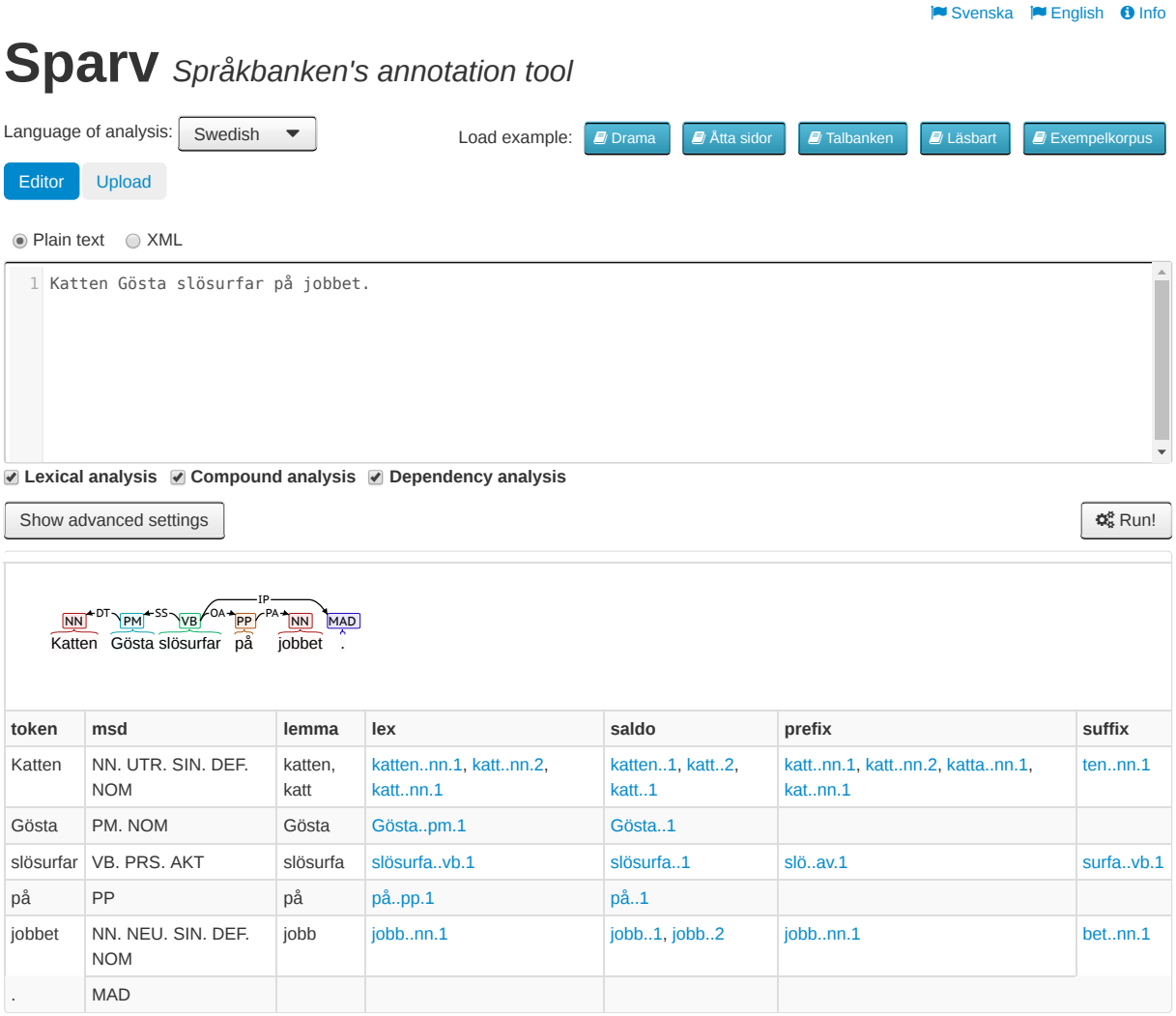
For parallel corpora Sparv also supports sentence linking based on the Gale-Church alignment algorithm (Gale and Church, 1993) and word linking using fast_align (Dyer et al., 2013).

6. Future work

We are planning to extend Sparv with document-oriented analyses, e.g., text classification and document clustering. The resulting annotations will be used in the document exploration tool *Strix* that is currently in development at Språkbanken. We also want to incorporate alternative tag sets such as Universal Dependencies (UD) (Nivre et al., 2016).

Another aim for Sparv is to improve the word linking for parallel corpora e.g. by training models for specific language pairs.

Moreover, we would like to work with ranking ambiguous annotations as for example the compound analysis and



token	msd	lemma	lex	saldo	prefix	suffix
Katten	NN. UTR. SIN. DEF. NOM	katten, katt	katten..nn.1, katt..nn.2, katt..nn.1	katten..1, katt..2, katt..1	katt..nn.1, katt..nn.2, katta..nn.1, kat..nn.1	ten..nn.1
Gösta	PM. NOM	Gösta	Gösta..pm.1	Gösta..1		
slösurfar	VB. PRS. AKT	slösurfa	slösurfa..vb.1	slösurfa..1	slö..av.1	surfa..vb.1
på	PP	på	på..pp.1	på..1		
jobbet	NN. NEU. SIN. DEF. NOM	jobb	jobb..nn.1	jobb..1, jobb..2	jobb..nn.1	bet..nn.1
.	MAD					

Figure 1: Testing the sentence *Katten Gösta slösurfar på jobbet.*, entered into the text area in the upper part of the image, in the Sparv web interface. After pressing the *Run!* button the pipeline is run on our servers and after a few seconds it responds with the analysis in the results table and the dependency tree.

the identification of word senses. A first attempt in performing word sense disambiguation will be integrated in Sparv soon.

7. Conclusion

We presented Sparv, our pipeline for making high quality text and corpora analyses. The analyses are used for our corpus search tool Korp, and are available for users to use on their texts as an online web service.

References

Yvonne Adesam, Malin Ahlberg, Peter Andersson, Lars Borin, Gerlof Bouma, and Markus Forsberg. 2016. Språkteknologi för svenska språket genom tiderna. *Kungliga Skytteanska Samfundets Handlingar*, 76(Studier i svensk språkhistoria 13):65–87.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp – the corpus infrastructure of språkbanken. In *Proceedings of LREC 2012. Istanbul: ELRA*, page 474–478.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013a. Saldo: a touch of yin to wordnet's yang. *Language resources and evaluation*, 47(4).

Lars Borin, Markus Forsberg, Leif-Jöran Olsson, Olof Olsson, and Jonatan Uppström. 2013b. The lexical editing system of karp. In *Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) 2013. Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.*, volume 2013, pages 503–516.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *In Proc. NAACL*.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19(1):75–102, March.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. HunPos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 209–212, Stroudsburg. Association for Computational Linguistics.

Dimitrios Kokkinakis, Jyrki Niemi, sam hardwick, Krister Lindén, and Lars Borin. 2014. Hfst-swener . a new ner resource for swedish. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC), Reykjavik 26 - 31 May 2014.*, pages 2537–2543.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nasjonalbiblioteket. 2011. NSTs leksikalsk database for svensk. https://www.nb.no/sprakbanken/show?serial=sbr-22&lang=nn.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2005. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, pages 24–26.

Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering*, 13(2):95–135.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK. ELRA.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: A web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.