SwissAdmin:
A multilingual tagged parallel corpus of press releases

Yves Scherrer, Luka Nerima, Lorenza Russo, Maria Ivanova, Eric Wehrli
LATL-CUI, University of Geneva, 7 route de Drize, Carouge, Switzerland
{yves.scherrer, luka.nerima, lorenza.russo, maria.ivanova, eric.wehrli}@unige.ch

Abstract
SwissAdmin is a new multilingual corpus of press releases from the Swiss Federal Administration, available in German, French, Italian and English. We provide SwissAdmin in three versions: (i) plain texts of approximately 6 to 8 million words per language; (ii) sentence-aligned bilingual texts for each language pair; (iii) a part-of-speech-tagged version consisting of annotations in both the Universal tagset and the richer Fips tagset, along with grammatical functions, verb valencies and collocations.

The SwissAdmin corpus is freely available at www.latl.unige.ch/swissadmin.

Keywords: multilingual corpus, sentence alignment, POS-tagging, collocations

1. Introduction
In this paper, we present a part-of-speech tagged parallel corpus of press releases from the Swiss Federal Administration. The press releases are available in the three official main languages of Switzerland (German, French and Italian) and partially in English.

After presenting the data source (Section 2.), we describe the preprocessing steps applied to the raw data as well as the sentence alignment process (Section 3.). In Section 4., we present the annotated version of the SwissAdmin corpus containing part-of-speech tags, lemmas, grammatical functions, verb valencies and collocations. We conclude by comparing our corpus to other similar resources.

2. The SwissAdmin corpus
The SwissAdmin corpus is a new language resource for the three official languages of Switzerland: German, French and Italian. About 20% of the texts are also available in English. It takes the form of a quadrilingual corpus consisting of press releases from the Swiss Federal Administration. Its web site http://www.news.admin.ch provides archives of the press releases since 1998.1

The web site aggregates press releases from the Federal Chancellery and various federal departments and offices. The documents contain news items concerning political matters. They are intended for a large audience and do not contain large amounts of specialized language. In practice, documents are written in one language and then translated to the other languages by the federal translation service. This ensures a high quality of the translated texts, but unfortunately the original language cannot be recovered from the publicly available data. We assume however that the original language of the majority of texts is German, and that none of the texts has English as original language.

We provide SwissAdmin in three versions: (i) plain text files that have been preprocessed and cleaned, but not annotated; (ii) sentence-alignment files for each language pair; (iii) text files annotated with POS-tags, using both the Universal tagset and the richer Fips tagset. The annotations also contain grammatical functions, valency information for verbs, as well as collocations (cf. Figure 1).

3. Preprocessing and sentence alignment
The corpus is constituted as follows:

- All documents are downloaded, and plain text is extracted from the HTML files using the BeautifulSoup library for Python.

Table 1: Number of press releases (documents) per language, after validation by the language identification tool. The last row shows the total number of words per language.

| Year | DE  | FR  | IT  | EN  |
|------|-----|-----|-----|-----|
| 2013 | 2082| 2073| 1932| 615 |
| 2012 | 2163| 2140| 1981| 563 |
| 2011 | 2100| 2077| 1887| 538 |
| 2010 | 2178| 2128| 1932| 525 |
| 2009 | 2286| 2236| 1964| 491 |
| 2008 | 2204| 2178| 1926| 409 |
| 2007 | 2133| 2064| 1794| 289 |
| 2006 | 1968| 1937| 1735| 259 |
| 2005 | 1085| 1060| 920 | 82  |
| 2004 | 1072| 1052| 866 | 75  |
| 2003 | 1082| 1049| 822 | 101 |
| 2002 | 761 | 724 | 530 | 65  |
| 2001 | 538 | 439 | 303 | 50  |
| 2000 | 570 | 550 | 340 | 43  |
| 1999 | 386 | 372 | 228 | 19  |
| 1998 | 136 | 134 | 46  | 1   |
| 1997 | 47  | 42  | 22  | 1   |
| Total| 22791| 22255| 19228| 4126|
| Words| 6.6M| 8.2M| 6.6M| 1.3M|

1The archive of older press releases may be found at http://www.admin.ch/cp, but the different language versions are not linked, which makes the alignment process difficult. For the moment, we do not take into account this additional resource.
| Year | DE-FR | DE-IT | FR-IT | DE-EN | FR-EN | IT-EN |
|------|-------|-------|-------|-------|-------|-------|
| 2013 | 27 504| 24 695| 24 262| 7 524 | 7 320 | 6 397 |
| 2012 | 27 021| 23 966| 23 682| 6 261 | 6 079 | 5 500 |
| 2011 | 27 612| 24 370| 24 265| 6 136 | 6 084 | 5 411 |
| 2010 | 29 385| 25 938| 25 589| 6 246 | 6 131 | 5 734 |
| 2009 | 29 826| 25 224| 25 034| 5 925 | 5 712 | 5 426 |
| 2008 | 29 453| 25 271| 25 028| 5 203 | 5 083 | 4 603 |
| 2007 | 26 487| 22 636| 22 555| 4 014 | 3 798 | 3 306 |
| 2006 | 24 220| 21 604| 21 512| 3 263 | 3 218 | 2 935 |
| 2005 | 14 054| 12 414| 12 181| 1 156 | 1 131 | 1 068 |
| 2004 | 14 132| 11 663| 11 511| 945   | 951   | 887   |
| 2003 | 13 769| 10 929| 10 741| 1 263 | 1 258 | 974   |
| 2002 | 10 044| 7 454 | 7 300 | 784   | 752   | 675   |
| 2001 |  6 574|  4 653|  4 548|  415  |  380  |  372  |
| 2000 |  8 452|  5 232|  5 137|  446  |  432  |  386  |
| 1999 |  5 566|  3 617|  3 544|  203  |  187  |  140  |
| 1998 |  2 185|  876 |  892 |  26   |  25   |   25  |
| 1997 |   755 |  399 |  401 |   31  |    0  |    0  |
| Total | 297 039| 250 941| 248 182| 49 841| 48 541| 43 839|

Table 2: Numbers of aligned sentences per language pair, after all cleaning and preprocessing steps.

- The different language versions of the documents are aligned; this is done by looking at the ID number that is shared by the different language versions of a press release.
- The text files are then cleaned up: empty lines are deleted, as well as paragraphs that contain less than 20 words. Such short paragraphs are mainly list items, addresses and disclaimers that we prefer to exclude from the final corpus.
- Moreover, we use a language identifier to guess the language of the file. If the guess does not match with the language indicated in the file name, the file is skipped.

The raw version of SwissAdmin corresponds to the result of these processing steps. Table 1 shows the number of documents per year and language.

The second version of SwissAdmin contains sentence-aligned data for each language pair. Sentence alignment was performed using Hunalign (Varga et al., 2005). As a preliminary step for sentence alignment, the texts had to be split in sentences. This was done with a specific tool provided with the Moses toolkit (Koehn et al., 2007). While this tool already contains lists of non-sentence-breaking prefixes (like "Mr." or "i.e.") for German and English, we created similar lists for French and Italian on the basis of abbreviations included in the Fips lexicon. The statistics of the sentence-aligned version are given in Table 2.

### 4. Annotation

In order to provide a suitable corpus for cross-linguistic studies, for development of NLP tools and in particular for the training of statistical systems, we also offer an annotated version of the corpus. While there are already many parallel corpora, only few of them are available for language pairs such as German–French or German–Italian, as claimed by Göhring and Volk (2011).

The annotation was performed automatically by the Fips parser (Wehrli, 2007; Wehrli and Nerima, 2014) used here as a POS tagger (henceforth referred to as the Fips tagger).\(^2\) The annotation includes lexical and morpho-syntactic information, as well as collocations. Two examples of annotated English sentences are given in Figure 1 below. For each token, the following information is displayed, spread over seven columns:

1. The **orthographical form** (token). Notice that Fips may group together words that form complex lexical units, for instance French compound nouns such as *Conseil fédéral* ("Federal Council") or *pomme de terre* ("potato"), complex conjunctions such as *as soon as*, fixed adverbial phrases such as *by and large*, or the German preposition *je nach* ("according to"). On the other hand, Fips may treat single words as multiple tokens. For instance, German compounds are decomposed, so that *Medaillengewinner* ("medal winner") will be presented as two tokens (*Medaillen* and *Gewinner*, similarly *Gebärdensprache* ("sign language") is represented as *Gebärden* ("hand sign") and *Sprache* ("language").\(^3\)

2. The **POS tag** in Universal Tagset format (Petrov et al., 2012), i.e., one of the following twelve POS tags:

\(^2\)The Fips tagger performs a complete syntactic analysis of the input document, using the whole grammar of the Fips system, but outputs results in a word-by-word manner without constituent information.

\(^3\)In other words, the tokenisation process adopted here differs in granularity from the one adopted by standard POS taggers. It considers a token to be a linguistically significant lexical unit rather than a sequence of characters between two separators.
The POS tag in the richer Fips format, which includes morphological information such as tense and mode, agreement features such as gender, number, person and case, as well as language-specific tags such as the infinitival marker to, or the possessive marker ’s in English, or clitic pronouns in Italian and French. Each tag consists of a category, optionally followed by a type and agreement features. For instance, the tag NOUN-COM-PLU designates a plural common noun, while VERB-AUX-INDE-IND-PRE-3-PLU indicates an auxiliary verb in third person plural indicative present. A complete list of the tags is available on the corpus website.

4. The lemma, which is the citation form associated with the token. Notice that the lemma includes the particle in the case of phrasal verbs in German and English.

5. The main grammatical functions are given, associated with the highest node of a constituent (for instance, for a noun phrase, the grammatical function will appear with the determiner, if there is one, otherwise with the noun). The grammatical function labels are SUBJ for subject, DO for direct object, IO for indirect object, PrepO for prepositional object, and ADJUNCT.

6. Each verb is annotated with valency information, which takes the form of a list of arguments, specifying the grammatical function (using the same labels as above) and the semantic head of the constituent. For instance, in the second example of Figure 1, the verb draw has a valency table with two arguments (DO:conclusions and PrepO:disasters). This means that the constituent critical conclusions, which is the grammatical subject of the sentence, is analyzed as the (deep) direct object of the verb, due to the passive construction. The second argument of the verb is the prepositional object natural disasters, with disasters as (semantic) head.

7. The collocation, if the word belongs to a collocation detected by the tagger (see below).

4.1. Collocations

Collocational knowledge is widely recognized as a useful information for a variety of NLP applications. This is why we decided to add that knowledge to the SwissAdmin corpus. The collocation detection process developed for Fips has been described in several publications (Seretan, 2011; Wehrli and Nerima, 2013). Suffice it to say here that this procedure can identify “known” collocations, i.e. collocations that have been lexicalized, even when their constituents are far apart or in non-canonical order, due to grammatical processes such as passivization, relativization, fronting, etc. The collocation detection procedure can also recognize a collocation, say of the verb-object type, when the object has been pronominalized. For instance, take it will count as an occurrence of the collocation to take a break when it refers to break, as discussed in detail in Wehrli and Nerima (2013).

Given the fact that only lexicalized collocations can be identified, the number of detected collocations crucially depends on the number of collocations in our database for a given language. Table 3 shows the number of collocations detected in a relatively large fragment of the corpus. The figures clearly confirm the importance of collocations, with nearly three collocations in every hundred words in French, for instance. Finally, it should be noticed that contrary to current practice, phrasal verbs in English and German were not counted as collocations; rather, they are treated as specialized lexeme forms in our database.

Table 3: Coverage of collocations. The table shows the absolute number of identified collocations, the number of tokens of the corpus fragment used to identify them, as well as the number of collocations per 100 tokens.

|                | DE  | FR  | IT  | EN  |
|----------------|-----|-----|-----|-----|
| Collocations   | 8594| 44673| 26578| 19162|
| Analyzed tokens| 1.1M| 1.5M| 2.9M| 1.3M|
| Coll./100 tokens| 0.76| 2.89| 0.91| 1.36|

Table 4: Manual evaluation of Fips tagging accuracy, measured on the universal tagset. The first two rows give details about the corpus fragment used for the evaluation.

|                | DE  | FR  | IT  | EN  |
|----------------|-----|-----|-----|-----|
| Sentences      | 164 | 2343| 434 | 123 |
| Tokens         | 3534| 60918| 11254| 3046|
| Accuracy       | 96.1%| 98.4%| 97.4%| 97.2%|

4.2. Evaluation

In order to validate the performance of the Fips tagger, we manually evaluated the POS tags on small excerpts of the SwissAdmin corpus. The accuracy was measured on all tokens, including punctuation symbols and unknown words, on the basis of the Universal Tagset. The evaluation was done by a native speaker of each language. The tagging accuracy for each language is shown in Table 4. The differences between languages reflect the current state of the respective grammars. Even if it is difficult to compare the results due to different tagging architectures and corpora used, the scores show state-of-the-art performance with regard to collocations of more than two units, such as weapons of mass destruction.

NOUN, VERB, ADJ, ADV, PRON, DET, ADP (adpositions, i.e., prepositions and postpositions), NUM (numerals), CONJ, PRT (particles), ‘’ (punctuation) and X (other).

---

4 We assume here a fairly broad definition of collocation, as an arbitrary and conventional combination of two syntactically related lexical units (not counting function words), such as adjective-noun (heavy smoker), verb-direct object (to take a break), noun-preposition-noun (flag of convenience), etc. Notice that in our definition, a lexical unit can be a lexeme or a collocation.
5. Related resources and availability

While the SwissAdmin corpus is not the only one available for the given languages, it differs in various ways from several related resources:

- **Europarl** (Koehn, 2005) is a very large parallel corpus that is also available in the four languages covered by SwissAdmin. However, it is of a slightly different genre (Parliament proceedings). While several subsets of it have been annotated for various purposes, there is no canonical annotation available for all languages covered by Europarl.

- The **WaCky** collection (Baroni et al., 2009) contains large amounts of POS-tagged text from the Web in the four SwissAdmin languages. However, its data sources are much more diverse, and the resulting text is noisier.

- **Text+Berg** (Göhring and Volk, 2011) is a German–French parallel corpus of mountaineering reports, part of which has been annotated as a parallel treebank. It also contains a small amount of Italian data.

- The **Allegra** corpus (Scherrer and Cartoni, 2012) has been extracted from a similar data source of press releases, but covers the Swiss minority language Romansh in addition to German and Italian. It has not been annotated.

Some of its characteristics make the SwissAdmin corpus particularly appealing for NLP research. For instance, it is one of the rare parallel texts of the news genre, which happens to be the genre mostly used in treebanks, and hence for training parsers. Also, to the best of our knowledge, this is the first multilingual corpus containing collocation annotations.

To conclude, SwissAdmin is a new multilingual corpus, freely available in three versions: a cleaned unannotated version, a sentence-aligned version and a POS-tagged version.⁵

The SwissAdmin corpus can be freely downloaded from our website [www.latl.unige.ch/swissadmin](http://www.latl.unige.ch/swissadmin).

---

⁵The copyright of the source texts remains the property of the Swiss Confederation, as stated on [http://www.disclaimer.admin.ch/terms_and_conditions.html](http://www.disclaimer.admin.ch/terms_and_conditions.html).
6. References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation, 43(3):209–226.

Göhring, A. and Volk, M. (2011). The Text+Berg corpus: An alpine French-German parallel resource. In Proceedings of Traitement Automatique des Langues Naturelles (TALN) 2011, Montpellier, France.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In Proceedings of the ACL 2007 demonstration session.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In Proceedings of MT Summit 2005.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Proceedings of LREC 2012, Istanbul, Turkey.

Scherrer, Y. and Cartoni, B. (2012). The trilingual AL-LEGRA corpus: Presentation and possible use for lexicon induction. In Proceedings of LREC 2012, Istanbul, Turkey.

Seretan, V. (2011). Syntax-Based Collocation Extraction. Springer.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In Proceedings of RANLP 2005, pages 590–596.

Wehrli, E. and Nerima, L. (2013). Sentence analysis and collocation identification. In Proceedings of the Workshop on Multiword Units in Machine Translation and Translation Technology, MT Summit XIV, pages 12–17, Nice, France.

Wehrli, E. and Nerima, L. (2014). The Fips multilingual parser. In Gala, N., Rapp, R., and Bel, G., editors, Festschrift in honour of Michael Zock. Springer.

Wehrli, E., Seretan, V., and Nerima, L. (2010). Anaphora resolution, collocations and translation. In Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), pages 27–35, Beijing, China.

Wehrli, E. (2007). Fips, a deep linguistic multilingual parser. In Proceedings of the ACL 2007 Workshop on Deep Linguistic Parsing, pages 120–127, Prague, Czech Republic.