Corpus-Based Acquisition of Support Verb Constructions for Portuguese

Britta D. Zeller   Sebastian Padó

Department of Computational Linguistics,
Heidelberg University, Germany

April 20, 2012

What are Support Verb Constructions (SVCs)?

- Subgroup of multiword expressions
  ⇒ SVCs form a syntactic and semantic unit
- Complex predicates (CPs) consisting of verb and noun; prepositional and non-prepositional
  Estou na dúvida. – I am in doubt.
  Vamos dar um passeio! – Let‘s take a walk!
  ⇒ Verbs in SVCs are often semantically impoverished (light verbs)
  ⇒ Hard to distinguish from other CPs and arbitrary constructions
  ⇒ Our focus: non-prepositional SVCs
- Often replaceable by an individual full verb
  Maria deu a resposta correcta. – Maria respondeu correctamente.
  Maria gave the correct answer. – Maria answered correctly.
  ⇒ Syntactic modifications

Motivation

SVCs have effect on the performance of many NLP tasks, e.g. for:

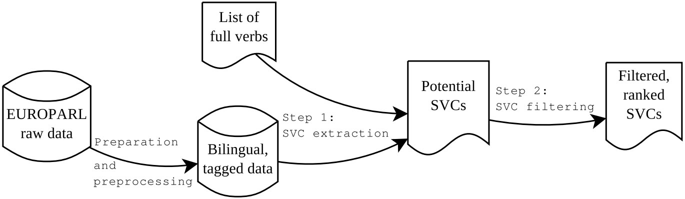
- Language generation (syntax):
  ∗ *Ela levou o amigo a casa e ao desespero.* [Athayde, 2001]
  ∗ *She drove her friend home and to despair.*
- Recognition of selectional preferences (semantics):
  $X_{\theta AGT}$ *decidiu* $Y_{\theta THM} \leftrightarrow X_{\theta AGT}$ *tomou a decisão de* $X_{\theta THM}$
  $X_{\theta AGT}$ *decided* $Y_{\theta THM} \leftrightarrow X_{\theta AGT}$ *made the decision to* $Y_{\theta THM}$

  $X_{\theta AGT}$ *decidiu* $Y_{\theta THM} \not\leftrightarrow X_{\theta AGT/CAUSE}$ *atrasou a decisão de* $Y_{\theta THM}$
  $X_{\theta AGT}$ *decided* $Y_{\theta THM} \not\leftrightarrow X_{\theta AGT/CAUSE}$ *delays the decision to* $Y_{\theta THM}$

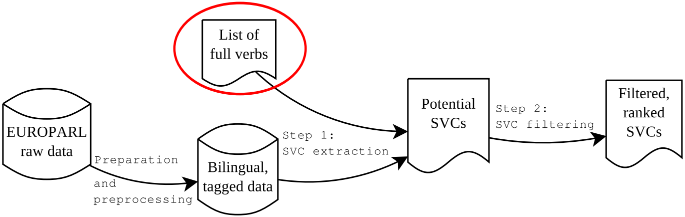
⇒ Recognition of SVCs is desirable

Our idea: Acquisition of Portuguese SVCs by combining cross- and monolingual methods with shallow preprocessing

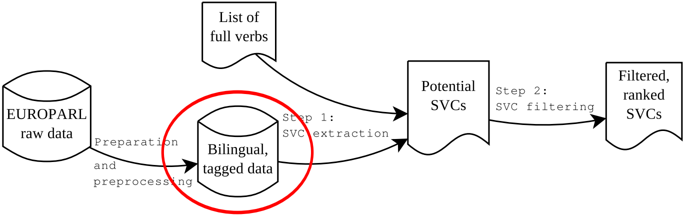
Overall structure of the SVC acquisition procedure



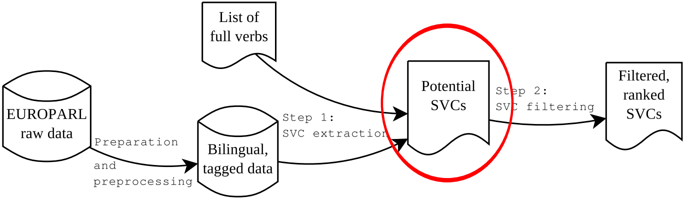
Overall structure of the SVC acquisition procedure



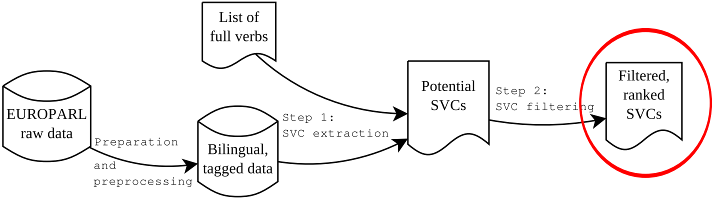
Overall structure of the SVC acquisition procedure



Overall structure of the SVC acquisition procedure

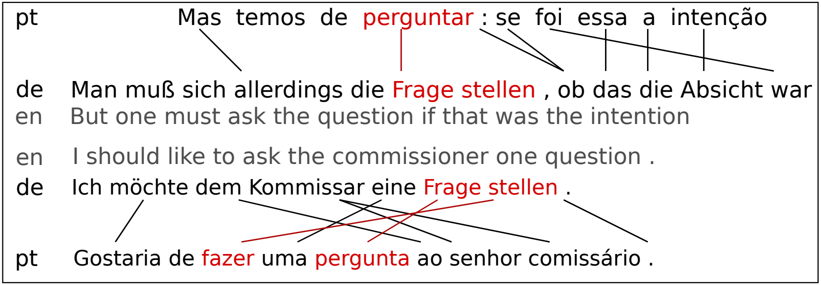


Overall structure of the SVC acquisition procedure



- Goal: acquiring Portuguese SVCs with little preprocessing (POS tagging [Schmid, 1994, Carreras et al., 2004, Padró et al., 2010])
- Parallel corpus: PT-DE portion of EUROPARL, v.3 [Koehn, 2005]
- Starting point: full verbs which semantically correspond to at least one SVC
- Cross-lingual extraction: foreign language as pivot [Bannard and Callison-Burch, 2005]
  ⇒ semantic equivalence and syntactic status
- Monolingual filtering: association measures [Krenn and Evert, 2001, Evert and Krenn, 2001]
  ⇒ strength of correlation

Setting of the cross-lingual step



pt Mas temos de perguntar : se foi essa a intenção

de Man muß sich allerdings die Frage stellen , ob das die Absicht war

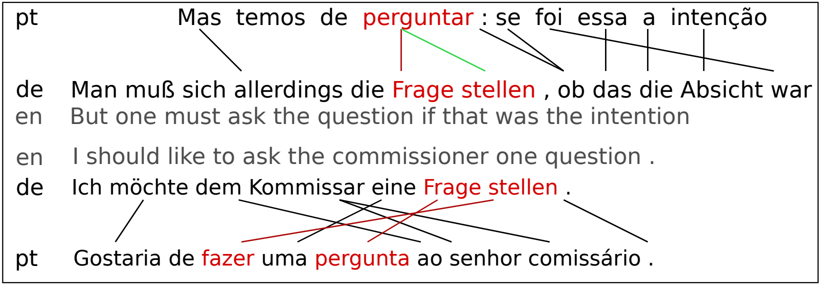
en But one must ask the question if that was the intention

en I should like to ask the commissioner one question .

de Ich möchte dem Kommissar eine Frage stellen .

pt Gostaria de fazer uma pergunta ao senhor comissário .

Setting of the cross-lingual step



pt Mas temos de perguntar : se foi essa a intenção

de Man muß sich allerdings die Frage stellen , ob das die Absicht war

en But one must ask the question if that was the intention

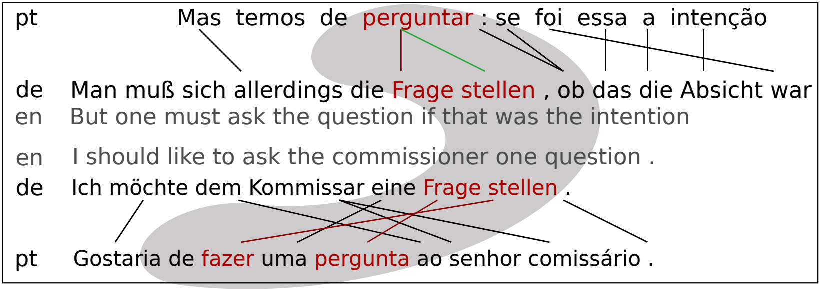
en I should like to ask the commissioner one question .

de Ich möchte dem Kommissar eine Frage stellen .

pt Gostaria de fazer uma pergunta ao senhor comissário .

- Heuristic extension of word alignments

Setting of the cross-lingual step



pt Mas temos de perguntar : se foi essa a intenção

de Man muß sich allerdings die Frage stellen , ob das die Absicht war

en But one must ask the question if that was the intention

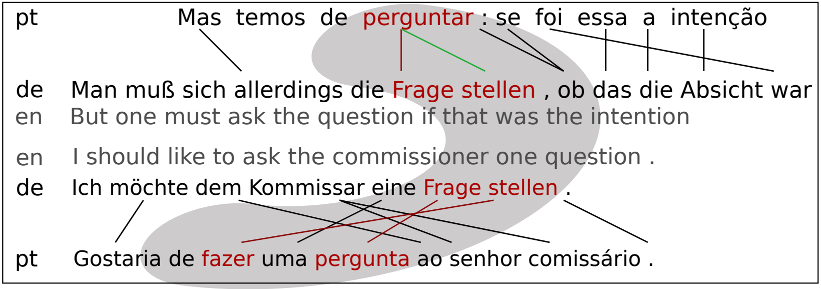
en I should like to ask the commissioner one question .

de Ich möchte dem Kommissar eine Frage stellen .

pt Gostaria de fazer uma pergunta ao senhor comissário .

- Heuristic extension of word alignments

Setting of the cross-lingual step



pt Mas temos de perguntar : se foi essa a intenção

de Man muß sich allerdings die Frage stellen , ob das die Absicht war

en But one must ask the question if that was the intention

en I should like to ask the commissioner one question .

de Ich möchte dem Kommissar eine Frage stellen .

pt Gostaria de fazer uma pergunta ao senhor comissário .

- Heuristic extension of word alignments
- Occurrence thresholds for alignment pairs, restriction to verb-noun (V-N) pairs

Benefits and shortcomings of the cross-lingual step

- Extraction of Portuguese V-N pairs
- Detection of semantically correct SVCs:
  apoiar → *dar apoio*, *dar assistência*
  to support → *to provide support*
- Remaining false positive, compositional V-N pairs:
  apoiar → *exigir apoio*
  to support → *to ask for support*

⇒ Necessary: distinction between compositional and fixed V-N combinations, removal of compositional ones

⇒ Identify correlations between V and N with association measures

Setting and benefits of the monolingual step

PMI ranking of cross-lingual results for *apoiar* ($\times 10^{-7}$):

01.	prestar assistência	160
02.	prestar ajuda	45
03.	conceder ajuda	28
04.	granjear apoio	19
05.	prestar apoio	18
06.	receber ajuda	13
07.	receber apoio	9.8
08.	providenciar apoio	7.5
09.	conceder apoio	7.5
10.	fornecer apoio	6.9
11.	disponibilizar apoio	5.6
12.	dar assistência	4.4
13.	proporcionar apoio	4.3
...		
40.	garantir apoio	0.18
41.	retirar apoio	0.17
42.	prever apoio	0.15
43.	demonstrar apoio	0.11
44.	esperar apoio	0.10
45.	ter apoio	0.041

- Association measures reveal correlations between words

Setting and benefits of the monolingual step

PMI ranking of cross-lingual results for *apoiar* $(\times 10^{-7})$:

01. prestar assistência : 160
02. prestar ajuda : 45
03. conceder ajuda : 28
04. granjear apoio : 19
05. prestar apoio : 18
06. receber ajuda : 13
07. receber apoio : 9.8
08. providenciar apoio : 7.5
09. conceder apoio : 7.5
10. fornecer apoio : 6.9
11. disponibilizar apoio : 5.6
12. dar assistência : 4.4
13. proporcionar apoio : 4.3
. . .
40. garantir apoio : 0.18
41. retirar apoio : 0.17
42. prever apoio : 0.15
43. demonstrar apoio : 0.11
44. esperar apoio : 0.10
45. ter apoio : 0.041

- Association measures reveal correlations between words
- Filtering to increase precision:

Setting and benefits of the monolingual step

PMI ranking of cross-lingual results for *apoiar* ($\times 10^{-7}$):

01. prestar assistência : 160
02. prestar ajuda : 45
03. conceder ajuda : 28
04. granjear apoio : 19
05. prestar apoio : 18
06. receber ajuda : 13
07. receber apoio : 9.8
08. providenciar apoio : 7.5
09. conceder apoio : 7.5
10. fornecer apoio : 6.9
11. disponibilizar apoio : 5.6
12. dar assistência : 4.4
13. proporcionar apoio : 4.3

. . .

~~40. garantir apoio : 0.18~~
~~41. retirar apoio : 0.17~~
~~42. prever apoio : 0.15~~
~~43. demonstrar apoio : 0.11~~
~~44. esperar apoio : 0.10~~
~~45. ter apoio : 0.041~~

- Association measures reveal correlations between words
- Filtering to increase precision:
  → minimum V-N co-occurrence threshold

Setting and benefits of the monolingual step

PMI ranking of cross-lingual results for *apoiar* ($\times 10^{-7}$):

01. prestar assistência : 160
02. prestar ajuda : 45
03. conceder ajuda : 28
~~04. granjear apoio : 19~~
05. prestar apoio : 18
06. receber ajuda : 13
07. receber apoio : 9.8
~~08. providenciar apoio : 7.5~~
09. conceder apoio : 7.5
10. fornecer apoio : 6.9
~~11. disponibilizar apoio : 5.6~~
12. dar assistência : 4.4
13. proporcionar apoio : 4.3
. . .
~~40. garantir apoio : 0.18~~
~~41. retirar apoio : 0.17~~
~~42. prever apoio : 0.15~~
~~43. demonstrar apoio : 0.11~~
~~44. esperar apoio : 0.10~~
~~45. ter apoio : 0.041~~

- Association measures reveal correlations between words
- Filtering to increase precision:
  → minimum V-N co-occurrence threshold
  → remove entries if verb is unlikely to occur in SVCs (diversity)

Setting and benefits of the monolingual step

PMI ranking of cross-lingual results for *apoiar* ($\times 10^{-7}$):

01. prestar assistência : 160
02. prestar ajuda : 45
03. conceder ajuda : 28
~~04. granjear apoio : 19~~
05. prestar apoio : 18
06. receber ajuda : 13
07. receber apoio : 9.8
~~08. providenciar apoio : 7.5~~
09. conceder apoio : 7.5
10. fornecer apoio : 6.9
~~11. disponibilizar apoio : 5.6~~
12. dar assistência : 4.4
13. proporcionar apoio : 4.3
. . .
~~40. garantir apoio : 0.18~~
~~41. retirar apoio : 0.17~~
~~42. prever apoio : 0.15~~
~~43. demonstrar apoio : 0.11~~
~~44. esperar apoio : 0.10~~
~~45. ter apoio : 0.041~~

- Association measures reveal correlations between words
- Filtering to increase precision:
  → minimum V-N co-occurrence threshold
  → remove entries if verb is unlikely to occur in SVCs (diversity)

⇒ Rejection of arbitrary constructions

⇒ Different settings: restrictive thresholds for high precision (hiPrec), loose thresholds for high recall (hiRec)

Evaluation setting

Gold standard:

- 6 initial full verbs:
  ameaçar, *apoiar*, *faltar*, *perguntar*, *prometer*, *responder*
- V-N pairs resulting from cross-lingual step as reference set
- Judged by two native speakers on semantic similarity to full verb (IAA $\kappa = 0.74$ [Cohen, 1960])
- 22 V-N pairs judged as true positive SVCs

Evaluation:

- Computation of precision, (relative) recall and f_1
- Evaluation of results of cross-lingual step (relative recall = 100%)
- Evaluation of final results including monolingual step

Results for the cross-lingual step

	6 full verbs
Precision	**0.26**
Recall	**1.00**
F_1	**0.42**

- Variable precision for individual verbs: $\text{prec}_{faltar} = 1.00$, $\text{prec}_{apoiar} = 0.16$
- Reason: *apoiar* occurs frequently and in many contexts

⇒ 22 SVCs retrieved for 6 full verbs
⇒ Success depends on initial full verb
⇒ Goal: Increase precision while not overly lowering recall

Final results

	Cross-lingual only	PMI hiPrec	PMI hiRec
Precision	0.26	**0.91**	0.61
Recall	1.00	0.45	**0.86**
F_1	0.42	0.61	**0.72**

- Restrictive filtering increases precision, loose filtering hardly lowers recall while improving precision
- Most responsible for improvement: V-N co-occurrence threshold on PMI-ranked list

⇒ Considerable improvement over cross-lingual results: f_1 0.42 → f_1 0.72

⇒ Reliable scores for both precision and recall in different settings

Conclusions:

- Synergy effects by combining cross- and monolingual techniques:
  1. extraction of syntactically and semantically correct expressions
  2. filtering to keep only SVCs
- No complex preprocessing necessary: parallel corpus and POS tags
- Applicable to the need for both solid precision and recall, e.g. language generation and lexicon expansion

Related studies about CP extraction:

- Monolingual: with POS information [Grefenstette and Teufel, 1995, Duran et al., 2011] or association measures [Krenn and Evert, 2001, Evert and Krenn, 2001]
- Cross-lingual: paraphrase detection with pivot idea [Bannard and Callison-Burch, 2005], with deep linguistic analysis [Zarrieß and Kuhn, 2009]

Athayde, M. F. (2001).
Construções com verbo-suporte (Funktionsverbgefüge) do português e do alemão.
Cadernos do cieg, (1):5–68.

Bannard, C. and Callison-Burch, C. (2005).
Paraphrasing with bilingual parallel corpora.
In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 597–604, Stroudsburg, Pennsylvania. Association for Computational Linguistics.

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004).
FreeLing: an open-source suite of language analyzers.
In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004).

Cohen, J. (1960).
A coefficient of agreement for nominal scales.
Educational and Psychological Measurement, 20(1):37–46.

Duran, M. S., Ramisch, C., Aluísio, S. M., and Villavicencio, A. (2011).
Identifying and analyzing brazilian portuguese complex predicates.
In Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World, pages 74–82. Association for Computational Linguistics.

Evert, S. and Krenn, B. (2001).
Methods for the qualitative evaluation of lexical association measures.
In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pages 188–195, Stroudsburg, Pennsylvania. Association for Computational Linguistics.

Grefenstette, G. and Teufel, S. (1995).
Corpus-based method for automatic identification of support verbs for nominalizations.
In Proceedings of European Chapter of the Associaton of Computational Linguistics, pages 98–103.

Koehn, P. (2005).
Europarl: a parallel corpus for statistical machine translation.
In Conference Proceedings: The 10th Machine Translation Summit, pages 79–86, Phuket, Thailand. AAMT.

Krenn, B. and Evert, S. (2001).
Can we do better than frequency? A case study on extracting PP-verb collocations.
In Proceedings of the ACL Workshop on Collocations. Association for Computational Linguistics.

Padró, L., Collado, M., Reese, S., Lloberes, M., and Castellón, I. (2010).
FreeLing 2.1: five years of open-source language processing tools.
In Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'2010), Valletta, Malta.

Schmid, H. (1994).
Probabilistic part-of-speech tagging using decision trees.

In Proceedings of the International Conference on New Methods in Language Processing, Manchester, United Kingdom.

Zarrieß, S. and Kuhn, J. (2009).
Exploiting translational correspondences for pattern-independent MWE identification.
In Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, pages 23–30, Singapore. Association for Computational Linguistics.

Addendum: Specification of the second monolingual filter – context diversity

- Two categories of SVCs:
  - SVCs with light verbs. Verb has very high context diversity, e.g. dar apoio, dar um passo, dar uma resposta, . . .
  - SVCs with nearly idiomatic meaning. Verb has very low context diversity given the mininimum V-N co-occurrence threshold, e.g. correr um risco
- Remove V-N pairs with verbs which have a medium amount of co-occurring nouns

⇒ Filter 1: Consider only V-N pairs with a specific minimum frequency
⇒ Filter 2: Consider only V-N pairs with very high / low context diversity