Corpus-based Acquisition and Analysis of Support Verb Constructions

Magisterarbeit

zur Erlangung des Grades
Computerlinguistin (M.A.)

Institut für Computerlinguistik
an der Neuphilologischen Fakultät
der Ruprecht-Karls-Universität Heidelberg

vorgelegt von Britta Dorothee Zeller

Gutachter: Prof. Dr. Sebastian Padó
Prof. Dr. Anette Frank

Heidelberg, den 15. August 2011

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, dass alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht sind und dass die Arbeit in dieser oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.

Heidelberg, den 15. August 2011

Contents

List of Figures

List of Tables

List of Listings

List of Abbreviations

ADJ	Adjective
AM	Association measure
CC	Conjunction
COMP	Sentential complement
DET	Determiner
DIRREF	Directional reference
EAGLES	Expert Advisory Group on Language Engineering Standards
ELDA	Evaluations and Language resources Distribution Agency
FV	Full verb
FVG	Funktionsverbgefüge (Support verb construction)
LE-PAROLE	Language Engineering - Preparatory Action for linguistic Resources Organization for Language Engineering
LVC	Light verb construction
MWE	Multi-word expression
MWU	Multi-word unit
NLP	Natural language processing
NP	Noun phrase
OWE	One-word expression
PAROLE	Preparatory Action for linguistic Resources Organization for Language Engineering
PERSP	Personal pronoun
PI	Indefinite pronoun
PMI	Pointwise mutual information
POS	Part-of-speech
PP	Prepositional phrase
PREPOBJ	Prepositional object
PRON	Pronoun
PX	Possessive pronoun

RG	Adverb
SbG	Substantive group
STTS	Stuttgart-Tübingen Tag Set
SVC	Support verb construction
SV	Support verb
TT	TreeTagger
VP	Verb phrase
Z	Cardinal

Acknowledgements

I am grateful to Prof. Dr. Sebastian Padó for his supervision. It was a perfect balance between freedom to realise my own ideas and useful suggestions on difficult points, and he always showed interest in my thesis topic. Conversations during coffee/hot chocolate breaks with Thomas, Hiko, Sascha, Michael and Matthias resulted in one or the other flash of inspiration – and in good mood. Thanks as well to Maria Teresa Cartaxo-Haußig and Júlio Cezar Rodrigues who annotated the gold standard. Prof. Dr. Christine Hundt (Universität Leipzig) gave me literature tips and information about SVCs in Portuguese. Dr. Iris Hendrickx provided me with information about preprocessing tools for Portuguese. Cilli, Sina, Johannes and Uli, thanks for your feedback on this text.

Apart from professional support, I would like to thank my parents and my aunt for encouraging and supporting me during my studies. Eva and Heike, without you I would be only half myself; thanks for your impulses, advice and (sister) time. And finally, thanks to my friends – some of them mentioned above –, for making life worthwhile every day and giving me energy and joy.

Obrigada –

Dankeschön –

Vergelt's Gott!

Abstract

This thesis deals with Support Verb Constructions (SVCs) and their automatic acquisiton. SVCs are verbal strucures, consisting of a verb and a noun, which form a unit in both syntactic and semantic aspects. As SVCs are hard to interpret on both counts, they are especially challenging for natural language processing.

We test the possibilities of the acquisiton of SVCs by means of corpus-based methods with few linguistic resources. In particular, we investigate the phenomenon in Portuguese.

The acquisition is carried out in a two-stage approach. First, we extract SVCs using a bilingual parallel corpus. Starting from a list of Portuguese full verbs which approximately correspond to the meaning of an SVC, we use the alignment information to retrieve Portuguese expressions which are semantically appropriate SVCs. In this context, the parallel language acts as a 'pivot' to connect the Portuguese full verb and SVCs. In the next step, we analyse the possibilities to refine the retrieved expressions. It turns out that it is difficult to use information about the support verb's arguments to do such a filtering. Instead, we calculate association measures (e.g. pointwise mutual information) and compile a ranking. This second step, thus, is conducted on the monolingual level.

The experiments show that the presented approach works very well: we retrieve semantically appropriate SVCs and achieve a maximum precision of 91% and a maximum recall of 86% in two different settings. However, the applicability of the approach depends on the contextual diversity of the initial full verb. Heterogeneity complicates the acquisition of high quality SVCs.

Zusammenfassung

Diese Magisterarbeit beschäftigt sich mit Funktionsverbgefügen (FVGs) und ihrer automatischen Akquise. FVGs sind Verbalgefüge aus Verb und Substantiv, die sowohl syntaktisch als auch semantisch eine Einheit bilden. Da FVGs sich in beiderlei Hinsicht an der Grenze eindeutiger Definitionen befinden, sind sie eine besondere Herausforderung für die maschinelle Sprachverarbeitung.

In dieser Arbeit wird ein Ansatz getestet, der mittels korpusbasierter Methoden und mit wenig linguistischer Information FVGs akquiriert. Insbesondere steht dabei das Portugiesische im Zentrum der Aufmerksamkeit.

Die Akquise erfolgt in zwei Schritten. Zuerst werden die FVGs durch den Einsatz eines bilingualen parallelen Korpus extrahiert: Ausgehend von einer Liste portugiesischer Vollverben, die semantisch in etwa einem FVG entsprechen, werden über die Alignierung portugiesische Ausdrücke aufgefunden, die dem zugehörigen FVG semantisch entsprechen. Dabei dient die parallele Sprache als 'Angelpunkt', um das Vollverb und die FVGs im Portugiesischen zu verbinden. Im Anschluss werden die Möglichkeiten ermittelt, diese Ausdrücke zu verfeinern. Es stellt sich heraus, dass es schwierig ist, Informationen über die Argumente des Funktionsverbs für eine solche Filterung zu verwenden. Stattdessen werden Assoziationsmaße (z.B. pointwise mutual information) berechnet und auf diesen basierend ein Ranking erstellt. Der zweite Schritt erfolgt also im Gegensatz zum ersten auf monolingualer Ebene.

Es zeigt sich, dass dieses Verfahren sehr gut funktioniert: Es werden semantisch korrekte FVGs aufgefunden, wobei in zwei verschiedenen Einstellungen eine maximale Genauigkeit von 91% bzw. eine maximale Trefferquote von 86% erreicht werden. Allerdings hängt die Eignung des Verfahrens davon ab, ob das zur Akquise verwendete Vollverb in sehr heterogenen Kontexten verwendet wird. Dies erschwert die Extraktion qualitativ hochwertiger FVG-Listen.

1 Introduction and theoretical background

Support Verb Constructions – verb-noun complexes like the expression *to take a walk*, which act as a syntactic and semantic unit – are a linguistic phenomenon which occurs in many languages. They are a special challenge to natural language processing (NLP): On the syntactic level, their verbs act as a central element of the sentence but are not really a predicate. On the semantic level, they evoke very fine-grained connotations. Thus, it is sometimes difficult to separate SVCs from freely combined words or other complex predicates. On the semantic level, it is hard to adequately interpret their fine-granularity: there are subtle differences between SVCs which contain the same noun but different support verbs (SVs). Hence, finding adequate synonyms for an SVC is not always easy.

In many NLP areas, it is important to consider collocations of any kind – thus, also SVCs. For example, in coreference resolution, SVCs are recognised as potential mentions: SVCs enable anaphoric references where other syntactic constructions like full verbs (FVs) do not, as shown in examples (1) and (2).

(1) Anna made a [proposal]$_1$. John did not like [it]$_1$.

(2) * Anna [proposed to go to the pub]$_1$. John did not like [it]$_1$.

Furthermore, frame semantics and semantic role labelling can be affected by SVCs. Consider the following examples (3) and (4). While semantic frames (in capitals) are mainly assigned to the verb, whose roles (in lower case) are filled by the nouns, such an assignment is not appropriate for the sentence containing an SVC, even if both sentences have the same meaning; see Johnson et al. (2002)[1].

(3) [Priscilla and Gwyneth Molesworth]$_{self\_mover}$ **walked**$_{SELF\_MOTION}$ [in the park]$_{area}$.

(4) [Priscilla and Gwyneth Molesworth]$_{self\_mover}$ took a **walk**$_{SELF\_MOTION}$ [in the park]$_{area}$.

For such applications, it is important to reliably identify SVCs. Quite some work has already been conducted in this area from different points of view, and there are some automatic approaches which achieve good results (see section 1.2). This encourages us to apply a cross-lingual approach to the task of automatically acquiring SVCs for Portuguese, a relatively resource poor language, using only flat syntactic structures (i.e. part-of-speech-tagged data). Our approach is language independent, provided a part-of-speech tagger (POS tagger) for the respective languages is available.

Our aim is to investigate whether cross-lingual techniques and parallel corpora are suitable for the processing of data with few linguistic information for both

[1]The frame annotations in the sample sentences are taken from FrameNet, v.1.5: `http://framenet.icsi.berkeley.edu/` (August 2011, date last accessed).

syntactic and semantic issues, and whether it is possible to create new lexical resources for complex constructions like SVCs in this way. However, we also account for monolingual approaches. In particular, we explore to what extent monolingual methods can improve the bilingually retrieved information, and if there are indications that a combination of mono- and bilingual approaches generally leads to better results.

The remainder of this thesis is organised as follows: this chapter points out the pursued ideas and used methods, presents studies related to the present research question and gives linguistic background about SVCs. Chapter 2 sketches the used corpus and the preparation of the corpus data for our task, and evaluates the performance of these steps. Then, chapter 3 describes the first part of the two-stage approach of this thesis, which is an adaptation of the cross-lingual 'pivot' approach of Bannard and Callison-Burch (2005). Chapter 4 introduces our gold standard based on the results of the pivot procedure and provides intermediate results. Subsequently, chapter 5 describes the second, monolingual step to filter these results: first, we report on our analysis of the context of the SVCs, i.e. their arguments, and the conclusions we draw from it. Then, we present the application of various association measures in different settings and their performance on the overall task. Finally, chapter 6 concludes our work and gives an outlook on possible further investigations.

1.1 Idea and overview of the approach

The goal of our approach is to generate lists of SVCs, i.e. lists of verb-noun pairs which semantically correspond to a given full verb. As dataset, we use the German and Portuguese portion of EUROPARL (Koehn, 2005), a well-known parallel corpus (for a more detailed description, see chapter 2.1). As Portuguese is a relatively resource poor language, a bilingual approach should enable insights which are hard to achieve monolingually. Our basic assumption is that there are FVs which approximately correspond to the semantic meaning of one or several SVCs (cf. section 1.3).

Starting from these FVs, we extract expressions that ideally resemble SVCs on the syntactic level and correspond to the FV's semantic meaning, making a detour via the other, parallel language – in our case, German. Therefore, we exploit the alignment information produced with machine translation techniques and merged with a standard alignment symmetrisation algorithm. This idea is clearly inspired by Bannard and Callison-Burch (2005)'s proposal to use parallel data in another language as a pivot for paraphrase extraction. Their approach has been adapted for our purpose which has a more narrow scope, and achieves good results on the SVC acquisition task. To the best of our knowledge, such a bilingual setting for SVC acquisition has not been attempted before.

Although the pivot approach is a good basis for the acquisition of SVCs, the

quality of the results is not yet satisfying: there are many false positives in the resulting SVC list which should be eliminated.

Thus, we try to subsequently improve the results of the bilingual approach in a second step. We investigate the feasibility of two monolingual techniques: *i)* filtering by using the information about the arguments of the SVC's support verb, i.e. their surrounding context, and *ii)* filtering by calculating association measures on the SVCs. A detailed analysis of the practicability of *i)* reveals that the SV's arguments are realised very differently and that it is hard to cover all relevant constructions. Hence, the second option is implemented. As we will show, the association measure strategy indeed leads to considerable improvements.

Figure 1 illustrates the overall workflow for SVC processing proposed in this thesis.

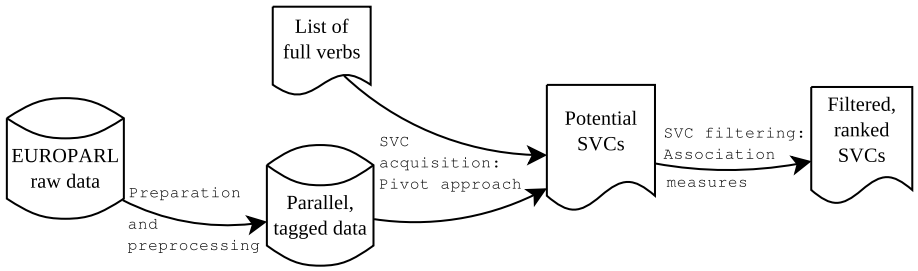

Figure 1: Overall workflow of the SVC processing

1.2 Related work

Many studies have been conducted about different kinds of collocations. A collocation is a phrase whose meaning is not only the sum of its parts, but which has also an own specific meaning. Thus, it is a sort of a superclass for SVCs, idiomatic expressions, verb-particle combinations, etc. (Manning and Schütze, 1999, p. 29).

An important basic task for collocation processing is the design and creation of collocation lexica. Ideally, such lexicographic components should be as complete and correct as possible. There have been some manual efforts to create a complete list of SVCs, e.g. Herrlitz' list for German (Herrlitz, 1973). However, the author points out that completeness is simply infeasible. Nonetheless, the question of how dictionary entries for SVCs should look like has been discussed. For the description of one proposal, see Hanks et al. (2006).

But which methods help to acquire and gather information about SVCs on a large scale? One way is the development of manual or semi-automatic annotations of corpus data. Working with corpora enables an empirically based impression of the distribution, features and behaviour of SVCs. Kamber (2008) carried out an extensive manual corpus study on German SVCs, where a by-product is a large list of SVCs. Hendrickx et al. (2010) developed a specific annotation layer for

Portuguese SVCs on the CINTIL corpus[2], and carried out several studies on the annotated data. In a further study (Duarte et al., 2010), they investigate the SVCs' behaviour and syntactic and semantic features in detail. As well, Fellbaum et al. (2006) present the design and implementation of a resource which makes corpus information about German multi-word expressions (MWE, or MWU for 'multi-word unit') accessible. In their article, the authors point out that one of the main problems is the distinction between SVCs and common verb-noun pairs.

Instead of manually annotating SVCs, Cinková et al. (2006) develop a component to extract lexical information for Swedish SVCs semi-automatically, which is more efficient than purely manual approaches. Collocation lexica as presented in this article, in turn, can be used for further steps, e.g. to improve language generation systems as Smadja and McKeown (1990) did. The idea of establishing such a lexical information resource is also the focus of our approach, however, in a fully automatic way.

Various other approaches rather concentrate on such an automatic SVC acquisition. As this requires a large quantity of data, all these approaches are corpus-based.

Pearce (2002) summarises five approaches since the 1970's to compute collocational probabilities. He also makes an own proposal, using sense information derived from WordNet (Fellbaum, 1998) to determine the substitutability of a word within a (candidate) collocation.

Grefenstette and Teufel (1995) extract SV-argument structures. They search for nominalised forms of a given list of FVs, e.g. *to appeal* → *appeal*, and then locate the corresponding SV, e.g. *make + appeal*. This approach leads to good results. However, the authors claim that it is hard to decide if an extracted SV is a 'real' SV or if they just extracted a frequent but lexically free compositional verb-noun phrase.

Krenn and Evert (2001) conducted a study on the capability of various association measures to establish rankings for different types of collocations, applied in different corpus domains. For the calculations, a minimum cooccurrence threshold of 3 is used. They reveal that the applicability of association measures is domain-independent but observe differences between various collocation patterns. Interestingly, the approach achieves better precision on full form data than on base forms, i.e. lemmatised words. In a second study (Evert and Krenn, 2001), the authors investigate the measures' behaviour for hapax legomena (i.e. words which occur only once in the corpus) and words occurring only twice. The poor performance of these words leads to the conclusion that the exclusion of low-frequency data is legitimate.

As common association measures are linguistically rather unmotivated, Wermter and Hahn (2004) propose an approach for measuring collocativity, which accounts

[2]available from ELRA, see `http://catalog.elra.info/product_info.php?products_id=1102` (August 2011, date last accessed)

for linguistic information. In particular, their proposed calculation includes the modifiability – or semantic fixedness – of a collocation. That is, it is measured whether additional lexical material can be introduced into the collocation's nominal group and if so, to what extent this material is prescribed. The authors report that their measure performs better on three different kinds of collocations (including SVCs) than log-likelihood, t-test and frequency.

All approaches presented up to this point are based on monolingual techniques. There are, however, also some bi- and multilingual approaches which have been carried out to acquire collocations.

Smadja et al. (1996) use an automatically aligned corpus to extract a broad range of collocations (including SVCs) for the creation of a bilingual collocation lexicon. The idea of generating a bilingual lexicon is different from our approach, and of course, the authors must choose another starting point than we do: using an initial list of English *collocations*, they compute the most probable corresponding collocations in French with statistical methods. They achieve up to 70% of correctly translated collocations.

Bannard and Callison-Burch (2005), whose idea serves as main motivation for the present thesis, do not concentrate on collocations but carry out their acquisiton method on multi-word paraphrases. However, these paraphrases are meaningful units and thus, some kind of collocation. Their approach is explained in detail in section 3.1.

Moirón and Tiedemann (2006) also use multilingual information to retrieve different kinds of MWEs and is thus akin to our approach. Their aim is to distinguish literal ('transparent') from idiomatic ('opaque') MWEs, measuring translational entropy and the proportion of correct bidirectional alignments to determine the differences between literalness and idiomaticity. Their assumption is that expressions have a literal meaning, if their translation is a combination of the words' isolated translations (i.e. bidirectional alignments). In contrast, translating the individual words of an idiomatic expression does not reflect the overall sense. Contrary to our initial list of full verbs, Moirón and Tiedemann start with a list of support verbs, known for triggering both idiomatic and literal expressions. Furthermore, the usage of full syntactic parses is an important difference to our approach. The experiments have been carried out on V-PP pairs and achieved up to 93% uninterpolated average precision, disregarding the PP's prepositions.

Zarrieß and Kuhn (2009) exploit automatically created 1:n alignments to acquire candidate MWEs, like SVCs or verb-preposition combinations, which semantically correspond to a given FV. In a second step, the retrieved MWEs are filtered using dependency-parses in both languages: only those target language words which are situated on a common parse tree branch are considered. Finally, they discard all target words which lead to a decrease in correlation between source and target expression. The usage of dependency parses is the main difference to

our method.

The results of all these studies show that it is worthwile to employ the knowledge contained in parallel data to process MWEs of any kind.

We have already presented some examples for typical SVCs and mentioned that these constructions occur in many languages. As examples, see Athayde (2001) for Portuguese, von Polenz (1963) for German, Butt (2003) for Urdu, Hong et al. (2006) for Korean, Danlos (1992) for French and Cinková et al. (2006) for Swedish and Czech.

But what is the exact definition of SVCs? There exist different terminologies and explanations which we try to sort and condense into a common definition in the following section. Then, we provide information concerning Portuguese SVCs.

1.3 The syntax and semantics of Support Verb Constructions

As von Polenz (1963) notes, language tends to nominalisation. SVCs – or 'Light Verb Constructions', LVCs, as some authors call them – are a prime example for such nominalisations. They are basically defined as a structure consisting of a support verb and a nominal predicate (Athayde, 2001, p. 10):

$$\text{SV} + \text{N}_{pred}$$

The SVC as a whole functions as predicate. The support verb is a special kind of verb which can partially or completely lose its meaning[3]. The nominal predicate, in turn, is actually reflecting the meaning of the collocation. It can consist of a non-prepositional (example (5)) or prepositional (example (6)) object of the SV (Athayde, 2001).

(5) dar esperança
to give hope

(6) estar na dúvida
to be in doubt

Another variant to classify SVCs is the question whether they are semantically substitutable by FVs. While there are no corresponding FVs for the examples in (5) and (6), examples (7) and (8) show that other SVCs can be replaced by FVs. The FVs can have another word stem than the SVC ($fim \rightarrow acabar$) as well as the same ($resposta \rightarrow responder$); see Athayde (2001).

(7) (a) Vou *pôr fim* a isto.
I will *put an end* to this.

[3]However, Brugman (2001) claims that SVs are not meaningless but just highly abstract and that they even have polysemous meaning.

(b) Vou *acabar* com isto.
I will *stop* this.

(8) (a) *Dá*-me *uma resposta*!
Give me *an answer*!

(b) *Responda*-me!
Answer me!

Cases as in examples (7) and especially (8) serve as basis for our approach, as they establish a connection between SVCs and FVs.

The replaceability of SVCs by individual verbs attests an aspect about SVCs brought up from a psycholinguistic perspective: they are perceived as one coherent unit by the speaker, resulting in an easy perception by language learners, but posing problems in production. These problems are due to the fine-granular differences in meaning of the SVs (see Grefenstette and Teufel (1995), Cinková et al. (2006)), illustrated in examples (9) and (10)[4].

(9) to *take* a bath

(10) to *have* a bath

One might ask whether there are any rules to determine which SVs may occur within a specific SVC and which may not. To the best of our knowledge, there are none – this fact makes the SVC delimitation task more difficult for NLP. However, we expect that the perception of SVCs as one unit is quantitatively reflected in the corpus data and thus provides some indication.

An SVC can be used for syntactic variation, e.g. passive constructions or anaphoric references. As a specific unit, it also underlies specific syntactic restrictions. For example, it is not possible to establish a coordination construction between an SVC and a common verb-noun complex, even if they use the same verb as in example (11), taken from Athayde (2001, p. 14):

(11) * *Ela levou o amigo a casa e ao desespero.*
* She took the friend home and to despair.
* She drove her friend home and to despair.

Apart from the syntactic features, SVCs lead to a specific behaviour on the semantic level: they enable a fine-grained adjustment of the aktionsart, where a simple FV permits only one option (see Eisenberg (2006)). Butt and Geuder (2001) and Butt (2003) carry out comprehensive studies on SVs and SVCs in both syntactic and semantic aspects.

[4]The examples are taken from Butt (2003).

Although the formal definition of SVCs shown at the beginning of this section sounds simple and straightforward, there is a lot of controversy about their definition and delimitation from other phenomena on both the syntactic and the semantic level. These disagreements stem from the SVCs being on the border between functional and lexical issues. For example, Bußmann (2008, p. 209) as well as other authors (for a summary, see Athayde (2001, p. 42)) only consider constructions with a prepositional object as real SVCs. Even the questions if SVCs can be regarded as a separate syntactic class or not (see Butt (2003, p. 6) versus Eisenberg (2006, p. 309)) and whether there is a reason to pay so much attention to SVCs (van Pottelberge, 2001), are disputed.

As to the delimitation of SVCs to other constructions, e.g. so-called 'Streckformen' in German as well as idiomatic expressions or periphrases in general, it is difficult to define clear criteria. Athayde (2001) and Döll and Hundt (2002) present extensive analyses for Portuguese (compared to German) and point out that no criterion is clear-cut. The authors also present several tests which help to decide whether an expression is an SVC or not. The most widely accepted test is the test for anaphoricity, see examples (12) and (13), taken from Döll and Hundt (2002, p. 154).

(12) Tenho livros. – O que tenho? – Tenho-os.
I have books. – What do I have? – I have them.

(13) Tenho dúvidas. – * O que tenho? – * Tenho-as.
I have doubts. – * What do I have? – * I have them.

The difficulties in defining and demarcating SVCs shown in this section suggest that automatic processing of these complex structures will be complicated as well. However, we expect to be able to acquire lexical information about Portuguese SVCs at least at a coarse level. Therefore, the next section will take a closer look at Portuguese SVCs and their challenges in particular.

1.4 Support Verb Constructions in Portuguese

Examples for high-frequent Portuguese SVs are *fazer* ('to make'), *dar* ('to give'), *trazer* ('to bring'), *tomar* ('to take'), *ter* ('to have'), *pôr* ('to put') and *oferecer* ('to offer') (Gärtner, 1998, p. 78 f.). There are many SVCs enumerated in the literature containing these SVs, e.g.[5]:

(14) fazer um brinde
to give a toast

(15) dar apoio
to give help

[5]Examples (14) and (15) taken from Döll and Hundt (2002).

Gärtner (1998, p. 112 f.) provides the syntactic models representing the underlying structure to realise a correct Portuguese sentence with an SVC[6]:

(SV + SubstantiveGroup)
→ Faz frio. – It is cold.

Subject + (SV + SubstantiveGroup)
→ A mulher deu um grito. – The woman let out a scream.

Subject + (SV + PrepositionalGroup)
→ A rapariga caiu no pranto. – The girl burst into tears.

Subject + (SV + SubstantiveGroup) + indirectObject
→ O chefe pôs fim à discussão. – The boss put an end to the discussion.

Subject + (SV + SubstantiveGroup) + prepositionalObject
→ O orador fez referência à situação. – The speaker refers to the situation.

Subject + (SV + SubstantiveGroup) + directionalReference
→ O João deu um passeio pelo centro. – John took a walk to the centre.

According to Athayde (2001, p. 48), these patterns can be modified for some SVCs, i.e. they allow for the insertion of adverbs, like *muito* in *estar muito em voga* (lit. 'to be very in vogue').

The SVC examples presented in Athayde's introduction, inlcuding examples (16)-(18), suggest two facts: *i)* SVCs with a prepositional nominal predicate are common in Portuguese, and *ii)* such SVCs are likely to occur in newswire and political texts. This fact is not surprising, as these domains stand out because of numerous nominalisations.

(16) A TAP *está em greve.*
TAP airlines *is on strike.*

(17) A taxa de desemprego *continua em queda.*
The unemployment rates *continue to fall.*

(18) A Angola *está* de novo *em guerra.*
Angola *is* again *at war.*

These characteristics could also apply to the corpus we use (see section 2.1). However, prepositional SVCs will be discounted for our studies. The reason is simple: as Moirón and Tiedemann (2006) stated, prepositions are highly amibguous

[6]Additionally, there are some variances of these patterns, not listed here, which are due to specific word order restrictions.

in translation and pose problems to automatic word alignment. Considering prepositional objects for SVC acquisition would lead to more noise in our data. Hence, using only non-prepositional SVCs is more promising for high-quality SVC extraction. Nonetheless, our approach is applicable to prepositional SVCs as well.

2 Corpus and data preparation

This chapter introduces the corpus we use and explains the preparation of the corpus data which is necessary to implement our SVC acquisition approach. Finally, we provide a qualitative evaluation of the performance of these steps.

2.1 The corpus

Figure 2: 'Choose your corpus', IKEA Walldorf (source: author)

First of all, a data-driven approach needs a text corpus. One big advantage of data-driven methods is the proof of concept: sufficient corpus evidence for a specific phenomenon justifies the assumptions one has made.

For bi- or multilingual methods, it is necessary to have access to data in several languages, i.e. to a comparable or parallel corpus. Comparable corpora consist of texts in several languages having the same topic. The texts do not necessarily contain exactly the same sentences. In contrast, parallel corpora contain texts which are translations of each other or of the same source. (Lemnitzer and Zinsmeister, 2006, p. 198) Parallel corpora have a more general applicability, but it is easier to acquire comparable corpus data. However, there are natural resources for parallel texts. We will come back to this point shortly.

Cross-lingual methods are useful for resource-poor languages, i.e. languages for which hardly any or even no NLP tools (taggers, parsers, etc.) are available. One can draw inferences from a parallel corpus without such tools. Exploiting parallel data can happen explicitly, e.g. by means of annotation projection from a resource-rich into a resource-poor language (as an example, see Padó and Lapata (2009) for the projection of semantic roles), or implicitly by making use of automatically created alignments (for example, Kuhn (2005) inducts syntactic information out

of the alignments).

We opt for the Portuguese and German portion of the EUROPARL corpus[7].

EUROPARL – a parallel corpus. The texts contained in EUROPARL stem from the proceedings of the European Parliament since 1996 and are predominantly speeches of its members. As the texts must be comprehensible to the representatives of all member states, they are translated. By this, a huge collection of parallel data is created which – due to the EU enlargement – grew up to 21 languages (EUROPARL v.6). However, there is a caveat concerning the term 'translation': although the parallel texts have the same content on a discourse-semantic level, free translations are frequent (see Zarrieß and Kuhn (2009)). As an example, consider the following parallel sentences in Portuguese, German and English in their literal translation and their actual cooccurrence in EUROPARL[8]:

(19)

Solicitamos à	*mesa*	*que*	*investigue*	*este*	*facto.*
Wir bitten den	Tisch	dass	untersucht werde	dieser	Fakt.
We ask the	table	that	will be investigated	this	fact.

Wir fordern das Präsidium auf, sich mit dieser Angelegenheit zu befassen.

Can we ask the bureau to look into this fact .

The sentences do not have the same content on the word level. Especially the Portuguese word *mesa* ('table') cannot be translated literally as this word is not used in such an expression in the target languages. Such differences are both an opportunity and an obstacle: slight variations of individual words or short expressions can reveal new lexical information and are exactly what we are looking for in our approach. However, lexical investigations are aggravated if the complete sentence is freely translated.

Reasons for choosing EUROPARL. There are various reasons to use EUROPARL for our study. First, there are only few parallel corpora for the language pair in question. Of course, parallel corpora exist for both languages separately (e.g., an overview of corpora in Portuguese is available on the Linguateca web page[9]; for German, see Lemnitzer and Zinsmeister (2006)). But Portuguese-German material is rare, and most of the corpora are either not freely available (e.g. the ELDA MLCC corpus[10]) or only useable via web interfaces (e.g. Tiedemann's OPUS project[11]). EUROPARL is freely downloadable.

Furthermore, the corpus should be big enough and not too domain specific. Although EUROPARL indeed covers a specific type of text, the covered domain can be used more generally than the alternatives, e.g. the JRC-Acquis corpus, a

[7] version 3 (September 2007, see `http://www.statmt.org/europarl` (August 2011, date last accessed)

[8] If not otherwise indicated, the following translations into English are always the original sentences in EUROPARL.

[9] see `http://www.linguateca.pt/corpora_info.html` (August 2011, date last accessed)

[10] see `http://www.elda.org/catalogue/en/text/W0023.html` (August 2011, date last accessed)

[11] see `http://opus.lingfil.uu.se/` (August 2011, date last accessed)

strongly specific corpus (Steinberger et al., 2006) which mainly contains legal texts of the European Union. Another parallel corpus which covers the desired language pair is the EMEA corpus, providing texts about the nature and application of medicine (Tiedemann, 2009). This collection, in turn, is discarded due to its text structure: it is not likely to find SVCs in such factual texts. To the best of our knowledge, there is no parallel corpus for the desired language pair which is balanced in terms of the content.

Hence, EUROPARL seems to be the best fitting choice.

EUROPARL's data design and tools. EUROPARL's data consists of plaintext and XML tags. The latter contain metainformation and mark speakers and sections in the text (chapters, paragraphs). Additionally to the text data, EUROPARL's web page provides some Perl scripts which execute necessary corpus preprocessing steps.

The Portuguese portion amounts to ca. 1,441,000 sentences, the German one to ca. 1,516,000[12].

The EUROPARL data that we used for our study has an average sentence length of 25.99 words in Portuguese and 22.61 words in German. Note that there are many sentences which are rather headlines for specific operations or events during the sessions, e.g. 'resumption of the session' or 'votes'. As to the fact that the lion's share of the corpus consists of speeches, phrases like 'Mr president, commissioners, ladies and gentlemen' are frequent and occur repeatedly.

2.2 Preprocessing

For the realisation of the current approach, we carry out several preprocessing steps. We word-align the parallel corpus and generate additional information about the text data, i.e. POS tags and lemmas.

Finally, all available information and data should be presented in a compact and well-processable format. It is important that the different annotation levels (i.e. lemmas, POSes and alignments) can be mapped onto each other and that these annotations are as reliable as possible.

The steps and the employed machinery used for preprocessing are described in this section. Figure 3 gives an overview of the preprocessing steps.

2.2.1 Aligning EUROPARL

Sentence alignment. The first step is the alignment of the Portuguese and German EUROPARL data, using the Perl script `sentence-align-corpus.perl` provided as described in section 2.1. It aligns sentences of two given languages according to the algorithm of Gale and Church (1993) which assumes that the length of corresponding sentences in language_a and language_b is about the same, i.e.

[12] see `http://www.statmt.org/europarl/archives.html#v3` (August 2011, date last accessed)

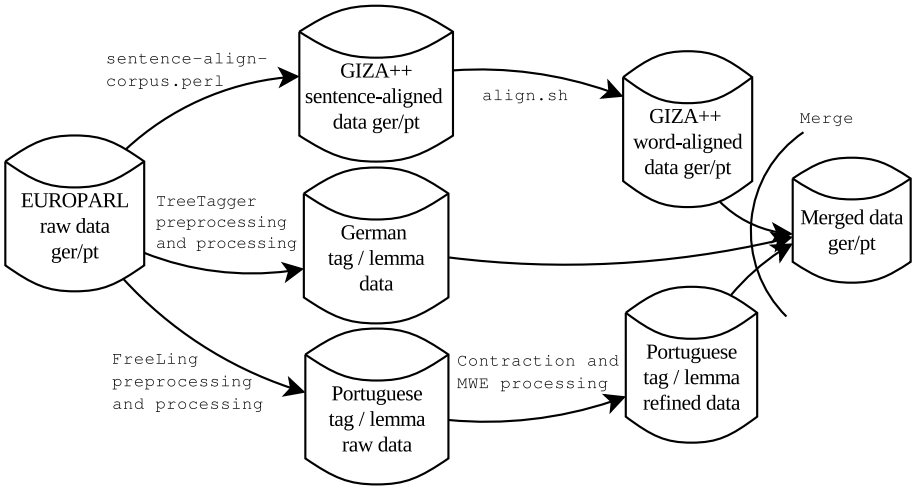

Figure 3: Preprocessing pipeline

'that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences' (Gale and Church, 1993, p. 1). The algorithm results in the sentence alignment which achieves the highest probability under this assumption, having removed those document parts for which no clear match was found in the other language. Moreover, it executes some data cleansing, such as deletion of blank lines, double blanks etc.

In this way, the initial amount of potential parallel data (see section 2.1: about 1,441,000 sentences in Portuguese) is reduced to about 1,268,000 effectively usable aligned sentence pairs.

Word alignment Word alignment is done with GIZA++ (Och and Ney, 2003), a statistical machine translation toolkit which implements the training of IBM translation models 1-5 (Brown et al., 1993) and a Hidden Markov alignment model (Vogel et al., 1996). GIZA++'s execution is processed with a shell script (`align.sh` in figure 3), internally accessible in the Department of Computational Linguistics in Heidelberg (author unknown). It preprocesses the text (XML tag removal, lower case conversion of all characters, again removal of datasets with unaligneable sentences etc.) and then runs GIZA++ itself. GIZA++ is used with the standard settings, producing Viterbi alignments for IBM model 4. The outcome are two unidirectional alignments, i.e. one alignment $\text{language}_a \rightarrow \text{language}_b$ and one alignment $\text{language}_b \rightarrow \text{language}_a$. For every sentence alignment, a probability value is assigned (do not confuse this probability value with *word* alignment probability).

The word alignment step results in 1,106,987 word-aligned sentence pairs.

2.2.2 POS tagging and lemmatisation

POS tags indicate a word's class, for example noun, verb or determiner. Most POS tagsets – definitions about which POSes can be assigned to a token – discriminate word classes both on a coarse level (e.g. nouns, verbs, adjectives) and on a fine-grained level (e.g. possessive pronouns like 'my' or 'its', and personal pronouns like 'you' or 'he'). Word class information is useful for various NLP tasks as it gives indications for a word's neighbours (Jurafsky and Martin, 2008, p. 137). In this way, one can derive typical POS patterns for a specific construction in a specific language. For example, in English, determiners normally precede nouns or adjectives but not verbs.

Lemmatisation is the process of mapping a word of any morphological form to its root, e.g. shoes → shoe. Knowing a word's root helps to observe its behaviour independent of the surface realisation (Jurafsky and Martin, 2008, p. 80).

Recalling the objective of extracting and analysing SVCs, we obviously need to work on a more abstract level than the word level. For this reason, the Portuguese and German text portions are processed by a POS tagger and lemmatiser, providing information about sequences of the word's POSes and basic forms. This generalisation ensures a more abstract access to the corpus data and consequently, higher occurrence frequencies for each observed token.

Although it would be interesting to evaluate how well the applied taggers perform on our data, such an analysis is out of the scope of this thesis.

Processing for German. For German, there are several POS taggers or models for statistical taggers available, e.g. Stanford (Toutanova and Manning, 2000) or TnT (Brants, 2000). We use TreeTagger (Schmid, 1994), developed at the University of Stuttgart. TreeTagger (TT) processes both probabilistic POS tagging and lemmatisation. It uses a binary decision tree to compute transition probabilities for POSes. As for lemmatisation, it uses a fullform and a suffix lexicon, both built from a tagged training corpus, and a fallback default entry.

The used tagset is the Stuttgart-Tübingen Tag Set, or STTS (Schiller et al., 1995). As for the model, TT's web page supplies a parameter file for German. However, we run TT with the parameter file provided by the Department of Computational Linguistics in Heidelberg, which is an earlier version of the file distributed on TT's web page and comprises about 35 MB. The tagger process should not be executed with lower case text as GIZA++ returns by default, because the upper case information is important to detect German nouns and nominalised verbs and adjectives[13].

As TT is based on Latin1 encoding, but the EUROPARL data are in UTF8 format, a conversion with the Linux shell command `iconv` is necessary. However, some UTF8 characters have no corresponding Latin1 character, e.g. different

[13]Thus, for the German corpus portion, the lower case conversion in `align.sh` was deactivated.

UTF8 variants of hyphens. `Iconv` simply ignores these cases. In order to avoid too many ignored characters, some clear and intuitive conversions are treated with a preceding `sed` command, e.g. the conversion of '–' to '-'[14]. The major part of the unconvertable and hence removed characters occurs in proper names and named entities, e.g. ć in 'Mladić', the name of a delegate. Thus, the token given to the tagger is 'Mladi'. Despite this modification, the tagger's behaviour remains the same as for proper names: the token is tagged as named entity with an unknown lemma.

Before running the tagger, sentence boundaries are marked so that they cannot be changed due to the tagging process. For the German portion of word-aligned data, TT needs about three minutes on a Linux server with two Intel Xeon E5520 CPUs à 2.27 GHz and 24 GB RAM.

Processing for Portuguese. For Portuguese, there are hardly freely available tools and trained models for tagging and lemmatising. According to Iris Hendrickx, researcher at the Universidade de Lisboa, the reason for this is the PAROLE dictionary which is used in several POS taggers for Portuguese. It is not freely available but distributed by ELDA[15] (personal communication). We decided to use FreeLing, version 2.2 (Carreras et al. (2004), Padró et al. (2010)), for both tagging and lemmatising. This comprehensive toolkit was developed at the Universitat Politècnica de Catalunya, covering many NLP tasks, such as POS tagging, chunking, parsing, named entity extraction and coreference resolution. It works for several, predominantly Romance languages, with Portuguese being added in v.2.1. However, not all tasks are available for all languages, possibly due to a lack of training data. E.g., there is no Portuguese chunking component.

There are few reliable information about which tagset FreeLing uses for Portugese. According to FreeLing's user manual[16], the default tagset is PAROLE, but there is no tagset documentation for Portuguese[17].

Let us have a look at the origins of the tagset. LE-PAROLE (**L**anguage **E**ngineering - **P**reparatory **A**ction for linguistic **R**esources **O**rganization for **L**anguage **E**ngineering[18]) is a project ordered by the European Commission running from 1996 to 1998 (ELRA, 1996). It aims at developing standardised annotated corpora and lexica in the languages of the member states of the European Union. There is no centrally administered PAROLE web page and the information provided by the formerly participating universities are partially contradictory. Nonetheless, the following consistent information could be found: the tagset used is based

[14] The whole list of `sed` conversions is shown in appendix A.

[15] see `http://catalog.elra.info/product_info.php?products_id=765` (August 2011, date last accessed)

[16] see `http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf` (August 2011, date last accessed)

[17] see `http://nlp.lsi.upc.edu/freeling/index.php?option=com_content&task=view&id=18&Itemid=47` (August 2011, date last accessed)

[18] see `http://www.ilc.cnr.it/viewpage.php/sez=ricerca/id=63/vers=ita` (August 2011, date last accessed)

on the standards established by EAGLES (Expert Advisory Group on Language Engineering Standards[19]). It consists of basic tags for all languages which are then customised language-specifically (ELRA, 1996). There is an overview of the general morpholocigal features available on `http://www.ub.edu/gilcub/SIMPLE/reports/parole/parole_morph/paromor_2.html#2.4` (August 2011, date last accessed), but for the language-specific adaptations in Portuguese, no information has been found. However, comparing Portuguese text tagged with FreeLing and the tagset for Spanish provided by FreeLing shows that these two tagsets correspond to each other to a large extent. So, one can widely refer to the tagset explanations given in `http://nlp.lsi.upc.edu/freeling/doc/userman/parole-es.pdf` (August 2011, date last accessed). There are only some small differences, e.g. specific tags in Spanish for auxiliary verbs (V**S**IP1S0 for *soy* ('I am') and V**M**IP1S0 for *ando* ('I go')), whereas Portuguese uses the same tags for both auxiliary and main verb (V**M**IP1S0 for both *sou* ('I am') and *ando* ('I go')).

FreeLing comes with a dictionary which contains about 908,000 word forms corresponding to about 105,000 POS-lemma combinations[20]. As to the encoding, it also requires Latin1 format. However, FreeLing can convert the text from UTF8 to Latin1 internally. Once again, we execute a preconversion of clear cases with the `sed` command.

Sentence boundaries are again marked before running FreeLing for the Portuguese data with the default settings[21]. This process takes about 47 minutes on a Linux system with an Intel Pentium E5300 CPU à 2.6 GHz and 2 GB RAM.

2.2.3 Further processing of the tagged data

FreeLing's output involves some features which are linguistically reasonable but problematic for our purposes. So, after having tagged the data, some revisions are necessary. They are described in this section.

FreeLing carries out a retokenisation of the given data in two ways. On the one hand, it merges fixed MWEs to one entity, e.g. *em vez de* ('instead of') becomes *em_vez_de*. On the other hand, FreeLing retokenises the text by decomposing contractions into its individual components. For example, the preposition-determiner contraction *do* ('of the') is split into the preposition *de* ('of') and the definite article *o* ('the'). Example (20) illustrates the original sentence (in italics), its lemmatised output and the English translation, the focused phenomenon being highlighted in boldface.[22]

(20) *Declaro reaberta a sessão **do** Parlamento Europeu ...*
Declarar reaberto o sessão **de o** Parlamento Europeu ...
I declare resumed the session of the European Parliament ...

[19] see `http://www.ilc.cnr.it/EAGLES96/intro.html` (August 2011, date last accessed)

[20] see `http://nlp.lsi.upc.edu/freeling/index.php?option=com_content&task=view&id=23&Itemid=58` (August 2011, date last accessed)

[21] The default settings employ an HMM-based tagger similar to the approach of Brants (2000).

[22] Note that the lemmatised form of the feminine article *a* is the masculine form *o*.

There are more than 30 preposition-determiner contractions of that kind, without counting morphological variance concerning gender and number. However, the words forming such a contraction can also occur in succession without being contracted, as in example (21).

(21) *Finalmente , quero salientar a importância* ***de o*** *debate sobre*
Finalmente , querer salientar o importância **de o** debate sobre
a bioética [...]
o bioética [...]
Finally, I should like to stress that the debate on bioethics [...]

FreeLing behaves in the same way for reflexive pronouns and indirect objects realised as a pronoun. In Portuguese, these pronouns are usually attached to the verb by a hyphen; see example (22).

(22) ***Refiro-me*** *, nomeadamente , à alteração nº 2 [...]*
Referir me , nomeadamente , a o alteração nº 2 [...]
I am thinking in particular of amendment no 2 [...]

Both token modifications are undesirable as they impede us to map the tagged data on the original text, which is necessary in order to use both POS/lemma and alignment information at the same time. Contraction decomposition occurs in 78.7% of the sentences, MWE composition in 26.5% of the sentences. Since such a high amount of data is concerned, we did the following:

As to the contractions and reflexive pronouns, it was found that the contracted forms (example 20) in the output are much more frequent than the respective words in succession (example 21). So, we basically recompose all occurrences of the respective preposition-determiner and verb-reflexive sequences to contractions. For the four preposition-determiner pairs which have by far the highest frequency in the successive form (i.e. as in example (21)), we prevent this recomposition[23]. Recomposed words are labelled with a newly created POS tag, consisting of the prefix 'CP' indicating the composition, and the tag of the determiner; e.g. 'CPDAFS0' stands for a compound with a definite article, feminine, singular (see example (23), showing both lemmas and POS tags).

(23) *Aprovação* ***da*** *acta*
Aprovação **do** acta
NFS000 **CPDAFS0** NCFS000
Approval of the minutes

For reflexive pronouns and attached indirect objects, we modify the beginning of the verbs' POS tag from 'V' to 'VREF'; e.g. 'VREFMIP1S0' for a combination of verb and reflexive pronoun, 1^{st} person, singular; see example (24).

[23]The costs would outweigh the benefits if we did this for all potential non-contractions. For example, the sequence *em isto* ('in this here') occurs in only one sentence, while the contraction *nisto* occurs in 207 sentences. In contrast, the sequence *de um* ('of a') occurs about 36,000 times.

(24) ***Refiro-me*** *[...]* *à* *alteração* *nº* *2* *[...]*
Referir-me [...] ao alteração nº 2 [...]
VREFMIP1S0 [...] CPDA0FS0 NCFS000 NCMS000 Z [...]
I am thinking [...] of amendment no 2 [...]

Concerning the merged MWEs, we can simply split the expression into its individual units. Every token is labelled with the tag prefix 'DIV' indicating the divison, and the tag which has been assigned to the MWE – in most cases adverb and preposition tags –, e.g. 'DIVRG' for the tokens of a split adverbial MWE like *por exemplo* ('for example'); see example (25).

(25) ***Por*** ***exemplo*** , *fala-se* *de* *passaportes* *para*
Por **exemplo** , falar-se de passaporte para
DIVRG **DIVRG** Fc VREFMIP3S0 SPS00 NCMP000 SPS00
bovinos .
bovino .
NCMP000 Fp
For example , there is talk of cow passports .

Although the described heuristics apply in most cases, some cases of overgeneration have to be sorted out. Therefore, we compare the number of lemmas produced and the number of aligned tokens in each sentence. If there is not exactly the same amount of lemmas and tokens, the dataset is removed. We will give some performance figures in section 2.2.4.

The preprocessing steps described up to this point are largely performed by a shell script; some steps are part of a Java[24] implementation which also contains all following procedures.

2.2.4 Merging all data

As a final step of the preprocessing, all information collected are put together, i.e. the original sentences in both languages, their POS and lemma information and their unidirectional alignments and alignment probabilities. Every sentence is labelled with a unique ID. The data are written into a single file of the structure shown in figure 4 (illustrated by the first sentence in the corpus).

Starting with an amount of 1,106,987 word-aligned sentences, the tagging and data merge process leaves us with 982,039 datasets (about 23,200,000 tokens in Portuguese and 21,600,000 tokens in German). This corresponds to a loss of 11.28%, stemming from the restrictions explained above. In return, the resulting datasets are supposed to be very clean.

Henceforth, we refer to this amount of data if we talk about the *corpus*. The data about *one* sentence pair as in figure 4 is called a *dataset*.

[24] `http://www.oracle.com/technetwork/java` (August 2011, date last accessed)

```
<sentID=1>
reinício da sessão
reinício do sessão
NCMS000 CPDAFS0 NCFS000
wiederaufnahme der sitzungsperiode
Wiederaufnahme d Sitzungsperiode
NN ART NN
0: 0    1: 1    2: 2    3: 3
0: 0    1: 1    2: 2    3: 3
0.00357598
8.42753E-4
```

Figure 4: Collection of all generated data

2.3 Alignment symmetrisation

As mentioned in section 2.2.1, GIZA++ produces two unidirectional word alignments, $A1$ and $A2$, for every sentence pair of language$_a$ and language$_b$:
$A1 = a \rightarrow b$ and
$A2 = b \rightarrow a$.

These alignments often contain contrary information, i.e. alignment $A1$ achieves better results for some tokens than alignment $A2$, but leaves other tokens unaligned which are well solved by alignment $A2$. It is desirable to symmetrise the two alignments into a so-called phrase-based alignment which is consistent with both word alignments. Phrase-based alignments typically achieve wider coverage and better alignments.

Techniques for the creation of phrase-based alignments are well-known. For example, Koehn et al. (2003) present a comparison of phrase-based translation models created on the basis of word alignments, syntactic phrases and phrase alignments. The former approach is based on the 'refined alignment' of Och et al. (1999) and is frequently used for the reconciliation of unequal word alignments. It is also applied in our approach and is briefly described in the following paragraphs.

The initial set A of alignments $A1$, $A2$ between language$_a$ and language$_b$ is the intersection of $A1$ and $A2$. It contains the most reliable information we get from GIZA++, i.e. high precision, but covers a small amount of sentences, i.e. low recall (see Koehn et al. (2003)). Thus, A is incrementally extended in two ways: We add *i)* all alignment points which are neighbours to an alignment in A and occur the union of $A1$ at $A2$ (but not in the intersection) and *ii)* alignments of the union for all tokens which are not yet aligned.

In Koehn et al. (2005), the approach has been given the name 'Grow-Diag-Final' algorithm. Listing 1 (taken from Koehn (2010, p. 118)) illustrates the algorithm in pseudocode. We implemented the algorithm exactly in that way and applied it to the whole corpus without any restrictions on minimal or maximal sentence

```
GROW-DIAG-FINAL(e2f,f2e):
  neighboring = ((-1,0),(0,-1),(1,0),(0,1),
                 (-1,-1),(-1,1),(1,-1),(1,1))
  alignment = intersect(e2f,f2e);
  GROW-DIAG(); FINAL(e2f); FINAL(f2e);

GROW-DIAG():
  iterate until no new points added
    for english word e = 0 ... en
      for foreign word f = 0 ... fn
        if ( e aligned with f )
          for each neighboring point ( e-new, f-new ):
            if ( ( e-new not aligned or f-new not aligned ) and
                 ( e-new, f-new ) in union( e2f, f2e ) )
              add alignment point ( e-new, f-new )

FINAL(a):
  for english word e-new = 0 ... en
    for foreign word f-new = 0 ... fn
      if ( ( e-new not aligned or f-new not aligned ) and
           ( e-new, f-new ) in union ( e2f, f2e ) )
        add alignment point ( e-new, f-new )
```

Listing 1: Pseudocode for the Grow-Diag-Final algorithm, Koehn (2010)

length[25].

The symmetrisation procedure needs about 8 minutes on a Linux server with two Intel Xeon E5520 CPUs à 2.27 GHz and 24 GB RAM. Every dataset object as shown in listing 4 is extended by the resulting merged alignment. Thus, the first sentence of the corpus now looks as in figure 5.

```
<sentID=1>
reinício da sessão
reinício do sessão
NCMS000 CPDAFS0 NCFS000
wiederaufnahme der sitzungsperiode
Wiederaufnahme d Sitzungsperiode
NN ART NN
0: 0   1: 1   2: 2   3: 3
0: 0   1: 1   2: 2   3: 3
0: 0   1: 1   2: 2   3: 3
0.00357598
8.42753E-4
```

Figure 5: Collection of all generated data incl. merged alignment

Now, there are three alignments available: *pt2de*, *de2pt* and *refined* (see terminology in Moirón and Tiedemann (2006)).

[25]This is a contrast to experiments described in the literature, e.g. Koehn et al. (2003), who consider a maximal sentence length of 15 in order to achieve better alignments.

Additionally, we carry out test runs which implement just a part of the Grow-Diag-Final implementation. One test excludes the 'Final' step which would add all words that are not yet aligned after the 'Grow-Diag' step but in the alignments' union. Another one only considers the alignment intersection which results in sparse but precise alignments. Intersection has been used as symmetrisation strategy in Moirón and Tiedemann (2006)[26]. We will refer to the effects of these settings to the SVC acquisition in chapter 3.

2.4 Qualitative evaluation of the alignments

This section describes an evaluation which examines the quality of the alignment and symmetrisation steps. As we plan to switch between Portuguese and German in order to retrieve Portuguese SVCs, it is important to get an impression of how the alignments between Portuguese FVs and their German counterparts look like, and which of these German counterparts lead to Portuguese SVCs. Based on this evaluation, we make decisions for the further proceeding.

We compare the amount and extent of the alignments before and after the alignment symmetrisation, and the quality of the alignments from different points of view.

We evaluate three Portuguese FVs, i.e. *ler* ('to read'), *perguntar* ('to ask') and *apoiar* ('to support/to help'), and their alignments. They were chosen because they are expected to act as a bridge between FVs and synonymous SVCs in Portuguese. Specifically, we aim at the SVCs *dar apoio* ('to provide support', lit. 'to give help'), *fazer uma pergunta* ('to ask a question', lit. 'to do a question') and *fazer uma leitura* ('to read', lit. 'to do a reading'). All of these SVCs also exist in EUROPARL. The German counterparts of these verbs are *lesen*, *fragen* and *unterstützen/helfen*.

There have been extracted 17,943 datasets containing at least one of these FVs. In the following, if not otherwise indicated, we consider the alignments of all occurrences of these three portuguese FVs.

2.4.1 Coverage of the symmetrised alignments

The effect of the alignment symmetrisation described in section 2.3 is, as expected, a more widespread alignment. Example (26) illustrates this effect for the verb *apoiar*, comparing the *pt2de* and the *refined* alignment.

(26) *pt2de* : apoiar – null
refined : apoiar – Beihilfe ('aid')

The alignment symmetrisation obviously improves the unidirectional alignment, i.e. it fills an alignment gap. On the other hand, the symmetrisation partly leads to unnecessarily or incorrectly aligned tokens, as in example (27): the alignment is extended by a token which has nothing to do with the source token.

[26]The reason for the poor performance of the *refined* word alignment in their work might be the intersection sparsity, coupled with the fact that the authors consider only 1:1 alignments.

# aligned German tokens	# occurrences	Cumulative percentage
1	12,167	67.0%
2	3,975	88.9%
3	1,203	95.5%
4	340	97.3%
5	258	98.7%
6	101	99.3%
7	56	99.6%
8	34	99.8%
9	30	99.96%
10	6	100%
	$\sum$ 18,170	

Table 1: Number of German tokens aligned with a Portuguese full verb

(27) *pt2de* : apoiar – Unterstützung ('support')
refined : apoiar – Unionsland Unterstützung ('EU country support')

Nonetheless, filling the null-alignments as in example 26 is very useful. The symmetrisation establishes new links between open-class tokens (especially V-N and V-V, but also V-ADJ) which have been unaligned before. As we are especially interested in open-class tokens for the SVC acquisition, a higher alignment coverage is helpful. Furthermore, it is possible to filter inappropriate alignments in later steps.

For the total of 17,943 extracted datasets, we count 4,114 (22.9%) differences between the *refined* and the *pt2de/de2pt* unidirectional alignments for the three FVs mentioned above. In 1,908 of these cases (10.6%), an alignment is created for a previously unaligned token.

The increase of the alignment which arises from the merge is strongly desired, so that we opt for the *refined* alignment as basis for our proceedings. In the following chapters, any mention of 'alignment' refers to the *refined* (i.e. symmetrised) alignment.

2.4.2 Quality of the symmetrised alignments

Next, we examine the distribution and quality of 1:1, 1:2... 1:n alignments. Table 1 shows the amount of German words with which a Portuguese FV is aligned[27]. Most Portuguese FVs are aligned with only one or at least few German words: $n = \{1..4\}$ make up more than 97% of the alignments. This is linguistically intuitive: for the language pair Portuguese-German, we expect in most of the cases that one word in language$_a$ can be expressed by one or few words in language$_b$.

[27]Note that the total number of observed alignments in table 1 is higher than the amount of examined datasets, as the tokens can occur multiple times per sentence.

For the verbs chosen for the evaluation as well as throughout all alignments produced by GIZA++, it can be observed that sometimes a single word in a sentence of language$_a$ is aligned with many words in the related sentence in language$_b$, as shown in figure 6.[28]

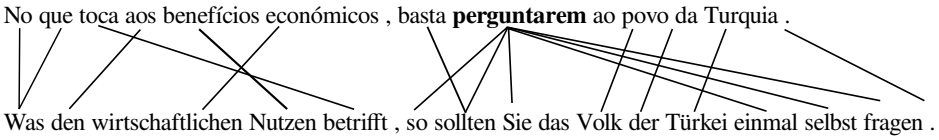

Figure 6: 1:n alignment with high n

In this sentence pair, the Portuguese FV *perguntar* is aligned with six German words. At first sight, such an alignment seems counterintuitive. But indeed, the morphological form of the Portuguese verb – imperative, 2^{nd} person plural – largely contains the contents and nuances of the German counterpart, i.e. the slightly impatient demand (*sollten* and *einmal*), the person (*Sie*), and the FV (*fragen*).

However, in most of the cases, such vast alignments are incorrect or at least not suitable for our objectives. According to Moirón and Tiedemann (2006), this behaviour might be due to non-literal translations, i.e. compositional meanings in at least one of the sentences. This is especially the case for opaque MWEs like idiomatic expressions, but might occur for SVCs as well, located on the border between literal and abstract. For our purposes, these alignments are not suitable as they introduce many misalignments. On the other hand, it is needless to say that we have to keep 1:n alignments in order to retrieve SVCs.

Hence, we decide to cut off and disregard all 1:n alignments with $n \geq 5$ in our study. As for our exemplary set of about 18,000 datasets, we reject 485 occurrences (2%) of the Portuguese FVs. Note that this restriction only refers to the alignment of *one* source word but not to the total number of alignments of an MWE in the source language; every word in an MWE can be aligned with up to 4 words.

2.4.3 Alignments between full verbs and SVCs

As the basic idea is to automatically extract SVCs starting with an FV, it is necessary to get an impression of *i)* how the alignments of SVCs look like, *ii)* how the alignment between an FV and an SVC looks like, and *iii)* if there are enough occurrences of alignments between FVs and SVCs. Therefore, we conduct a manual analysis on some relevant examples. We extracted the alignments of:

- the three Portuguese FVs mentioned above (e.g. *apoiar*)
- their most appropriate German FV counterparts (e.g. *unterstützen*)
- the three expected Portuguese SVCs (e.g. *dar apoio*), and

[28]The English translation of this example in EUROPARL is: 'And as to the economic benefits, **just ask** the turkish people.'

- their expected German SVC counterparts (e.g. *Unterstützung geben*)[29]

In each of these four evaluations, we compare the extracted token(s) with the respective counterpart in the other language. Although the data might contain m:n alignments ($\text{SVC}_a \rightarrow \text{SVC}_b$), we only consider 1:n alignments here, as we always consider the alignment of one specific word (e.g. $\text{VERB}_{SVC_a} \rightarrow \text{SVC}_b$).

We extracted the alignments of 19,425 Portuguese and 22,973 German FVs, and of 1,988 Portuguese and 2,343 German SVCs. In order to get not only exact SVC token sequences but also occurrences with modifiers such as *fazer* ***mais*** *uma pergunta* ('to ask *another* question'), we created patterns with concrete and abstract elements for Portuguese. Such patterns lead to a higher coverage than other pattern-based approaches for SVC retrieval proposed in the literature (e.g. Grefenstette and Teufel (1995) or Cinková et al. (2006)). We extract SVCs on the basis of patterns consisting of tokens and (optional) POS patterns as shown in example (28)[30]. This pattern matches e.g. the expressions *fazer uma pergunta*, *fazer a pergunta*, *fazer a radical pergunta* and *fazer a minha segunda pergunta*. The whole list of patterns can be found in appendix B.

(28) fazer D (PX) (A) pergunta

How do the alignments of SVCs look like? Concerning the question *i)*, the following behaviour is observed: if an SVC_a in language_a is translated into an SVC_b in language_b, then VERB_a is frequently not aligned at all. However, frequent alignments occur between two nouns, like

$$\text{NOUN}_a \rightarrow \text{NOUN}_b$$

where NOUN_b is in most cases the appropriate counterpart, or between a noun and both verb and noun[31]

$$\text{NOUN}_a \rightarrow \text{NOUN}_b{+}\text{VERB}_b.$$

For example, the SVC *fazer uma pergunta* was extracted 773 times, with the SV *fazer* being unaligned at 21.1%, and being aligned to a verb at 63.9%. This counterpart verb is mostly an SV as well. Only in 1.9% of cases, there is an alignment of the following form:

$$\text{VERB}_a \rightarrow \text{NOUN}_b{+}\text{VERB}_b.$$

[29]For the Portuguese SVC *fazer leitura*, there is no exact SVC equivalent in German. Instead, we use the German noun *Lektüre* ('reading') which retains the semantic meaning.

[30]The POS tags follow the first letters of the PAROLE tagset. 'PX' stands for possessive pronouns. See appendix C for the complete list of initial letters.

[31]The word order in language_b is not necessarily NOUN-VERB, as the formula illustrates an alignment.

The numbers of the remaining 13.1% of alignments with other word classes (adjectives etc.) are approximately uniformly distributed and not important for this study. These figures show that the SV is often semantically impoverished, as it is either aligned with an SV or unaligned, and is rarely connected to a noun. In contrast, the noun *pergunta* is *always* aligned, filling three alignment patterns which seem all promising (see table 2 for details). This fact shows that especially the alignments of the SVC's noun is fruitful. Note that this behaviour is characteristic for both German and Portuguese.

Alignment pattern	# occurrences	Percentage
pergunta $\rightarrow$ NOUN$_b$	568	77.5%
pergunta $\rightarrow$ NOUN$_b$+VERB$_b$	104	14.2%
pergunta $\rightarrow$ VERB$_b$	61	8.3%
$\sum$	733	100%

Table 2: Alignments of the noun *pergunta*

How does the alignment between an FV and an SVC look like? As to question *ii)*, we draw the following conclusions: if SVC$_a$ is translated into a full verb FV$_b$, we observe in most cases alignments with the SVC's noun (e.g. *pergunta* 58 times with *fragen*):

$$\text{NOUN}_a \rightarrow \text{FV}_b$$

In contrast, the SVC's verb is rarely aligned with the FV but rather remains unaligned (e.g. *fazer* only 19 times with *fragen*):

$$\text{VERB}_a \rightarrow \text{FV}_b$$

Considering the alignment of the FVs, they are in most cases translated directly and hence aligned with corresponding FVs. However, especially for Portuguese FVs, there are also many 1:n alignments. Table 3 indicates the most frequent alignment patterns and their frequencies for the FV *perguntar*.

Alignment pattern	# occurrences	Percentage
perguntar $\rightarrow$ VERB$_b$	975	81.25%
perguntar $\rightarrow$ NOUN$_b$	225	18.75%
$\sum$	1200	100%

Table 3: Alignments of the FV *perguntar*

As to the 975 alignments with a verb in German, most of the cases link *perguntar* with the appropriate FV. The 225 aligned German nouns are frequently

the nominalised form of the FV (i.e. *Frage*, which means 'question') and thus the semantically correct counterpart. As the nominal predicate in an SVC also contains the semantic core, we expect that a heuristic extension of the alignments between FVs and nouns could quantitatively improve the retrieval of SVCs. We cover this point in section 3.2.2.

Concerning the 1:n alignments, we discover in about 30% of the cases an aligment between a verb and a noun-verb combination, which obviously also matches SVCs:

$$\text{VERB}_a \rightarrow \text{NOUN}_b\text{+VERB}_b$$

Considering the remaining 1:n alignments, many cases will be rejected, as $n \geq 5$ (see section 2.4.2). Moreover, a lot of 1:n alignments for Portuguese FVs are due to the fact that, in Portuguese, the information about the person can be incorporated into the verb's conjugation (see Gärtner (1998, p. 118)), whereas German as well as English need a personal pronoun. These pronouns are co-aligned with the verb. Figure 7 shows such a case of 1:n alignment between Portuguese *Apoiemos* and German *Unterstützen wir*[32].

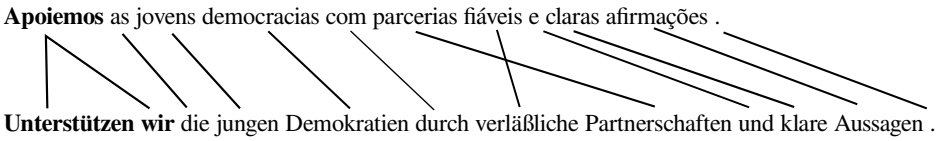

Figure 7: 1:n alignment for a co-aligned pronoun in Portuguese

These results show that the acquisition of SVCs, starting from an FV, is reasonable and promising, even though some effort additionally to the automatic alignment is necessary.

Are there enough occurrences of alignments between FVs and SVCs? The observations in this section suggest that question *iii)* can also be affirmed: if there is a proper SVC equivalent for an FV, there are enough alignments which help to reveal this equivalence. However, if there is no appropriate equivalent, the alignments are erroneous. As mentioned above, this is the case for the SVC *fazer leitura*. Therefore, it is not well utilisable.

Additional remark. We revealed one interesting side-fact in this manual study: the German FVs are often aligned with Portuguese constructions consisting of a copula verb and an adjective like *ser favorável* ('to be in favour') in example (29). As discussed in Döll and Hundt (2002), such structures operate as synonymous substitutions of SVCs. A systematic retrieval of such structures for semantic

[32]The English translation of this example in EUROPARL is: '**Let us support** the young democracies by means of reliable partnership and clear messages.'

categorisation might be interesting as well. However, we do not investigate this idea.

(29) *Ainda assim,* ***sou favorável*** *a grande parte do relatório que estamos a votar.*
Große Teile des zur Abstimmung vorliegenden Berichts kann ich dennoch **unterstützen** [...]
I **am**, nevertheless, **in favour** of large parts of the report.

3 Step one: SVC detection with the pivot approach

After having examined the alignment characteristics of SVCs and FVs which are synonymous with SVCs, we will now describe the implementation of the extraction. It is the first step of our two-stage approach for SVC processing. We use the concept of Bannard and Callison-Burch (2005) to retrieve SVCs starting from a FV list. This chapter explains briefly the basic idea of their article, its modified application in the present thesis and the concrete implementation.

3.1 Origin and adaptation of the approach

Bannard and Callison-Burch (2005) exploit the parallel data in EUROPARL to gather English paraphrases, like 'in check' for the initial phrase 'under control'. This article is one of the main inspirations of the present thesis. The alignment is based on the phrase-based machine translation approach by Och and Ney (2003). First, the 'pivot' algorithm locates all phrases f in the foreign language which are aligned with the initial phrase e_1; this part is henceforth called the first pivot step. Then, the algorithm locates all occurrences of these foreign phrases in the 'pivot language' (in our studies: German), goes back to the English data and gathers the English phrases e_2 which are aligned with the occurrences in the foreign language (this is called the second pivot step; both steps together are the pivot pipeline). According to the assumption that all extracted English phrases e_2 which have been retrieved in this manner have a similar meaning, they are considered as candidate paraphrases for e_1. Obviously, the approach allows for the extraction of various paraphrases.

To each of these candidate paraphrases, a probability value is assigned. Then, the single best paraphrase with the highest probability $\hat{e}$ is selected. Its computation is based on equation (1), but the authors have also implemented several modifications and extensions, such as the consideration of a language model for contextual probability, the use of multiple parallel corpora (i.e. several foreign languages) or word sense disambiguation.

$$\hat{e_2} = arg \max_{p(e_2|e_1)} = arg \max_{e2 \neq e1} \sum_f p(f|e_1)p(e_2|f) \tag{1}$$

Their approach achieves good results which are measured with regard to both meaning and grammaticality. The evaluation is carried out by means of two five point scales for the evaluation of machine translation systems, established by the Linguistic Data Consortium (LDC, 2002); see Callison-Burch (2007). The best model implements word sense disambiguation and achieves 70.4% of correct meaning for the extracted paraphrases.

We exploit EUROPARL's parallel data similar to Bannard and Callison-Burch

(2005), however, with a different source and target language (i.e. Portuguese and German) and a limitation to specific paraphrases, namely SVCs. We expect to end up with a list of extracted SVCs after the second pivot step. Hence, we also assume that the pivot approach leads to semantically similar constructions.

The same steps as in Bannard and Callison-Burch (2005)'s process have been implemented, but there are slight differences regarding some parameters.

First of all, we do not calculate probabilities, but simply use thresholds which indicate how many times an alignment pair must occur to be considered. Four different thresholds are determined. They are explained in detail in section 3.2.1.

Compared to Bannard and Callison-Burch (2005), the scope of our objective is much more narrow: while they locate paraphrases of any structure for phrases consisting of at least two tokens, we focus on one-word expressions (OWEs) as input (i.e. the FVs) and specific more-word expressions – SVCs – as output. These specific constructions require some heuristic constraints and filtering concerning the alignment tokens which are considered. They are illustrated in section 3.2.2.

Additionally, we experimented with different types of alignment models, i.e. partial implementations of the Grow-Diag-Final algorithm as mentioned in section 2.3. They are once again explained briefly in section 3.2.3.

3.2 Adjustable parameters

This section explains the adaptations we have made to Bannard and Callison-Burch's approach. The implementation of the pivot pipeline (as well as the subsequent filtering steps described in section 5.2) was done in Java.

3.2.1 Minimum occurrence thresholds

We work with four thresholds: two each for the first and the second pivot step, that is, for the alignment from the initial FV to the corresponding tokens in German, and – after having retrieved all occurrences of the selection of these German tokens – for the alignment of the German sequences and their Portuguese counterparts.

Both pivot steps can lead to OWEs – usually individual verbs or nouns – as well as MWEs – the candidate SVCs. As OWEs are naturally more frequent than MWEs, it is important to define higher thresholds for OWEs than for MWEs to achieve reasonable results. This leads to two thresholds for each pivot step. More precisely, as the objective is to retrieve SVCs, OWEs are completely excluded from the final output list. Consequently, the OWE threshold in the second pivot step is rather a filter, such designed that no OWE expression is selected at all.

We test different combinations of these thresholds with the initial Portuguese verbs announced in section 2.4. The experiments show that a higher threshold for OWEs in the first pivot step indeed improves the results as it excludes several high-frequent but semantically unspecific tokens – in most cases individual auxiliaries

and support verbs like 'to have' or 'to make'. The tested thresholds range between 300 and 350. However, a reasonable threshold can vary for different input verbs.

At this point, we encounter a central problem of automatic SVC processing: despite the preselection of OWEs via thresholds, high-frequent tokens like *ter* ('to have') cause problems in the first pivot step. They result in a huge amount of occurrences like *ter* + NOUN which, in large part, do not improve the SVC extraction. This effect is due to the high frequency of *ter* and thus the high probability that it is part of an alignment. We decided to explicitly exclude all results of the first pivot step which contain a token with a frequency of more than 100,000 in the corpus. This corresponds to an exclusion of tokens which make up about 0.4% of the corpus in each language.

Concerning the thresholds for MWEs, the values are considerably lower. The first pivot step should be less restrictive as it is followed by a second filtering step, and results in a threshold of 6. Even if this value seems low, it is enough to filter out most of the undesirable MWEs. For example, although being a probable word sequence in EUROPARL, the unlexicalised and compositionally free construction in example (30) is excluded. In the second pivot step, a slightly higher value of 9 is selected.

(30) Fraktion unterstützen
support the group

Obviously, the thresholds affect the system's precision and recall. They should be chosen with respect to the further intents and filtering methods. For example, it is well justifiable to choose high thresholds in order to improve precision, or to opt for low thresholds if further steps can filter out the false positives from the high recall.

The first column in table 4 on page 36 shows some results for the Portuguese input verb *apoiar* using the presented heuristic (thresholds: 300 (OWE) and 6 (MWE) in the first and 9 (MWE) in the second pivot step on symmetrised alignments). The SVCs which are regarded as semantically appropriate to the FV *apoiar* are highlighted in boldface[33].

3.2.2 Linguistic heuristics

Bannard and Callison-Burch (2005) typically handle sequential phrases which are not interrupted by other tokens. However, SVCs – in Portuguese (Athayde, 2001) as well as in other languages (e.g. see Fellbaum et al. (2006), Eisenberg (2006), Cinková et al. (2006)) – can be modified by adjectives, adverbs, possessive pronouns, negations etc., standing between the actual SVC tokens. This is not valid for all SVCs; some of them prohibit such modifications, as example (31) shows (taken from Döll and Hundt (2002, p. 154)).

[33]The decision on what is semantically appropriate is derived from an evaluation we undertook at this point. It is presented in detail in chapter 4.

(31) estar na dúvida – * estar na grande dúvida
'to be in doubt' – * 'to be in big doubt'

In our parallel data, such modifications are frequently co-aligned with the SVCs. Using the complete alignment would lead to data sparseness and worsen the results. Therefore, we focus on tokens which are expected to be part of the candidate SVC, i.e. nouns and verbs. Recall from section 1.4 that we ignore prepositions, because we concentrate on non-prepositional SVCs.

We focus on nouns and verbs insofar as we select only aligned tokens of these POS types; all other POS tokens are discarded. If more than one verb or noun are co-aligned, only the first hit is kept. For Portuguese, another restriction is effected: we exclude occurrences of the verb *ser* (one of the two Portuguese verbs for 'to be') as part of SVCs, as it never occurs in SVCs presented in the literature. To the best of our knowledge, there is no complete list of Portuguese SVCs available. However, the fact that neither Athayde (2001), nor Döll and Hundt (2002) or Gärtner (1998) specify any SVC examples with *ser* as SV, justifies this assumption.

For 1:n alignments starting from a German token (i.e. in the second pivot step), we only keep the nouns as aligned Portuguese tokens. Several experiments have shown that the extraction of Portuguese verbs from 1:n alignments do neither improve nor worsen the overall results. We think that this is due to a fact presented in section 2.4.3: the alignment strategy of GIZA++ and the Grow-Diag-Final symmetrisation algorithm frequently align the nouns which hold the major part of an SVC's semantic meaning, while SVs remain unaligned. In contrast, FVs are rarely co-aligned with other tokens and hence occur hardly in 1:n alignments, but rather in 1:1 alignments.

This leaves us with alignment selections of the following basic structures:

$X_a \rightarrow \text{NOUN}_b + \text{VERB}_b$
$X_a \rightarrow \text{VERB}_b$
$X_a \rightarrow \text{NOUN}_b$

Moreover, we pay attention to the fact that many SVs remain unaligned: we try to expand 1:1 noun alignments into a nearby verb that potentially is the appropriate SV. Carrying out similar expansions for 1:1 verb alignments is neither necessary nor fruitful, as single-aligned verbs tend to be FVs. An expansion would lead to overgenerating erroneous alignments, i.e. unlexicalised and compositionally free expressions.

The expansion step is implemented in different ways for alignments starting from a Portuguese or a German token, allowing for the language-specific syntactic structure in which SVCs may occur. Thus, the expansion can be carried out *i)* in both pivot steps and *ii)* for both languages as starting point. The procedures are as follows:

If a German single token is 1:1-aligned with a Portuguese noun, step backwards

up to 6 tokens which *precede* the noun and look for alignable verbs. Stop after the first discovery (i.e. the nearest verb to the noun). This expansion heuristic is safely applicable on Portuguese data, as its sentence structure is rather strict. As to Portuguese single tokens aligned with a German noun, we conduct a more careful extension because German syntax allows a more flexible sentence structure: the search for an appropriate verb is restricted up to a distance of 3 tokens. It is done forwardly, searching for a verb which *follows* the noun. For both languages, the search for a nearby verb is stopped as soon as a preposition or a sentence boundary is reached: in most cases, prepositions introduce a phrase which cannot be inserted into an SVC and, naturally, SVCs cannot be split across sentences. The distance of 3 tokens in German and 6 tokens in Portuguese, respectively, have been empirically identified as reasonable margins.

The heuristics clearly increase the recall of the overall pivot pipeline. For instance, there is not a single hit for the thresholds 350 (OWEs) and 9 (MWEs) in the first pivot step and 20 (MWEs) in the second pivot step if one leaves out these steps. The main reason for this effect is the exclusion of OWEs from the final results. Although there occur wrong verb extensions like * *dizer apoio* (* 'to say support'), the overall effect of the heuristic is positive.

We also run an expansion test with a reduced set of expansion candidates: only the addition of previously unaligned tokens is allowed. We particularly expect to avoid the addition of locationally near but semantically unrelated verbs which are already aligned with their appropriate (non-SVC) counterpart. However, this configuration worsens the SVC detection results, especially the recall. This means that in most cases, the support verbs of the SVCs are aligned, even if the alignment should be incorrect. We attribute this to an effect arising from the Grow-Diag-Final symmetrisation algorithm: as the unaligned neighbours of aligned tokens in the intersection get aligned as well, there are hardly any items left unaligned which are nearby the noun to be expanded. Due to the performance decline, this idea was rejected.

Note again that the decision on which heuristic is chosen and how it is adjusted, depends on the matter of interest, the language pair and the preference concerning precision and recall.

3.2.3 Different alignment algorithms

The Grow-Diag-Final algorithm comprises three possibilities to vary the symmetrisation method. These are:

- The *Complete* Grow-Diag-Final algorithm. Up to this point, we always used this setting.
- *Grow-Diag only.* It intersects the unidirecional alignments and adds all neighbouring alignment points of the union, but excludes the remaining (non-neighbouring) alignment points of the union.

- *Intersection only.*

We expect that the precision and recall values of the results of the pivot approach reflect the application of different symmetrisation strategies: *Grow-Diag* and *Intersection only* contain sparser alignments because they have more null-alignments, 1:1 alignments, and their 1:n alignments contain a lower amount of n aligned tokens than in the *Complete* symmetrisation procedure. Especially, high-frequent verbs as 'to have' are often 1:1-aligned with FVs but are then excluded due to the heuristics explained in section 3.2.1. Thus, a lot of SVCs should remain unrecognised as there are less occurrences of appropriate alignments (i.e. alignments containing a noun and/or a verb).

In fact, as assumed, *Grow-Diag only* returns less SVCs than the *Complete* symmetrisation; the result is a subset of *Complete*. However, the difference is smaller than expected: for the FV *apoiar*, *Complete* has only three more entries than *Grow-Diag*, and for the word *perguntar*, the results are even exactly the same. As to *Intersection only*, there is a surprise: not in all cases, it returns less results than *Complete*. Instead, it leaves out some of *Complete*'s SVCs and retrieves, in return, other candidate SVCs. We estimate that the heuristics described above lead to this effect, as they improve the SVC retrieval considerably. The unexpectedly high recall might also be due to the exclusion of 1:n alignments for $n > 4$ in all symmetrisation settings. This restriction levels the results of the different alignment methods.

Consider table 4 again. It shows an excerpt[34] of the SVCs retrieved by the different strategies for the word *apoiar*, using the thresholds and heuristics as defined in the previous sections.

Complete	Grow-Diag only	Intersection only
merecer apoio	merecer apoio	merecer apoio
receber apoio	receber apoio	receber apoio
prever apoio	-	-
ter apoio	ter apoio	ter apoio
conquistar apoio	conquistar apoio	conquistar apoio
apoiar proposta	-	-
dar ajuda	**dar ajuda**	**dar ajuda**
exigir apoio	exigir apoio	exigir apoio
reunir apoio	reunir apoio	reunir apoio
ser promoção	-	-
providenciar apoio	providenciar apoio	providenciar apoio
-	-	tornar apoio
dar assistência	**dar assistência**	**dar assistência**
prestar ajuda	**prestar ajuda**	**prestar ajuda**

Table 4: Some results of different alignment symmetrisations for *apoiar*

[34]The total of candidate SVCs retrieved comprises 65, 66 and 68 entries for the *Grow-Diag*, *Intersection* and *Complete* strategy, respectively. See appendix D for the whole list for *Complete* symmetrisation.

It is surprising that the partial symmetrisation strategies lead to more accurate results (returning a lower amount of false positives but the same amount of true positives). Thus, we conclude that – at least for relatively specific problems – the aligment quality is not necessarily a bottleneck if one has reasonable thresholds and/or heuristics which help to straighten out alignment mistakes. Callison-Burch (2007) and Zarrieß and Kuhn (2009) also draw the conclusion that sparse or partly incorrect alignments can be overcome, whereas Moirón and Tiedemann (2006)'s problem seems to be precisely the alignment quality.

3.3 Final pivot setting and additional remarks

The preceding investigations show that it is worth exhaustively analysing the effects and mutual influence of the specified parameters. It turns out that there are various settings for the pivot pipeline which all make sense, depending on the overall objectives, (i.e. high precision or high recall). Especially the heuristic filters can become very fine-grained if one needs to adjust the system to a certain language and phenomenon.

After having carried out all these tests, we opt for a setting which retains a high recall so that we do not lose potentially correct SVCs, but can apply further (monolingual) filter methods in the following steps to get a higher precision. Consequently, the *standard parameters*, as we call them henceforth, are set as follows:

- Thresholds: 300 for OWEs and 6 for MWEs in the first pivot step; 9 for MWEs in the second pivot step
- Heuristics: extraction of only verbs and nouns, expansion of 1:1 alignments of nouns into nearby verbs
- Alignment symmetrisation strategy: complete Grow-Diag-Final algorithm

During the tests, one of the initial Portuguese verbs – *ler* – turned out to be unusable, although this verb and its most frequent German alignment counterpart *lesen* occur frequently enough in EUROPARL to be used for the pivot technique. *Ler* occurs about 2,100 times and *lesen* about 1,000 times, respectively, with an intersection – i.e. alignment between *ler* and *lesen* – of exactly 1,000. As well, the most expected SVC – *fazer leitura* – occurs in the corpus (about 50 times). However, this amount seems to be too small: all but seven of these occurrences are *not* translated into the German verb *lesen*, which is the only expression retrieved by the first pivot step for the initial FV *ler* using the standard parameters. This might be due to the fact mentioned in section 2.4.3, that there is no German SVC expressing exactly the meaning of *ler*. Trials with very low thresholds (10 for OWEs and 1 for MWEs in the first pivot step; 1 for MWEs in the second pivot step) of course reveal the corresponding SVC, but the overall results perform poorly: out of 39 expressions extracted for *ler*, *fazer leitura* is the only valid SVC.

Thus, the following tests are carried out only with the initial FVs *apoiar* and *perguntar*, in parts expanded by other FVs which are more fruitful for our purposes than *ler*. If so, the additional FVs are indicated.

Obviously, the pivot pipeline approach presented in this chapter is quite promising for SVC extraction: we are able to acquire syntactically valid SVCs which are semantically equivalent to the given FV. As is typical for data-driven approaches, we detect unexpected but correct SVCs which can hardly be found in a monolingual manner or even be made up by humans – e.g. Lin and Pantel (2001) report similar findings for paraphrase extraction. Nevertheless, there are still many false positives: the pivot approach tends to extract not only semantically equivalent SVCs but also their antonyms. For example, the expression *exigir apoio* ('to demand help') in table 4 has the opposite meaning of *apoiar*. Thus, we filter the results in a second step to eliminate the false positives.

Chapter 5 explains the ideas we pursue to achieve such a refinement, and their feasibility. But before, it is necessary to evaluate the results we have achieved so far to have a well-founded basis for a final evaluation. The next chapter describes the evaluation we established on the output of the pivot procedure, and its intermediate results.

4 Gold standard and intermediate results

This chapter describes the creation of a gold standard for the SVCs retrieved with the pivot pipeline for a list of FVs, and indicates the performance achieved so far.

4.1 Evaluation data and setting

We carried out our evaluation on the output of the pivot pipeline, ran with the *standard parameters* defined in section 3.3. The six following FVs have been chosen: *ameaçar* ('to threaten'), *apoiar* ('to support/help'), *faltar* ('to lack'), *perguntar* ('to ask'), *prometer* ('to promise') and *responder* ('to answer'). They have approximately the same meaning as at least one Portuguese SVC. Table 5 indicates how many different candidate SVCs have been acquired by the pivot pipeline for each of these FVs.

Initial FV	# candidate SVCs retrieved
ameaçar	1
apoiar	64
faltar	2
perguntar	7
prometer	3
responder	7
	$\sum$ 84

Table 5: Number of candidate SVCs per full verb found by the pivot pipeline

Two Portuguese native-speakers, one from Portugal and another one from Brazil, carried out the annotation. They were given annotation guidelines (see appendix E) to ensure that their annotations are as consistent as possible. The annotators also got six files, one each per FV, containing the acquired candidate SVCs and sample sentences for each of these SVCs. For illustration, appendix F shows the file for the verb *prometer*.

The annotators evaluated two aspects:

1. Is the expression an SVC or not?
2. If yes: is the SVC semantically substitutable with the respective FV?

One could derive the quality of an SVC from the number of times it can be replaced by the respective FV in the sample sentences. Instead, we asked the evaluators to additionally create a qualitative ranking for every SVC list, i.e., to determine which SVC replaces best the meaning of the respective FV. However, we do not use the rankings since there is hardly any inter-annotator overlap. Details are provided in section 4.2.

4.2 Evaluation quality and gold standard

Calculating quality. We calculate the inter-annotator agreement to measure the annotation's reliability. Therefore, we use Cohen's Kappa (Cohen, 1960), also referred to as κ, as in equation (2):

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{2}$$

where p_0 is the measured agreement and p_e is the probability of agreement by chance. This measure is especially suitable in our setting, as we have two annotators and two evaluation tasks (one for each question illustrated in the previous section) with a two-valued nominal scale each (yes/no).

Thus, we calculate two κ values: one for the evaluation if an expression *is* an SVC, and another one for the evaluation if an SVC is *semantically equivalent* to the FV. Table 6 indicates the results.

Evaluated feature	κ value
Expression *is* an SVC	.604
SVC is *semantically equivalent* to the FV	.744

Table 6: κ inter-annotator agreement for the evaluation of candidate SVCs

It is not surprising that the first κ value is lower, as the decision if an expression is an SVC or not is a more general and thus a more difficult task than the second one. However, these are fairly good agreement rates, regarding the facts that SVC determination is a difficult task because of fuzzy borders for SVs (see Grefenstette and Teufel (1995)), and that the annotators come from different regions what might cause different use of SVCs.

Generally, the reliability of inter-annotator agreements is controversial, depending on the field of research: an agreement rate can be classified as fairly reliable by one evaluation standard, while another standard proposes to refuse the results. We base our interpretation on the criteria of Landis and Koch (1977) but compare our results as well with the classification of Krippendorff (1980). According to Landis and Koch's classification, the second κ value obtained here is 'substantial' ($0.61 < \kappa < 0.80$) and, thus, in the second-best of six categories; the first κ value is just slightly below this margin and classified as 'moderate'. Krippendorff's more rigorous scale would definitely reject our inter-annotation agreement for the first task, while the second ($\kappa > .67$) would allow tentative conclusions.

Establishing the gold standard. For the composition of the gold standard, we use the manually annotated data in the following way. First, we take the intersection of the two annotations: all expressions which have been judged as true positives (true negatives) by both annotators are classified as true positives

(true negatives) in the gold standard. For all cases in which the evaluators did not concur, we classify the expression by ourselves[35]. The main criterion for the decision is, whether the verb can be interpreted as SV in the given expression. As SV, the verb loses its general meaning – or a part of it – and is getting semantically impoverished. If this is the case, we consider the expression as true positive, or more precisely, as SVC.

The 22 SVCs finally judged as true positives in our gold standard are shown in figure 8. Note that the SVC *dar apoio* is represented three times, namely as *dar apoio*, *dar-lhe apoio* and *dar+lhe apoio*. In the latter two cases, the indirect object *lhe* is directly attached to the verb *dar*, which is due to the tokenisation issues mentioned in section 2.2.3. The concatenation of *dar* and *lhe* is carried out by the FreeLing tagger in two ways, i.e. with + and -, but there are no apparent specific rules which determine the difference between these symbols. The same applies for *dar resposta* and *dar+lhe resposta*.

ameaçar: constituir ameaça;
apoiar: conceder apoio, dar apoio, dar assistência, dar-lhe apoio, prestar ajuda, conceder ajuda, prestar apoio, dar ajuda, dar+lhe apoio, prestar assistência;
faltar: haver falta, ter falta;
perguntar: apresentar pergunta, levantar questão, fazer pergunta, colocar pergunta, colocar questão;
prometer: fazer promessa;
responder: dar resposta, tomar posição, dar+lhe resposta;

Figure 8: True positive SVCs in the gold standard

Examining the manual ranking. As mentioned in section 4.1, we asked the annotators to establish a ranking, judging the replaceability of the FV by the related SVCs. Our goal was to unify these rankings to a 'gold' ranking and use it to evaluate the ranked SVCs returned by our system at the very end. For this purpose, we planned to compute a ranking-order correlation coefficient, e.g. Spearman's ρ.

Unfortunately, it turns out to be difficult to unify the manual rankings: for three of the six initial FVs, the annotators selected a different set of gold SVCs, which they, moreover, ranked differently. For example, only one annotator considers *tomar posição* an SVC which semantically corresponds to *responder*, and while one annotator places *dar ajuda* on top of the list for *apoiar*, the other annotator assigns the third position to this SVC.

For the SVCs extracted for *ameaçar*, *faltar* and *prometer*, the selection and ranking of the annotators is consistent. However, *ameaçar* and *prometer* result

[35]Special thanks to Christine Hundt for her helpful opinion in this task.

in merely one SVC each (see figure 8) and are thus uninteresting for ranking correlation calculations. Similarly, the ranking for *faltar* with two entries is trivial.

Thus, we refrain from establishing a gold ranking. However, the manual rankings reveal an interesting fact: in all rankings which have more than one entry (i.e. for all initial FVs except for *ameaçar* and *prometer*), both annotators judge several SVCs to be equally good replacements for the FV. As an example, table 7 shows the manual ranking for the SVCs detected for *apoiar*.

Rank	Annotator 1	Annotator 2
1	conceder apoio, dar apoio, prestar apoio, conceder ajuda, dar ajuda, prestar ajuda	dar apoio
2	dar assistência, prestar assistência	conceder apoio
3		prestar apoio, prestar ajuda, dar ajuda, dar assistência, prestar assistência

Table 7: Qualitative manual ranking for SVCs replacing the FV *apoiar*

This uniform classification of the competing SVCs suggests that the fine-grained semantic difference of the SVCs is not relevant for the language speakers.

4.3 Intermediate results

This section provides an explanation of the measures calculated in the pivot pipeline evaluation, and their results.

Computed measures. Based on the gold standard shown in figure 8, we can from now on evaluate the performance of our SVC acquisition both qualitatively and quantitatively. For the following explanations, see Manning and Schütze (1999, p. 268 f.). Quality is expressed by the *precision* value (equation (3)), showing the ratio of correct versus incorrect items retrieved by the system. Quantity is measured by the *recall* value (equation (4)), measuring the amount of correct items retrieved by the system versus all items in the gold standard. The f_1 measure (equation (5)) combines precision (p) and recall (r), computed here with an equal weighting of precision and recall ($\alpha = 0.5$):

$$p = \frac{truePositives}{truePositives + falsePositives} \tag{3}$$

$$r = \frac{truePositives}{truePositives + falseNegatives} \tag{4}$$

$$f_1 = \frac{2pr}{p+r} \tag{5}$$

As noted in section 1.2, it is not possible to say how many SVCs exist and thus, how many SVCs make up the total of 'gold SVCs' for a given FV. The total amount of true positives, however, is necessary to compute recall. One can remedy this defect by defining the list of candidate SVCs resulting from the pivot pipeline to be the 'universe', thus, the relative 100% recall basis of all further steps: for us, only the true positives found in that universe exist; no other true positives which could be (in the real world) located outside the universe, are retrievable.

The experimental set-up in the following chapter contains settings which return not only a part of the universe (i.e. the items judged to be true positives), but all expressions retrieved by the pivot step as a ranked list. For these settings, we additionally compute the *(uninterpolated) average precision*, which evaluates the ranking quality of the returned results (Manning and Schütze, 1999, p. 535 f.).

Results for the pivot pipeline. Section 3.3 concludes in a rather intuitive way that the pivot approach is promising for the extraction of SVCs. Now, we concretely evaluate the pivot pipeline's performance in order to verify *i)* how well the pivot algorithm works for the current problem and *ii)* to what extent the following steps lead to improvements.

Table 8 shows the precision and f_1 values for the whole gold-annotated dataset ('all') and the individual FVs. As there has not been established a ranking so far, we cannot compute average precision. According to the above definition, recall is always 100%.

Setting	Results	
	Precision	F_1
ameaçar	1.	1.
apoiar	.16	.27
faltar	1.	1.
perguntar	.71	.83
prometer	.33	.5
responder	.43	.6
all	.26	.42
all but *apoiar*	.6	.75

Table 8: Precision and f_1 for the pivot pipeline

The overall precision is at .26, which is rather low. However, it is easy to see that the result is heavily influenced by the high amount of returned candidate SVCs for *apoiar*, with a high false positives rate. Leaving out the data of *apoiar*, precision increases to .60. Nontheless, these results are not yet satisfying. Thus, the main objective for the following steps is to detect and reject as many false positives as possible.

5 Step two: filtering the SVCs

As explained in chapter 3, we carried out the pivot procedure with rather loose restrictions in order to achieve a higher recall, and to eliminate false positives in further steps. Now, we try to increase precision as well. There are several possibilities how a filtering could look like, all of them monolingual strategies.

One approach exploits contextual information of the SVCs, i.e. the SVC's neighbouring words. For example, defining restrictions for the neighbours by means of recurrent POS patterns should narrow down the *semantic* context of the SVCs and filter the candidate expressions returned by the pivot pipeline.

Another idea is to reduce the result set by computing different association measures for the candidate SVCs, i.e. for the SVC's verb and noun, and to separate the true positives from the false positives in this way.

The following section describes the analyses we carry out to realise the context-related idea. Section 5.2 explains the steps we undertake regarding the association measure idea.

5.1 Filtering using the SVC context

The aim is to separate actual from apparent SVCs by analysing the contexts, i.e. the verb arguments of the SV. The arguments are compared to the arguments of the corresponding initial FV by means of cooccurrence frequencies: If the behaviour of the FV's and the SV's arguments are similar, we assume that FV and SVC are semantically similar as well. More precisely, if the realisation of an argument position is frequently the same for both FV and SV (e.g. the usage of the same preposition or the same substantive), then FV and SVC are supposed to have the same meaning. In order acquire the SV's arguments, we try to detect syntactic patterns which typically surround an SVC. A positive side effect of this approach is the possibility to generate not only a list of SVCs, but also their syntactic participants. From a lexicographical point of view, such an enrichment is desirable.

Danlos (1992) carries out a comprehensive theoretical study about the arguments of SVCs and shows how versatile they can be realised. Similarly, the usage of patterns for the extraction of subcategorised arguments in general received attention in the literature. For example, Haselbach (2010) uses patterns derived from word order typologies – the so-called 'Topologische Feldermodell' (lit. 'topological field model') – to extract verbal subcategorisations in Afrikaans. Gildea and Palmer (2002) compare the effect of different syntactical information on the quality of argument, or rather, semantic role extraction. They show that chunk information lead to good results, but that parses work much better. Punyakanok et al. (2005) do further investigations and confirm Gildea and Palmer (2002)'s hypothesis that full parse information are especially helpful for the extraction task, as the knowledge about argument boundaries has a major influence on the

extraction quality.

We are also interested in revealing the SV's arguments, i.e. the subject and (in-)direct object on the syntactic level, and the θ-roles on the semantic level. Despite the findings in the literature about versatile contexts of SVCs and the improvement of argument extraction by means of full syntactic parses, we propose that these arguments can be extracted by means of POS patterns, looking for noun phrases (NPs) on specific positions with characteristics that are the same as for the semantically equivalent FV.

This section explains the initial intuitions we had about the characteristics of the SVC context, and the actual observations we made. Unfortunately, the analysis shows that the situation is not as easy as we expected so that we only draw theoretical conclusions but refrain from an implementation.

5.1.1 Expectations

Our intuition is that there are typical patterns for SVs within an SVC, and that there are parallels between the SVC and FV text environments. This is also what Grefenstette and Teufel (1995) detected: the arguments of full verbs are frequently overlapping with the arguments of the extracted pairs of SV and nominalised verb. Above all, we expect to often retrieve the direct or indirect object of the FV as an indirect object of the SVC's support verb, and that the subject remains the same. Recalling the fact that we only consider non-prepositional SVCs, the SVC's substantive ought to be the SV's direct object. As to the θ-roles, we expect that in most cases, the same roles are realised, but that *i)* there might be differences in the existence and non-existence of θ-roles in adjunctive prepositional phrases, and that *ii)* the syntactic position of their fillers switches if there is a change in focus. Consider the examples (32), (33) and (34), showing sentences with the FV *apoiar* and two SVCs derived from this FV: *dar apoio* ('to give support') and *pedir apoio* ('to ask for support').

(32) *O grupo socialista apoia uma política industrial .*
The group socialist supports a politics industrial .
The socialist group supports an industrial policy.

(33) *A comissão dará o seu apoio a essa iniciativa .*
The commission will give its support to this initiative .
The commission will support this initiative.

(34) [. . .] *peço o apoio desta câmara para as alterações*
[. . .] I ask the support of this house for the amendments
aprovadas [. . .]
approved [. . .]
[. . .] I call upon this house to support the amendments which have been approved [. . .]

The meaning of the SVC in example (34) is just the opposite of the two preceding examples: the SVC's subject asks for support instead of giving it. Thus, the θ-roles for the act of supporting are switched. In sentence (33), the subject incorporates the agent and the indirect object incorporates the beneficiary. In contrast, the subject in example (34) only remains the agent of the act of asking, but is the (potential) beneficiary of the act of supporting. The indirect object is now the agent; the switch of the indirect object's θ-role is also marked by another preposition (*a* versus *para*). As we intend to retrieve SVCs which semantically correspond to a given FV, SVCs as in example (34) are desired to be discarded and *ii)* servers as criterion for exclusion.

5.1.2 Analysis of sample sentences

Analysis setting. For all 22 gold standard-SVCs found for the six input FVs[36], we check up to eight sample sentences, that is, 22×8 sentences in a rather superficial way. This broad but shallow analysis assures that our assumptions are general enough to hold for many syntactic and semantic contexts.

Furthermore, we evaluate two SVCs, namely *dar apoio* and *fazer pergunta*, in detail across 50 randomly chosen sentences each. We also investigate the internal structure of the NPs which occupy the argument positions. The calculations and counts we report below are based on these 100 sentences extracted from *dar apoio* and *fazer pergunta*; the superficial observations on the remaining sentences are not considered.

Finally, we examine eight sample sentences for each FV, i.e. 6×8 sentences. A comparison of the argument structure of the FVs and the SVCs should enable us to justify – or refute – certain assumptions.

General remarks. In the following, we present the recurrent patterns within, preceding and following an SVC, focusing on the SV's arguments. All the referenced examples are listed at the end of the following paragraphs (pages 50 f.; for clarity, the literal translations have been omitted in most cases). The Portuguese SVCs in the examples are highlighted in italics, the considered phenomena in boldface.

The sentence structure around the SVCs corresponds to the structures given in Gärtner (1998, p. 112 f.) (see section 1.4). Out of these, we mostly discover the following three patterns, the second and third being the most frequent:

Subject + (SV + SubstantiveGroup)
Subject + (SV + SubstantiveGroup) + indirectObject
Subject + (SV + SubstantiveGroup) + prepositionalObject

We will use the abbreviations SV and SbG to refer to the SVC components.

[36]Recall from section 4.1 that the FVs are *ameaçar* ('to threaten'), *apoiar* ('to support'), *faltar* ('to lack'), *perguntar* ('to ask'), *prometer* ('to promise') and *responder* ('to answer').

The SVs in the observed SVCs basically have trivalent valency. That is, they require a subject, a direct object (being the SbG), and an indirect object. However, there is high corpus evidence that they can be used as well in a divalent way. For example *fazer pergunta* is used without an indirect object in 58% of the observed cases, as in example (35).

4 of the 100 sentences are interrogative. Compared to the overall distribution of interrogative sentences in the Portuguese portion of our corpus, this percentage is representative: out of 1,106,987 sentences, there are 42,000 interrogative which corresponds to 3.79%.

Passive constructions are rare: in 260 observed sentences (2×50 plus 20×8, not counting twice the cases for *fazer pergunta* and *dar apoio*), there are only three passive constructions, one of them shown in example (36). This corresponds to a percentage of 1.1%. We do not provide comparative figures for the whole corpus as it is a separate task to reliably extract passive constructions.

The analysed data reveals that some expressions surrounding the SVCs seem to be specific, either for the SVCs, for the corpus domain or for both. For example, the VP *gostar de* ('to like to') introduces the SVC *fazer pergunta* in 18 of 50 cases (36%). We believe that the corpus domain plays an important role here, because asking a question in a polite form – as the members of the European Parliament should do – requires such introduction.

Patterns within the SVC. As mentioned in section 3.2.2, the string of the actual SVC tokens can be interrupted by other tokens. These tokens – as well as the SbG, being one of the SV's arguments – can be relevant concerning the argument detection. Table 9 shows the most recurrent patterns and their occurrence in the 100 sample sentences; relevant information are in boldface.

	Pattern	# occurrences (in 100 sentences)
1.	**SV**$_{Subj}$ (ADJUNCT$_{SbG}$)* SbG	32
2.	SV ADJUNCT$_{SbG}$* SbG	84
3.	SV-**PRON**$_{IndObj}$ (ADJUNCT$_{SbG}$)* SbG	n/a

Table 9: Patterns, arguments and their occurrence within an SVC

The patterns are as follows:

1. The SV incorporates the subject, and neither a subjective pronoun nor an NP are realised; see example (37). This information is relevant for our purposes.

2. One or several adjuncts modify the SbG, as in example (38). Mostly, they

are determiners, adjectives, adverbs, pronouns and cardinals. These tokens are frequent – that is why they are mentioned in several rows in table 9 – but irrelevant.

3. The indirect object of the SV in form of a pronoun is put in front of the direct object (the SbG) as in example (39). Such movements happen because of syntactic rules. The pronoun is appended to the SVC's verb (recall from section 2.2.3 that we introduce a specific POS tag for such constructions, starting with the prefix 'VREF'). Thus, the construction is not, e.g., *dar apoio*, but *dar-lhe apoio* and we cannot provide figures relative to the expression *dar apoio*. Nonetheless, the pattern is relevant.

Patterns preceding the SVC. Above all, we expect the subject to precede the SVC. But there is also other relevant information. Table 10 shows the relevant and irrelevant patterns found; they are described below.

	Pattern	# occurrences (in 100 sentences)
1.	– SVC	9
2.	$\mathbf{VP}_{Subj}$ SVC	36
3.	$\mathbf{NP}_{Subj}$ SVC	8
4.	$\mathbf{PERSP}_{Subj}$ SVC	2
5.	IMPERSONAL SVC	2
6.	$\mathbf{PRON}_{IndObj}$ SVC	7
7.	SUBCLAUSE SVC	11
8.	SUBCONJ SVC	11

Table 10: Patterns, arguments and their occurrence preceding an SVC

1. No preceding context, the sentence starts with the SVC (example (40)). Irrelevant.

2. A VP incorporates the subject. This is the case for verbal periphrases (see Gärtner (1998, p. 32)) which consist of a verb acting as auxiliary, and a second verb in infinitive or gerund form. Verbal periphrases can be *i)* modal, e.g. *gostar de*, or *ii)* temporal, as in the Portuguese tense 'futuro composto'; see examples (41) and (42). Relevant.

3. An NP as subject, see example (43). Relevant.

4. A nominative personal pronoun as subject, see example (44). Relevant.

5. Impersonal constructions (Gärtner, 1998, p. 117, 119), i.e. impersonal or non-existent subjects. There are manifold realisations, e.g. V + *Adj* (*é necessário* – 'it is necessary') or V + *que* + V_{Inf} (example (45)). Irrelevant.

6. The indirect object as non-nominative pronoun, put in front of the SVC due to syntactic rules[37] as in example (46). Relevant.

7. Subordinate clauses and adjuncts inserted with commas, not belonging to the SVC's arguments (example (47)). Irrelevant.

8. Subordinating conjunctions, as the SVC is located in a subordinate clause (example (48)). Irrelevant.

Patterns following the SVC. We assume that the context following an SVC contains the indirect object. The most frequent patterns are shown in table 11 and described below.

	Pattern	# occurrence (in 100 sentences)
1.	SVC .	16
2.	SVC ADJUNCT$_{SbG}$*	18
3.	SVC , ADJUNCT ,	4
4.	SVC (ADJUNCT)* **NP**$_{IndObj}$	54
5.	SVC (ADJUNCT)* PREPOBJ/DIRREF	6
6.	SVC [SUBCLAUSE,MAINCLAUSE]	8

Table 11: Patterns, arguments and their occurrence following an SVC

1. Punctuation denoting the end of the sentence (example (44)). Irrelevant.

2. One or several adjuncts modify the SbG, see example (49). Possible modifiers are the same as within the SVC. Irrelevant.

3. Other adjunctive phrases delimited by commas or dashes, see example (50). Irrelevant.

4. A prepositional object as *complement* (indirect object) of the SV. The prepositions might depend on SVC and context; in the observed sentences,

[37] For example, the object has to precede the verb if the verb and its arguments are negated or situated in a subordinate clause. It is called the 'próclise' form; see Gärtner (1998, p. 87).

it is always *a* (∼ 'to'). The complement follows the SVC directly, as in example (51), or separated by adjuncts/short adjunctive phrases. Relevant.

5. A prepositional object or directional reference as *adjunct* of the SV. The prepositions again depend on both SVC and context; we retrieve *sobre* (∼ 'about'), *de* (∼ 'of') and *em* (∼ 'in'), see example (52). Irrelevant.

6. Subordinate or main clauses which do not belong to the SVC's arguments, appended by coordination, by a subordinating conjunction or by a comma; see examples (53) and (54). Irrelevant.

---

(35) [. . .] gostaria de *fazer* as seguintes *perguntas* :
[. . .] I would like to ask the following questions :

(36) **Foi** *feita* uma **segunda** *pergunta* sobre o mainstreaming .
You asked a **second** question , which concerns mainstreaming .
(lit.: 'A second question has been asked about mainstreaming.')

(37) ***Faço*** a *pergunta* , mas receio já conhecer a resposta .
I ask the question , but I am afraid I already know the answer .

(38) *Damos* **o nosso total** *apoio* ao relatório do senhor deputado Napolitano .
We give **our unqualified** support to the report by our colleague Mr Napolitano .

(39) *Dou*-**lhe** o meu sincero *apoio* .
I wholeheartedly endorse **it** .

(40) ***Fazer*** *esta pergunta* é também respondê-la simultaneamente .
To ask the question is to answer it .

(41) **Gostaria de** *fazer* várias *perguntas* ao senhor comissário .
I should like to ask the commissioner the following points .

(42) [. . .] **vou** *fazer* algumas *perguntas* incisivas .
[. . .] **I am going to** put several urgent questions to him .

(43) [. . .] **o nosso grupo** *dará* o seu *apoio* a ambos os relatórios .
[. . .] **our group** supports all three of these reports .

(44) Foi por isso que **eu** *fiz* a *pergunta* !
This is why **I** asked the question !

(45) Ainda assim , penso que **é importante** *fazer a seguinte pergunta* :
I nonetheless think **it is important to** *ask the question* .

(46) Senhor presidente , gostaria de **lhe** *fazer* uma *pergunta* .
Mr president , I should like to ask a question .

(47) Devíamos **, do lado europeu ,** *dar* aqui o necessário *apoio* .
We in Europe should provide the necessary support .
(lit.: 'On the part of Europe, we should provide here the necessary support.')

(48) [. . .] é com todo o prazer **que** *dou* o meu *apoio* ao relatório do senhor deputado Ford .
[. . .] it is a great pleasure **to** rise in support of Mr Ford's report .

(49) Gostaria de lhe *fazer* três *perguntas* **rápidas** .
I should like to ask him three **quick** questions .

(50) A comissão tem de tentar *dar apoio* **, por todas as formas ,** aos jornais que apontam esse problema .
The commission will have to support the newspapers which expose this **, in all sorts of ways** .

(51) Temos de *dar resposta* **a esta pergunta** !
We do need answers here !
(lit.: 'We have to give answer **to this question**!')

(52) [. . .] só quero *fazer* algumas *perguntas* **sobre este assunto** .
[. . .] there are a number of questions I should like to ask **on this topic** , [. . .]

(53) [. . .] *fiz* a *pergunta* delicadamente **, por isso peço uma resposta .**
[. . .] I asked him politely **, and so I would like an answer .**

(54) *Fiz* algumas *perguntas* **, mas vou deixar as coisas como estão** [. . .]
I have raised one or two questions **, but I shall just leave it at that** [. . .]

Noun phrase patterns. In the next step, we consider in more detail the NPs filling the SV's argument positions – subject, direct object and indirect object. Recall that typical NP components are determiners, adjectives, adverbs, pronouns and cardinals. The most frequent patterns, including respective examples, are listed in tables 12 and 13, separately for *fazer pergunta* and *dar apoio*. We separate the results of the SVCs to show any expression-specific phenomena.

Note that the reported sum of subject and indirect object arguments is below 50 for both SVCs because they do not occur as NPs in all observed sentences. Especially the subject position is often *not* realised as NP (see example (37)). The overall number of subject and indirect object NPs is higher than the number of NPs indicated in tables 10 and 11. This is because some NPs are located further away and thus have not been counted in the analyses above.

Pattern	Example	Occ. (in 50 sentences)		
		Subj	DirObj	IndObj
Noun*	senhor presidente (Mr president)	0	5	0
(Det)* Noun*	o senhor presidente (Mr president)	0	21	17
Z (Adj) Noun / Z Noun (Adj)	duas perguntas (two questions)	0	9	0
(Det) (Rg) Adj Noun / (Det) Noun (Rg) Adj	o presente relatório (the current report)	0	14	0
Pi Det Noun	ambos os relatórios (both reports)	0	0	0
(Det) Det Px Noun (Adj)	os nossos colegas (our colleagues)	1	0	0
(Det) Noun Cc (Det) Noun	Watts e outros (Watts and others)	1	0	0

Table 12: NP patterns and their occurrence in subject, direct and indirect object position of the SVC *fazer pergunta*

Pattern	Example	Occ. (in 50 sentences)		
		Subj	DirObj	IndObj
Noun*	senhor presidente (Mr president)	0	5	2
Det Noun*	o senhor presidente (Mr president)	9	3	24
Z (Adj) Noun / Z Noun (Adj)	duas perguntas (two questions)	0	0	0
(Det) (Rg) Adj Noun / (Det) Noun (Rg) Adj	o presente relatório (the current report)	0	8	6
Pi Det Noun	ambos os relatórios (both reports)	0	0	1
(Det) Det Px Noun (Adj)	os nossos colegas (our colleagues)	1	34	1
(Det) Noun Cc (Det) Noun	Watts e outros (Watts and others)	0	0	0

Table 13: NP patterns and their occurrence in subject, direct and indirect object position of the SVC *dar apoio*

Above all, the difference between the figures in table 12 and 13 show that there indeed happen expression specific assignments of argument positions: for example, the subject position in *fazer pergunta* is rarely filled by an NP (see the third column in table 12), whereas *dar apoio* chooses nine times a Det Noun* -NP as subject. *Dar apoio* has 34 contexts in which a possessive pronoun (PX)

occurs in the direct object, but for *fazer pergunta*, this is never the case.

The SVCs' usage of a specific NP pattern in a specific argument position is mostly different. Only the following behaviours are approximately similar:

- Occasionally: NOUN* -NPs in the direct object position (5 in both expressions)
- Frequently: DET NOUN* -NPs in the indirect object position (17 versus 24)
- Relatively frequent: NPs with an ADJ and optionally a DET and/or RG in the direct object position (14 versus 8)

Apart from that, we only detect *un*typical NP patterns for specific argument positions, i.e. the zeros in the table.

Patterns surrounding the full verbs. As mentioned above, we also analyse 6×8 randomly selected sentences containing the FVs used for SVC acquisition, in order to detect similarities between the arguments of FVs and SVs.

However, our study does not lead to clear results. Most FVs have even more diverse contexts than the SVCs examined above. Three characteristics can be registered, though:

- The verbs are often used with impersonal expressions. For example, both *apoiar* and *perguntar* are used in impersonal context in 37.5%.
- The verbs frequently introduce sub-clauses and thus do not realise a 'standard' object argument but have a sentential complement (COMP) as argument. For *perguntar*, this is the case in 62.5% (e.g. 'I ask if ...').
- The verbs are frequently used in participle form, so that they are not assigned any arguments at all. For *ameaçar*, this happens in 62.5%.

5.1.3 Conclusions

We can affirm the following assumptions made concerning the behaviour of the SVC's arguments:

- There *are* recurrent patterns surrounding an SVC.
- The SV's subject is frequently its agent, although the syntactical realisation varies.
- The SVC's SbG corresponds to the direct object.
- The indirect objects are, if realised, appended with a narrow range of prepositions, which would ease a pattern-based extraction.

However, we have to reject a central assumption: the text surrounding the SVCs is not as uniform as expected and thus cannot be abstracted for argument generalisation in an easily applicable way. There are three main problems:

1. Context diversity: although some of the figures in tables 9, 10 and 11 seem to be representative, the retrieved patterns are not salient enough to sufficiently characterise the behaviour of the SV's arguments: there are various patterns for each of the three examined areas (*within*, *preceding* and *following* the SVC). Additionally, tables 12 and 13 reveal that the NPs filling the respective argument positions are very diverse. Impersonal expressions pose another problem to that task (see also the ambiguous cases reported in Hendrickx et al. (2010)). Taking into account all these alternatives makes it difficult to achieve both reliable coverage and correct detection of argument boundaries (which is, according to the literature, a crucial point) without being too specific or too general.

2. Expression specificity: the analysis of the NPs shows that the NPs' realisation is expression-specific. There is evidence that these particularities can be applied to other SVCs, because the superficial analysis of the SVCs extracted with the verbs *ameaçar*, *faltar*, *prometer* and *responder* show similar behaviour. However, we cannot assume that these analyses are generally representative. Developing a system which is potentially highly adapted to specific SVCs is not desirable as we try to retrieve general structures.

3. Preposition usage: we report in section 5.1.2 that complementary and adjunctive NPs occur with distinct prepositions. It is tempting to assume that the prepositions are systematically separable into indicators for complementary and adjunctive NPs. However, there is no evidence for clear regularities (see also Danlos (1992) who points out the problem of distinguishing between complementary and adjunctive objects, and e.g. Döll and Hundt (2002, p. 147) who present SVCs with various prepositions for the indirect object). To the best of our knowledge, there are no studies whether specific prepositions are used for the complements of Portuguese SVs, so that we refrain from making uncertain assumptions.

We also did not retrieve similar contexts for FVs and the related SVCs: the contexts of the FVs are far too heterogeneous to draw any well-founded conclusion.

Regarding our idea to incorporate context information by developing pattern-based extraction methods, the results of this extensive study are disappointing. As there is no parser available for Portuguese, there are no straightforward and promising possibilities to use contextual information to filter the list of acquired SVCs.

However, there is still a chance to filter out the false positives of the pivot step by computing association measures and ranking the results according to these measures. The following section elaborates this approach.

5.2 Filtering with association measures

As reported in section 1.2, there are several monolingual approaches which compute different association measures (AMs) to establish a ranking for collocation retrieval. For example, Cinková et al. (2006) compute AMs to retrieve SVCs. Similarly, Krenn and Evert examine this field exhaustively (Krenn (1999), Krenn and Evert (2001), Evert and Krenn (2001)). They figure out that, for SVCs, the Student's t-test and simple frequency perform best, with the t-test hardly being significantly better. The χ^2, log-likelihood and pointwise mutual information (PMI) measures perform poorly in their studies.

The main difference of our approach is that our acquisition of candidate SVCs happened in a bilingual way. The filtering, however, can as well be carried out in a monolingual way. It is an interesting question to what extent monolingual methods can improve the multilingual procedure, and the work presented in this section aims at answering that question.

We compute several AMs – i.e. PMI, Student's t-test and frequency – for Portuguese verb-noun pairs extracted from our corpus. Then, the values of the verb-noun pairs from the pivot pipeline results (i.e. true and false positive SVCs) are compared with the values of the overall verb-noun pairs in order to detect specific behaviours for SVCs. Apart from the calculation of the AMs, we implement some parameters and preselection steps to ensure reasonable measure results.

The following section gives some information on the measures and how we apply them in our approach. Then, we describe the implementation, i.e. the adaptable parameters and preselections, and finally present the results.

5.2.1 Explanation of the AMs

Pointwise mutual information. Pointwise mutual information comes from information theory and is perfectly suitable for the identification of collocations (see Manning and Schütze (1999, p. 178 f.)). It is computed by observing the relationship between the frequency of a collocation and the frequency of its individual parts. Equation (6) shows how PMI is computed in our setting. Note that we omit the division by the total number of tokens in our calculations, as it would just increase the values but not affect their ranking.

$$PMI_{verb \bullet noun} = \frac{freq(verb, noun)}{freq(verb) freq(noun)} \tag{6}$$

The idea we pursue with the application of PMI is the following: we expect the PMI values for the verb-noun pairs of SVCs to be rather low, because SVs are expected to occur frequently and combined with many different nouns. However, the PMI of arbitrary verb-noun combinations should be even lower.

Our assumptions turn out to be correct. For example, *dar + apoio* has a PMI

of 4.472×10^{-5} on our corpus data[38]. In contrast, verbs and nouns which are used in a specific context, e.g. 'pay + bill' or 'read + book', are supposed to have higher PMI values. For the Portuguese portion of our corpus data, an example for such expressions is *pagar + imposto* ('pay + tax'), which has a PMI of 1.5×10^{-2}. On the other side of the scale, there are verb-noun pairs containing an SV which occur incidentally and are no meaningful verb-noun combinations. They have lower PMIs than SVCs, e.g. 1.652×10^{-8} for *fazer + terrorista* ('make + terrorist'), which cooccur only once.

Krenn and Evert (2001) report that the PMI values for SVCs lie in a relatively narrow range. However, the detection of candidate SVCs from a PMI-ranked list performs poorly, as this range is not at the top of the list. Therefore, the authors decide to conduct a shake-up of the ranking, allowing for the SVC-typical PMI range to be at the top of the list, which outperforms all other AMs tested.

As we are faced with the same problem, we also test Krenn and Evert's proposal; our findings are presented in section 5.2.3.

Student's t-test. We decided to apply Student's t-test on our task because it performs best in the proceedings of Krenn and Evert (2001).

This test is a statistical test which checks if a hypothesis is confirmed or not. Its calculation for cooccurrences is described in Manning and Schütze (1999, p. 163 ff.). The t-test takes account of an important question: is a frequent cooccurrence between two words by chance, i.e., are the two individual words so frequent that their cooccurrence is more probable than it would be for less frequent words? It is computed as shown in equation (7):

$$tScore_{verb \bullet noun} = \frac{\chi - \mu}{\sqrt{\frac{s^2}{N}}} \tag{7}$$

where χ is the *observed* cooccurrence of verb and noun, μ is their *expected* cooccurrence (i.e. $p(verb)p(noun)$), s^2 is the sample variance and N is the sample size (i.e. number of observed items).

The resulting t-score can be easily classified as significant or not. We use this side effect to immediately discard insignificant values.

We straightforwardly implement Manning and Schütze (1999)'s proposal for the application of Student's t-test on collocation extraction. Note, however, two modifications we have made:

- Instead of simply equating s^2 with χ, we use the – more correct – formula $s^2 = \chi(1 - \chi)$, as our values for χ might vary.
- As underlying sample size N, we do not use the overall number of corpus tokens, but the overall number of extracted verb-noun pairs, as these pairs are our data base and all other tokens are not of interest.

[38] Computation restricted to a minimum cooccurrence of 20.

Frequency. Counting and normalising frequencies, although not being an AM in the strict sense but a component for AM calculation, is a fairly intuitive way to evaluate collocations. Despite its simplicity, frequency achieved some of the best results in Krenn and Evert (2001). It is computed by counting the number of cooccurrences of the considered verb-noun pair and divide them by the number of all occurring verb-noun pairs; see equation (8).

$$freq_{verb \bullet noun} = \frac{freq(verb, noun)}{\sum_{i=verb_1}^{verb_m} \sum_{j=noun_1}^{noun_n} freq(verb_i, noun_j)} \tag{8}$$

5.2.2 Restrictions and adjustable parameters

Obviously, the choice of the AM itself is one influencing factor for the results. The differences measured in Krenn and Evert (2001) let us expect that the AMs lead to different results.

Moreover, we define some restrictions and parameters which can be combined to various experimental filtering settings. They are explained in the following paragraphs.

Extraction of verb-noun cooccurrence pairs. First of all, we experiment with different strategies to acquire the verb-noun pairs from our corpus, searching for pairs which are supposed to be pairs of verb and direct object. There are four implementations available:

1. Pattern-based extraction. For the extraction of verb-noun cooccurrences by means of POS patterns, we make use of the analysis in section 5.1, more precisely, of the patterns retrieved within the SVCs (table 9). We extract all verb-noun cooccurrences in the corpus which match the following regular expression, starting at the verb and searching for the first following noun:

V [Det Adj Rg Pronoun Z]* N

Note that we accept all pronouns, apart from relative pronouns like *cujo* ('whose') which introduce a sub-clause.

Other POS tags (i.e. all irrelevant sentence parts) are considered as breakpoints for the search. Although this restriction leads to the loss of some cases, it is *i)* impossible to cover all possible patterns and *ii)* more important to us to extract verb-noun pairs which are likely to be verb-DirObj pairs.

2. Bag-of-words extraction I: noPattern. Another, simpler approach is to search for verb-noun cooccurrences within a specific word window size (Lund and Burgess, 1996). Starting from the position of the current verb, the algorithm searches for the first following noun within this window. Obviously, it is less restricted, as it allows any POS pattern between the verb and the noun.

We aim at keeping this setting as unrestricted as possible; so we choose a virtually unlimited window size, reaching up to the end of the sentence.

Note that this is not a pure bag-of-words approach, which would work without consideration of any lexical information: we make use of the POS information to recognise verbs and nouns.

3. Bag-of-words extraction II: maxDistance. This extraction method additionally implements the aspect of the bag-of-words model which has been neglected in the noPattern method: the noun following a verb must fall within a certain range; otherwise, it is ignored. The distance can be set flexibly.

4. Pattern-based extraction with maximum distance. This setting is the combination of the first and the third proposal presented here: a pattern-based extraction – implementing exactly the same POS tag restrictions as above – with a maximum-distance bag-of-words approach.

Mininum cooccurrence threshold. Some AMs are known for being sensitive to the computation of low cooccurrence frequencies and getting a strong, unwanted bias. That is why we introduce a minimum cooccurrence threshold. The determination of the threshold value depends on the objective, i.e. if one wants to have a high precision or recall. Note that a 100% recall (pure ranking) is only guaranteed if one does not set any threshold at all. Section 5.2.3 indicates the settings with the chosen values.

The minimum occurrence restriction has a pleasant side-effect: it ensures – except for the 100% recall setting – that the resulting SVC list is cut off, as it removes all candidate SVCs whose cooccurrence frequency is not confident enough.

Keep only retrieved expressions in rank. The major source of information that we have for the filtering step is the result of the pivot pipeline itself: it provides us with a narrow preselection of possible SVCs. Of course, we take advantage of this information: we do not add any verb-noun pairs to the final result list that have not been acquired by the pivot step, even if they behave similar to SVCs.

Consideration of verb context diversity. Another idea is to presort the verbs in order to identify *i)* SVs and, within these, *ii)* SVs which are likely to occur in SVCs. For example, one could discard unfrequent verbs – they occur in too few contexts to be SVs. However, we find out that such a presort is not sensible: the expressions resulting from the pivot pipeline are situated within a narrow frequency range anyway. Within this range, one cannot draw a clear line which separates the true positives from the false positives. For example, the verbs *merecer* and *receber* which are (according to our gold standard) no participants of an SVC, have a corpus frequency of 4,344 and 6,504, respectively. The verbs

prestar and *colocar*, which are part of the SVCs *prestar apoio* and *colocar pergunta*, occur 4,200 and 7,401 times, respectively.

Apart from such a presorting, there is another possiblity to consider the distribution of candidate SVs: instead of counting the verb's overall frequency, one can count the amount of different verb-noun pairs in which it occurs (the verb's 'contexts'). Such a pair frequency provides useful information about the verb's behaviour.

The expectations for SVCs are the following: on the one hand, we assume SVCs to occur with verbs of a high context diversity, i.e. SVs which cooccur with many different nouns. On the other hand, SVCs can be realised as verb-noun collocations which cooccur only (or almost only) with each other. For example, the verb *correr* ('to run') occurs in merely one verb-noun pair with a cooccurrence frequency > 50, which is *correr + risco* ('run + risk'). According to Döll and Hundt (2002), this is a valid SVC. Note, however, that this example has a corpus specific bias. Newswire texts might use the verb 'to run' in more contexts, e.g., in reports about sports events.

Figure 9 shows an excerpt of the list of verbs and cooccurring nouns with their coccurrence frequencies extracted from our corpus. It has been extracted with the maxDistance strategy (maxDistance = 1) and a minimum cooccurrence threshold of 20. This excerpt illustrates that the assumptions made above are justified: SVs like *tomar* and *dar* occur in many different contexts (i.e. 15 and 66 nouns, respectively), partly containing SVCs. Verbs with only one or two cooccurring nouns are either lexically closely connected words, e.g. *comprar + produto* ('buy + product'), or an SVC, e.g. *correr + risco*[39]. Verbs which cooccur with an average number of nouns are not likely to comprise SVCs, e.g. *fixar* ('to fix').

Thus, we account for the verb's context size in the following way:

1. Count the number of different contexts (verb-noun pairs) per verb: $contextSize_v$
2. Calculate a statistical average value for the distribution over all verbs: $contextSize_{avg}$
3. Discard all verb-noun pairs whose verb context size is below this average: $if(contextSize_v < contextSize_{avg}) : discard(v)$

Additionally, we keep all verb-noun pairs with $contextSize_v = 1$, for example *correr + risco*, or *dar-lhe + apoio*[40]. The effect of taking into account these cases can be both good and bad, though, as there might be added false positives. However, it leads to higher recall.

[39]We would also denote *correr + perigo* ('run + risk'; lit. 'run + danger') an SVC, but did not find confirming examples in the literature.

[40]Note again that *dar-lhe apoio* has the indirect object *lhe* incorporated. This form occurs in less contexts than the semantic equivalent *dar + apoio*.

tomar: conhecimento=348, disposição=31, parte=196, precaução=26, nota=640, **posição**=306, conta=30, decisão=817, **medida**=1656, iniciativa=201, forma=31, consciência=179, lugar=28, posse=68, providência=79;
poupar: tempo=27, energia=36, esforço=27, dinheiro=39;
correr: perigo=48, risco=80;
fixar: limite=25, meta=20, prazo=21, objectivo=62, norma=21;
comprar: produto=22;
dar: expressão=43, informação=57, **importância**=30, emprego=34, força=54, instrução=59, lição=76, **passo**=218, vida=44, **resposta**=1285, acesso=24, garantia=121, luz=92, indicação=29, preferência=46, carta=27, espaço=22, conta=290, entrada=53, oportunidade=66, conselho=23, ênfase=69, conteúdo=33, resultado=57, prioridade=468, testemunho=32, esperança=40, azo=262, segurança=21, prova=501, altura=54, cobertura=25, exemplo=25, continuidade=175, dinheiro=62, origem=662, formação=38, cumprimento=110, sequência=23, momento=53, início=706, mostra=297, **assistência**=29, tempo=84, solução=106, confiança=26, atenção=104, **ajuda**=29, conhecimento=39, lucro=21, modo=27, seguimento=308, razão=105, lugar=243, aplicação=27, **apoio**=135, primazia=29, destaque=23, fruto=86, corpo=44, ordem=22, voz=54, execução=33, forma=84, sinais=55, quitação=84;

Figure 9: Examples for verbs and their cooccurring nouns (maxDistance = 1, threshold = 20)

We implemented several options. The verb context consideration can be:

- switched on or off
- based on the average context distribution of the verbs over the whole corpus or only of the verbs in the pivot pipeline results, and
- realised with median or (rounded-down) arithmetic mean as statistical average value.

These options result in five possible settings:

- none: no context restrictions
- allMean: context distribution computed with *all* corpus verbs, arithmetic *mean* as average value
- allMed: context distribution computed with *all* corpus verbs, *median* as average value
- expMean: context distribution computed only with verbs from pivot *expressions*, arithmetic *mean* as average value
- expMed: context distribution computed only with verbs from pivot *expressions*, *median* as average value

We will refer to these setting names in the following analysis. Their effect is explained in section 5.2.3.

Additional remark. Consider figure 9 again, in particular the words highlighted in boldface. The verb-noun pairs formed with *tomar* and *dar* comprise even more SVCs than those we are eager to find for our initial FVs. Some examples are *tomar medida* ('to take action'), *dar importância* ('to give importance') or *dar passo* ('to make a move'). The semantics of these constructions has nothing in common with the semantics of the SVCs we are interested in; they only share distributional and syntactic attributes with them. However, they are valid SVCs (for the mentioned examples, see Gärtner (1998, p. 79) and Döll and Hundt (2002, p. 151)).

Keep one-hit pivot results. Some input FVs for the pivot pipeline lead to only one candidate SVC. Apparently, these single results have good quality. For example, *arriscar* ('to risk') only leads to *correr + risco*, as well as *ameaçar* ('to threaten') just results in *constituir + ameaça* ('constitute + menace'), which is a valid SVC according to our gold standard. As *constituir + ameaça* occurs rarely, it would be excluded from both high-recall and high-precision threshold settings and we would not get any result for the input FV at all. To avoid such gaps, and based on the fact that single results seem correct in most cases, we decided to retain all verb-noun pairs which are the only output of the pivot pipeline.

We run several settings with different parameter assignments, always using the results of the pivot pipeline proceeded with the standard parameters (see section 3.3) as candidate SVC list.

5.2.3 Experimental settings

This section presents the behaviour of the parameters for different values and the settings we finally define for the AM filter. Right at the beginning, we intend to reveal the most salient finding: there are only a few parameter settings which cause clearly better or worse results than the others. Although there have been tested many settings for verb-noun pair extraction strategies, thresholds, and the consideration of the verb context, there are hardly any differences in the quality of the SVC rankings across many experiments. Even the variation of the AMs did not always yield striking effects. This contrasts with Krenn and Evert (2001)'s study which reveals noticeable particularities for some AMs.

Nonetheless, the filtering step leads to a substantial improvement. The lion's share of the refinement is due to the minimum cooccurrence threshold. Apart from that, the pivot pipeline preselection turns out to be mainly responsible for the good results: the range of expressions retrieved by the bilingual approach is already very narrow, so that only few AM experiments achieve further improvements

without notably lowering recall. Thus, our experiments show that the automatic acquisition of SVCs is a perfect field of application for multilingual methods like the pivot approach, but can indeed be filtered with monolingual techniques.

As mentioned in section 4.3, we calculate precision, recall, f_1 and – for the test runs with 100% recall – average precision.

In order to meet the main requirements one could impose on SVC detection, we define three basic settings, attaching importance to different measures:

- hiPrec: the filtering process focuses on precise results
- hiRec: the filtering process focuses on many results
- totalRec: the process does not filter the input list, but returns the whole list as a ranking

We will refer to these setting names in the following analysis.

AM choice. As mentioned above, there is no big difference between the overall results for different AMs. However, there are some tendencies. An interesting fact is that – contrary to Krenn and Evert (2001)'s results – the t-test performs worse than both frequency and PMI in most settings. This is especially visible in the average precision calculation of the *totalRec* setting.

Setting	AM	Threshold	Verb context	Results			
				Precision	Recall	F_1	Avg. Prec.
totalRec	PMI	0	none	.22	1.	.36	**.24**
totalRec	t-test	0	none	.24	1.	.39	.17
totalRec	freq	0	none	.22	1.	.36	.17

Table 14: Total recall results for all AMs (maxDistance = 1, *apoiar* only)

Consider table 14, showing the *totalRec* results of all AMs for the verb *apoiar*, having extracted the verb-noun pairs with the maxDistance strategy (the precision achieved here is higher than precision for *apoiar* in table 8 on page 43 because the maxDist extraction already filters out some false positive verb-noun pairs in our gold standard). While precision, recall and f_1 measures of course show no differences for the respective AMs in a 100%-recall task[41], the average precision does: obviously, PMI is more capable to establish a good ranking than t-test and frequency. However, note that the performance of the measures depends on the choice of the initial FVs to be paraphrased. This trend is observable in all tests we carried out. Especially the consideration or exclusion of the FV *apoiar* influences the results, probably due to the high amount of candidate SVCs (i.e. 64) for *apoiar* in the pivot step (see section 4.1).

The modification of the ranking for PMI proposed in Krenn and Evert (2001) does not lead to better results on our data: as the pivot results – the true positives

[41] The precision for t-test is slightly higher due to the exclusion of non-significant values.

as well as the false positives – are within a narrow PMI range, such a shake-up does not improve the ranking but just mixes it again. Thus, we decide to omit the proposed modification.

We will refer to the AM's influence once again in section 5.2.4 after having fixed the other, as yet unspecified parameters.

Verb-noun extraction strategy. As to the options for the extraction of verb-noun pairs, we observe that the more linguistically motivated pattern-based extraction does not lead to better results, but is computationally more expensive. Several tests reveal that, in most cases, the noun's distance from the verb is between 1 and 3, with the best results achieved with a distance of merely 1. Thus, we conclude that a POS pattern-based approach as well as the freedom of the noPattern setting are not necessary, if not counterproductive. For all test carried out below, we use the maxDistance, set to 1, as extraction method.

Minimum cooccurrence threshold. The definition of the cooccurrence threshold is highly corpus dependent. In particular, it depends on *i)* the corpus size and *ii)* the corpus balance. For example, corpora with a monotonous vocabulary might easily exceed such thresholds for a few prevalent expressions.

As to our data, we find the thresholds of 20 and 50 to be good values for the *hiRec* and *hiPrec* filter, respectively.

Consideration of verb context diversity. A careful choice of this parameter allows for rejecting another few false positives. For example, *exercer apoio* and *merecer apoio*, which are no SVCs according to our gold standard, are rejected by the *expMean* setting.

Table 15 shows the behaviour of the available settings. There is a simple reason why both *mean* settings perform visibly worse in recall than *allMed*, *expMed* and *none*: as is widely known, the arithmetic mean tends to be influenced by outliers while the median is less susceptible. There are many verbs which occur only in few contexts, especially for high cooccurrence thresholds like 20 or 50. For example, *comprar* (see figure 9) occurs in only one verb-noun pair. These verbs with sparse contexts cause a low median, which results in a low verb rejection rate and thus in higher recall. The figures in table 15 show that there is even no rejection for *median* thresholds at all, as they have the same results as *none*. At the same time, there are high-frequent verbs which have a broad range of contexts, e.g. *dar* with 66 contexts. These data affect the arithmetic mean, so that the *mean* settings have a high rejection rate. As a consequence, their recall is worse, but *expMean* achieves an outstanding precision value.

The effect of counting different amounts of verb-noun pairs, i.e. the verbs of all extracted verb-noun pairs (*all*) or only the verbs occurring in the candidate SVCs (*exp*), seems intuitive as well: the verbs occurring in the candidate SVCs are

Setting	Threshold	Verb context	Results		
			Precision	Recall	F_1
hiPrec	50	none	.76	.59	**.67**
hiPrec	50	allMean	.71	.45	.56
hiPrec	50	allMed	.76	.59	**.67**
hiPrec	50	expMean	**.91**	.45	.61
hiPrec	50	expMed	.76	.59	**.67**

Table 15: Results for different verb context consideration strategies (AM = PMI, maxDistance = 1, whole annotated set)

likely to be SVs and thus are used in many contexts; especially, they are used in more contexts than an ordinary verb (which is considered in the *all* count). Thus, counting only the context of potential SVs naturally leads to a higher arithmetic mean and thus, to higher precision for *expMean*.

Based on these figures, we decide the following: *expMean* is the appropriate strategy for a filter that focuses on high precision. As there is no difference between *allMed*, *expMed* and *none*, we choose *none* – having the least computational effort – as default option.

Pivot pipeline modification. In order to exhaust all setting possibilities, we also go one step back and modify the setting of the pivot pipeline: we choose a more restrictive setting for the cross-lingual step, combined with the *hiPrec* and *hiRec* filters in the monolingual step. We expect higher precision with such a setting. Therefore, we change the thresholds of the pivot standard parameters (see section 3.3) to higher values, i.e. 350 for OWEs and 9 for MWEs in the first pivot step, and 20 for MWEs in the second pivot step.

Pivot setting	AM setting	AM threshold	Verb context	Results		
				Precision	Recall	F_1
hiPrec	hiPrec	50	expMean	1.	.18	.31
hiPrec	hiRec	20	none	.65	.59	.62

Table 16: Results for a restrictive pivot pipeline setting with AM PMI

The results for these runs are shown in table 16. Although precision is perfect for the *hiPrec-hiPrec* setting, its recall is very low. More precisely, for the verbs *ameaçar* and *faltar*, there has no SVC been found at all. For each of the four other verbs, just one SVC has been found. As these are all correct, a 100% precision is achieved. The *hiPrec-hiRec* setting does not lead to striking results; however, it has a good balance between precision and recall. Thus, if one wants to achieve perfect quality, the *hiPrec-hiPrec* setting is recommendable.

Setting	Threshold	Verb context	Results			
			Precision	Recall	F_1	Avg. Prec.
hiPrec	50	expMean	**.91**	.45	.61	n/a
hiRec	20	none	.61	**.86**	**.72**	n/a
totalRec	0	none	.33	1.	.51	**.33**

Table 17: Best overall results with AM PMI (whole annotated set)

Setting	Threshold	Verb context	Results			
			Precision	Recall	F_1	Avg. Prec.
hiPrec	50	expMean	.9	.41	.56	n/a
hiRec	20	none	.6	.81	.69	n/a
totalRec	0	none	.36	1.	.53	.28

Table 18: Results with AM t-test (whole annotated set)

Setting	Threshold	Verb context	Results			
			Precision	Recall	F_1	Avg. Prec.
hiPrec	50	expMed	**.91**	.45	.61	n/a
hiRec	20	none	.61	**.86**	**.72**	n/a
totalRec	0	none	.33	1.	.51	.11

Table 19: Results with AM frequency (whole annotated set)

We think that these results justify the decision made in section 3.3: it is good to leave the (first) acquisition step less restrictive so that the (second) filtering step has a greater degree of freedom. Hence, we stick to the pivot standard parameters the as defined before.

5.2.4 Final setting and results

Summing up, the tests carried out on the parameter adjustments suggest the following settings for the SVC filtering:

- Verb-noun pair extraction: maxDistance = 1
- Minimum cooccurrence threshold: 50 for high precision (*hiPrec*), 20 for high recall (*hiRec*), 0 for full recall (*totalRec*)
- Verb context diversity: *expMean* for high precision, *none* as default

Tables 17, 18 and 19 show the evaluation results for the three basic settings with these parameters, carried out on the whole gold-annotated dataset. The figures show once again that PMI outperforms the other AMs: frequency performs

ameaçar: constituir ameaça;
apoiar: prestar assistência, prestar ajuda, prestar apoio, dar apoio;
faltar: haver falta, ter falta;
perguntar: fazer pergunta;
prometer: fazer promessa;
responder: dar resposta, tomar posição, *ter resposta*;

Figure 10: SVCs for the best *hiPrec* setting

ameaçar: constituir ameaça;
apoiar: prestar assistência, prestar ajuda, conceder ajuda, prestar apoio, conceder apoio, dar assistência, dar apoio, dar ajuda, *haver apoio*, *receber ajuda*, *receber apoio*, *oferecer ajuda*, *fornecer apoio*, *obter apoio*, *pedir ajuda*, *merecer apoio*;
faltar: haver falta, ter falta;
perguntar: levantar questão, colocar pergunta, colocar questão, apresentar pergunta, fazer pergunta, *formular pergunta*;
prometer: fazer promessa;
responder: dar resposta, tomar posição, dar+lhe resposta, *receber resposta*, *haver resposta*, *ter resposta*;

Figure 11: SVCs for the best *hiRec* setting

worse in the ranking (see value for average precision), and Student's t-test has basically slightly worse results. Thus, PMI is the best suitable AM in our task.

The results in table 17 are the final reference numbers for the success of our SVC acquisition approach proposed in this thesis, achieving a maximum precision of .91 and a maximum recall of .86. Figure 10 shows the extracted expressions for the best precision setting, figure 11 for the best recall setting (false positives are highlighted in italics). Expressed in total numbers, the *hiPrec* setting contains only one false positive of 12 returned SVCs. Of the 22 true positives, we find 20 correct expressions in the *hiRec* setting, with the two undiscovered SVCs being variations of a discovered SVC (*dar+lhe apoio* and *dar-lhe apoio*).

The AM filtering technique presented in this section achieves an enormous improvement of the results of the pivot pipeline: consider the figures for the whole annotated dataset in the first step, illustrated in the 'all' setting in table 8 on page 43, and the final results after the filtering in table 17, respectively. A comparison shows an increase of up to 65% in precision for the SVC acquisition task (i.e., an increase from .26 to .91). For the *hiPrec* setting, the f_1 value increases by 19%, even though there is a (theoretical) recall of 100% in the pivot results. Concerning the *hiRec* setting, the f_1 value's improvement even amounts to 30% (.42 versus .72).

This is a very satisfactory result for the overall process.

6 Conclusions and future work

In this thesis, various research aspects have been examined: first, we explored whether cross-lingual techniques are suitable for the extraction of syntactically and semantically valuable information for resource-poor languages.

Our adaptation of the pivot approach of Bannard and Callison-Burch (2005) for the acquisition of Portuguese SVCs proves that this question can be affirmed: the cross-lingual technique is perfectly applicable to the acquisition task, without requiring complex preprocessing: based on a simple POS pattern search, the approach leads to correct SVCs for every initial FV, detecting even unexpected but correct SVCs due to the data-driven method.

These resulting SVCs can be viewed as a newly created lexical resource. Every entry provides syntactic information, i.e. it shows which SV can be combined with which noun to construct an SVC. Similarly, the resource contains semantic information, as it represents one or several expressions which are semantically equivalent to the given FV. Thus, the pivot approach conducted on POS-tagged data serves as a solid base for the acquisition of SVCs.

Furthermore, we investigated to what extent the combination of mono- and multilingual methods yields synergy effects.

The results shown in the previous chapter prove that our combinatory approach achieves very good results in both precision- and recall-focused tasks. The AM filter improves the precision of the pivot approach by up to 65%, and an overall maximum precision of 91% and maximum recall of 86% are achieved. We have found that the PMI measure works best for the filtering, however, hardly performing better than the other AMs tested. Thus, while the bilingual approach seems to reliably acquire SVCs, the monolingual AM technique is capable to rank and refine these expressions and to separate real from apparent SVCs. In doing so, the bilingual step implements the syntactic and the raw semantic filter, the monolingual step implements the fine-granular semantic filter.

Alas, it was not feasible to extract information about the argument structure of the SVs in the acquired SVCs, which would have considerably increased the usefulness of the developed SVC lexicon, and maybe the results, as well. The possibilities to realise these arguments turned out to be too diverse and too case-specific to allow for an approach with only little linguistic information available.

In addition, it would have been desirable to discover the fine-grained differences between several SVCs which have been retrieved for the same FV, e.g. between *fazer pergunta* and *colocar pergunta* for the verb *perguntar*. However, this subtlety might not be necessary: both annotators found that many of the retrieved SVCs are semantically equally related with the initial FV (see section 4.2). This is an interesting fact, considering the numerous studies on the fine-granularity of SVCs.

Hence, we can come to a positive conclusion: we have achieved most of the original goals. We proved on the basis of several Portuguese FVs that an extraction

of SVCs with the procedure proposed in this thesis is possible, and that the results are reliable. There is just one caveat in terms of universal validity: a comparison of the results for the whole gold-annotated dataset and its subsets (e.g. consider the *totalRec* setting results in table 14 and tables 17, 18 and 19, respectively) shows notable fluctuations. Especially FVs that lead to many candidate SVCs in the pivot step, thus occur in heterogeneous contexts and have many various alignments, perform worse. In contrast, FVs with less potentially synonymous SVCs lead to better results. Hence, the success of our approach depends on the initial FV. In order to investigate the question, to what extent the FV affects the performance, it is necessary to carry out a broad study with many different FVs on a versatile corpus.

As the corpus limits the variety of traceable SVCs and we can only retrieve expressions which occur in EUROPARL, it is difficult to prove that the approach proposed in this thesis is applicable to the acquisition of any kind of SVCs from any semantic field. For example, *dar um grito* ('to let out a scream') does not occur in our corpus and can not be evaluated here. Hence, the ability of the developed software to detect SVCs should be interpreted with caution, despite its success in the present study.

Future work. The last aspect mentioned in the conclusions, i.e. the question if the presented approach has general validity and is applicable to other domains and semantic fields, can be tackled by using another bi- or multilingual corpus – if available for Portuguese data –, which covers another domain than EUROPARL. As argued in section 2.1, there are currently no good alternatives, however, the number of parallel corpora is growing and promises more data sources in the future.

Another way to avoid the corpus problem is to transfer the approach to another language pair, e.g. Portuguese and English or Portuguese and Spanish, for which parallel corpora might be available. Such a transfer can be done easily, as the only requirement – besides the corpus – is a POS tagger for both languages.

Apart from using different underlying data, it would be interesting to detect other SVC constructions, i.e. SVCs comprising a preposition like *pôr à prova* ('to put to the test'). In section 1.4, we argued that it is more difficult to extract these SVCs: the consideration of prepositional phrases leads to more possibilities and hence, to more noise in the extraction step. However, a more restrictive extraction method in the pivot step, a strict filtering and a sufficiently large corpus could reconcile these drawbacks. Additionally, one could test if the linguistic association measure proposed in Wermter and Hahn (2004) reliably recognises and ranks these more complex prepositional SVC structures.

We mentioned in section 5.2.2 that the verb-noun pairs which are focused by our filter strategies contain also other, semantically different but valid SVCs, e.g. *tomar medida* ('to take action'). We could take advantage of this fact and

implement an intelligent expansion of our SVC detection: instead of only acquiring SVCs which are semantically equivalent to the initial FV, we can additionally retrieve completely different SVCs, including their related full verb: all verb-noun pairs within the focused range are tested for the compliance of typical features of SVCs, e.g. different AM values, minimum cooccurrence of verb and noun, etc. In order to assign the proper meaning to each newly found SVC, i.e. to retrieve a synonymous FV where possible (recall from section 1.3 that not all SVCs have corresponding FVs), the pivot approach can be applied backwards: starting with a Portuguese SVC, one or more FVs are retrieved which are most frequently aligned with this SVC and hence have probably the same meaning. Such a reversal of the pivot approach is unproblematic; our evaluation in section 2.4.3 shows that the connection between SVCs and FVs is applicable in both directions.

Furthermore, another filtering technique can be used: instead of separating real from apparent SVCs by means of selected verb-noun pairs, one could go more into depth and apply a complete semantic vector space to that task. Such vector space models offer manifold varieties to compare the semantic similarity of two items. On the one hand, the similarity calculation can be carried out with different methods (e.g. space dimensionality reduction) and measures, of which cosine is the simplest. On the other hand, different kinds of target words can be focused (words, phrases, word pairs etc.), and their context information can be acquired in different ways, i.e. the dimensions of the space can be filled with varying information. For example, there are bag-of-words approaches (Lund and Burgess, 1996) or the usage of patterns of any kind, like pair patterns to retrieve relations as *CauseEffect(x, y)* in Turney (2008), or Padó and Lapata's dependency parse patterns, which lead to a linguistically more informed vector space (Padó and Lapata, 2007). But as well, whole documents can be mapped into the dimensions (Landauer and Dumais, 1997). We assume that the increase of available information and possibilities by using a semantic vector space improves the results and leads to a more accurate SVC acquisiton.

We hope that the success of the two-stage approach presented here, combining bi- and monolingual techniques, gives rise to further experiments and implementations of one or the other of our proposals, or completely new combinatory settings.

Appendix

A Conversion from UTF8 to Latin1 with `sed`

Table A.1 lists the UTF8 and the corresponding Latin1 characters which have been converted with the Linux shell command `sed` before executing the POS taggers.

UTF8 character	Latin1 character
–	-
…	...
"	"
“	"
”	"
‚	’
´	’
`	’
'	’

Table A.1: Character conversion from UTF8 to Latin1

B Token-POS-patterns for the analysis of alignments of Portuguese SVCs

This appendix lists the patterns which have been used for the acquisition of occurrences of *dar apoio*, *fazer leitura* and *fazer pergunta*. Lower case strings denote a token, strings with capital initial letters stand for a POS tag. A single 'X' acts as wildcard for any POS tag. POS tags in brackets are optional.

dar (D) (D) (PX) (A) apoio
dar A D PX apoio
dar RG (D) (PX) apoio
dar D D PX apoio
dar Fc X X Fc (D) (PX) apoio
dar Fc X Fc (D) (PX) apoio
fazer (RG) D leitura
fazer (D) (PX) A leitura
fazer (RG) (RG) D pergunta
fazer (RG) RG (Z) pergunta
fazer (Z) (A) pergunta
fazer D (PX) (A) pergunta

C Initial letters of the PAROLE POS tagset

This appendix points out the coarse word classes for the PAROLE tagset. As the initial letter of a POS tag is sufficient to indicate the respective word class, only these letters are specified in table C.1.

Initial letter	Part-of-speech
N	Noun
V	Verb
D	Determiner
A	Adjective
R	Adverb
P	Pronoun
S	Preposition
C	Conjunction
I	Interjection
Z	Cardinal
W	Date and time
F	Punctuation

Table C.1: Initial letters of the PAROLE tagset

D Extracted candidate SVCs for the verb *apoiar*, using Grow-Diag-Final symmetrisation

This appendix lists all 68 expressions which have been extracted for the FV *apoiar* in the pivot pipeline by means of Grow-Diag-Final symmetrisation. The expressions considered as SVCs due to the gold standard are highlighted in boldface.

ser ajuda
merecer apoio
receber apoio
prever apoio
ter apoio
conceder apoio
pedir apoio
dar apoio
solicitar apoio
considerar apoio
dar assistência
constituir apoio
ser apoio
ser favor
dar-lhe apoio
oferecer apoio
prestar ajuda
agradecer apoio
receber ajuda
conceder ajuda
registar apoio
expressar apoio
dizer apoio
manifestar apoio
reforçar apoio
garantir apoio
prestar apoio
existir apoio
dar ajuda
obter apoio
manter apoio
proporcionar apoio
incluir apoio
conseguir apoio
procurar apoio
recolher apoio
conquistar apoio
apoiar proposta
exigir apoio
encontrar apoio
aumentar apoio
requerer apoio
pedir ajuda
exprimir apoio
esperar apoio
fornecer apoio
retirar apoio
assegurar apoio
confirmar apoio
declarar apoio
granjear apoio
apreciar apoio
dar+lhe apoio
saber apoio
afirmar apoio
saudar apoio
prestar assistência
reiterar apoio
haver apoio
demonstrar apoio
justificar apoio
reunir apoio
ser promoção
providenciar apoio
reafirmar apoio
perder apoio
disponibilizar apoio
oferecer ajuda

E Annotation guidelines for the evaluation of candidate Support Verb Constructions

This appendix presents the original guidelines used by the annotators.

1 About Support Verb Constructions (SVCs)

- SVCs are syntactic constructions which consist of a prepositional or non-prepositional object and a so-called 'support verb'. Within this SVC, the support verb has lost almost all of its semantic meaning, whereas the noun of the object presents the semantic core.
- Often (but not always), there are main verbs which can semantically substitute the SVC.
- Examples in German and Portuguese are:
  - non-prepositional: eine Frage stellen (fragen); fazer uma pergunta (perguntar)
  - prepositional: in Kraft treten; entrar em vigor

2 About the data

- There is given one PDF with sample sentences for the extracted expressions of each verb:
  - the resulting (candidate) SVCs for the main verb 'apoiar'
  - the resulting (candidate) SVCs for the main verb 'perguntar'
  - ...
- For each expression, there is a headline starting with '***'. It indicates the respective expression and gives an opportunity to evaluate the expression. Then follow the sample sentences.
- For each expression, a maximum of eight sample sentences is printed; if its occurrence in the corpus is lower, there are less examples.
- At the end of each list, there is a form to establish a top-five ranking of the expressions.

3 What needs to be evaluated?

- If there is no doubt that the expression in the headline is nonsense, it is enough to mark 'no SVC' in the headline.
- If the expression is (possibly) an SVC, all sample sentences should be read carefully, having in mind the following restrictions.
- At the end, make a ranking (described below).

4 How should be evaluated?

- First of all, decide whether the expression is an SVC or not. If necessary, make use of the sample sentences. Sometimes, it is difficult to assess whether an expression is actually an SVC or just a frequent phrase. Please do not judge prematurely.
- If you judge the expression an SVC: is the SVC in the context of the respective sentences semantically substitutable with the respective main verb ('apoiar', 'perguntar', ...)? Write down the number of correct replacements in the headline.
  Please note: the SVCs should only be tested for semantic appropriateness. It is not important to keep a syntactically correct sentence (i.e. incorrect syntactic structure, missing or redundant prepositions, pronouns etc. can be disregarded).
  It is possible that there occur expressions which are valid SVCs but which are not semantically equivalent to the given main verb. In this case, mark the expression as 'SVC' but set the '# correct replacements' to zero.
- After having read all proposed expressions, please make a qualitative ranking of the top five SVCs. The better an SVC substitutes the main verb, the higher is its ranking. If you consider less than five proposals as valid SVCs, rank only these expressions and do not fill up the ranking with incorrect expressions. If you consider several SVCs equally good, put them onto the same ranking position.

Muito obrigada pela sua ajuda!

F Evaluation file for the verb *prometer*

This appendix presents one of the annotation files given to the annotators. Due to layout issues, it has been slightly modified.

*** 'fazer promessa': ___ SVC ___ no SVC ___ # correct replacements
infelizmente , não é a primeira vez que nos fazem tais promessas e que assumimos essa responsabilidade .
com efeito , a nível europeu corremos o risco permanente de fazer promessas que posteriormente não podemos cumprir por razões de natureza orçamental .
não podemos fazer promessas num dia , e rapidamente esquecê-las no dia seguinte .
confesso que , quando faço uma promessa , gosto de a cumprir .
fazem se muitas promessas , mas , infelizmente , muitas não são cumpridas .
os estados-membros tendem muitas vezes a fazer grandes promessas , mas depressa as esquecem quando chega a altura de pagar .
não o disseram . apenas fizeram uma promessa , mas não é de promessas que precisamos .
fez a promessa de acelerar esta questão no conselho .

*** 'cumprir promessa': ___ SVC ___ no SVC ___ # correct replacements
os doadores - entre os quais a união europeia - não teriam , alegadamente , cumprido as suas promessas .
estamos a ver um governo determinado , interessado em cumprir as suas promessas .
estou certo de que a senhora deputada sandbæk esperará que cumpra a sua promessa de lhe responder por escrito .
com essa política , não conseguirá cumprir a promessa do pleno emprego .
iremos acompanhar a situação , procurando certificar nos de que a comissão está a cumprir as promessas que fez .
temos de estar dispostos a cumprir esta promessa assim que a vontade do povo for respeitada .
espero que possa cumprir essas promessas , senhor primeiro-ministro .
cumpriu a sua promessa .

*** 'honrar promessa': ___ SVC ___ no SVC ___ # correct replacements
ficou claro que a comissão europeia não honrou a sua promessa .
o sucesso desta importante ronda de negociações irá em parte depender da vontade de honrar essa promessa .
poderá a comissão explicar por que razão não iniciou este processo e quando

tenciona honrar a sua promessa ?
e votei contra um estatuto de parceria , porque isso significaria que a ue não honraria as suas promessas .

*** Top 5 ranking of the valid SVCs

1: _____

2: _____

3: _____

4: _____

5: _____

Bibliography

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

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

Thorsten Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing (ANLP-2000)*, Seattle, Washington, 2000.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert. L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19:263–311, 1993.

Claudia Brugman. Light verbs and polysemy. *Language Sciences*, 23:551–578, 2001.

Miriam Butt. The light verb jungle. Number 9 in Havard Working Papers in Linguistics, pages 1–49. 2003.

Miriam Butt and Wilhelm Geuder. On the (semi)lexical status of light verbs. In Norbert Corver and Henk van Riemsdijk, editors, *Semi-lexical Categories*, number 59 in Studies in Generative Grammar, pages 323–370. Mouton de Gruyter, Berlin, 2001.

Hadumod Bußmann, editor. *Lexikon der Sprachwissenschaft*. Alfred Kröner Verlag, Stuttgart, Germany, 4th edition, 2008.

Chris Callison-Burch. *Paraphrasing and Translation*. PhD thesis, University of Edinburgh, Edinburgh, 2007.

Xavier Carreras, Isaac Chao, Lluís Padró, and Muntsa Padró. FreeLing: an open-source suite of language analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, 2004.

Silvie Cinková, Pavel Pecina, Petr Podveský, and Pavel Schlesinger. Semi-automatic building of Swedish collocation lexicon. In *Proceedings of the 5th Conference on International Language Resources and Evaluation (LREC'2006)*, Genova, Italy, 2006.

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

Laurence Danlos. Support verb constructions: linguistic properties, representation, translation. *Journal of French Linguistic Studies*, 2(1):1–32, 1992.

Inês Duarte, Anabela Gonçalves, Matilde Miguel, Amália Mendes, Iris Hendrickx, Fátima Oliveira, Luís Filipe Cunha, Fátima Silva, and Purificação Silvano. Light verbs features in European Portuguese. In *Proceedings of the 2nd Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Pisa, Italy, 2010.

Cornelia Döll and Christine Hundt. *Funktionsverbgefüge im Portugiesischen: Komplexe sprachliche Einheiten zwischen Syntax und Lexikon*, pages 145–170. Valentia, Frankfurt am Main, Germany, 2002.

Peter Eisenberg. *Der Satz. Grundriss der deutschen Grammatik*. Bd. 2. Metzler Verlag, Stuttgart, Germany, 3rd edition, 2006.

ELRA. The ELRA newsletter. 1(4), December 1996. URL `http://www.elra.info/nl/newsletters/V1N4.pdf`. August 2011, date last accessed.

Stefan Evert and Brigitte Krenn. 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, 2001. Association for Computational Linguistics.

Christiane Fellbaum. *WordNet: an electronic lexical database*. MIT Press, Cambridge, Massachusetts, 1998.

Christiane Fellbaum, Alexander Geyken, Axel Herold, Fabian Koerner, and Gerald Neumann. Corpus-based studies of German idioms and light verbs. *International Journal of Lexicography*, 19(4):349–360, 2006.

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

Daniel Gildea and Martha Palmer. The necessity of parsing for predicate argument recognition. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 239–246, Philadelphia, Pennsylvania, 2002. Association for Computational Linguistics.

Gregory Grefenstette and Simone Teufel. 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, 1995.

Eberhard Gärtner. *Grammatik der portugiesischen Sprache*. Niemeyer Verlag, Tübingen, Germany, 1998.

Patrick Hanks, Anne Urbschat, and Elke Gehweiler. German light verb constructions in corpora and dictionaries. *International Journal of Lexicography*, 19(4): 439–457, December 2006.

Boris Haselbach. A morphosyntactic tagset, corpus processing, and lexical data extraction for Afrikaans. Master's thesis, Institut für Maschinelle Sprachverarbeitung, Stuttgart, Germany, 2010.

Iris Hendrickx, Amália Mendes, Sílvia Pereira, Anabela Gonçalves, and Inês Duarte. Complex predicates annotation in a corpus of Portuguese. In *Proceedings of the 4th Linguistic Annotation Workshop*, LAW IV '10, pages 100–108, Stroudsburg, Pennsylvania, 2010. Association for Computational Linguistics.

Wolfgang Herrlitz. *Funktionsverbgefüge vom Typ 'in Erfahrung bringen'. Ein Beitrag zur generativ-transformationellen Grammatik des Deutschen.* Max Niemeyer Verlag, Tübingen, Germany, 1973.

Munpyo Hong, Chang-Hyun Kim, and Sang-Kyu Park. Treating unknown light verb construction in Korean-to-English patent MT. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala, editors, *Advances in Natural Language Processing*, volume 4139 of *Lecture Notes in Computer Science*, pages 726–737. Springer Berlin/Heidelberg, 2006.

C. R. Johnson, C. J. Fillmore, M. R. L. Petruck, C. F. Baker, M. Ellsworth, J. Ruppenhofer, and E. J. Wood. FrameNet: theory and practice. October 2002. URL `ftp://ftp.icsi.berkeley.edu/pub/techreports/2002/tr-02-009.pdf`. August 2011, date last accessed.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing. An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics.* Prentice Hall, 2 edition, May 2008.

Alain Kamber. *Funktionsverbgefüge empirisch. Eine korpusbasierte Untersuchung zu den nominalen Prädikaten des Deutschen.* Max Niemeyer Verlag, Tübingen, Germany, 2008.

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

Philipp Koehn. *Statistical Machine Translation.* Cambridge University Press, Cambridge, Massachusetts, 2010.

Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL '03, pages 48–54, Stroudsburg, Pennsylvania, 2003. Association for Computational Linguistics.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. Edinburgh system description for the 2005

IWSLT speech translation evaluation. In *International Workshop on Spoken Language Translation (IWSLT)*, 2005.

Brigitte Krenn. *The usual suspects: data-oriented models for identification and representation of lexical collocations.* PhD thesis, Universität des Saarlandes, Saarbrücken, 1999.

Brigitte Krenn and Stefan Evert. 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, 2001.

Klaus Krippendorff. *Content Analysis: an introduction to methodology.* Sage Publications, Inc., Beverly Hills, California, 1980.

Jonas Kuhn. Parsing word-aligned parallel corpora in a grammar induction context. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05, pages 17–24, Stroudsburg, Pennsylvania, 2005. Association for Computational Linguistics.

Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

LDC. *Linguistic data annotation specification: assessment of fluency and adequacy in translations*, 2002. Revision 1.5.

Lothar Lemnitzer and Heike Zinsmeister. *Korpuslinguistik – Eine Einführung.* narr studienbücher. Gunter Narr Verlag, Tübingen, 2006.

Dekang Lin and Patrick Pantel. DIRT – discovery of inference rules from text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining(KDD-01)*, pages 323–328, 2001.

Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28:203–208, 1996.

Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing.* MIT Press, Cambridge, Massachusetts, 1999.

Begoña Villada Moirón and Jörg Tiedemann. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL '06 Workshop on Multiword Expressions in a Multilingual Context*, Stroudsburg, Pennsylvania, 2006. Association for Computational Linguistics.

Franz J. Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Franz J. Och, Christoph Tillmann, and Hermann Ney. Improved alignment models for statistical machine translation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP99)*, pages 20–28, University of Maryland, College Park, Maryland, 1999.

Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33:161–199, 2007.

Sebastian Padó and Mirella Lapata. Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research*, 36:307–340, 2009.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 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, 2010.

Darren Pearce. A comparative evaluation of collocation extraction techniques. In *3rd International Conference on Language Resources and Evaluation, LAS*, 2002.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. The necessity of syntactic parsing for semantic role labeling. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1117–1123, 2005.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. *Vorläufige Guidelines für das Tagging Deutscher Textcorpora mit STTS. Draft.* Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung / Universität Tübingen, Seminar für Sprachwissenschaft, 1995.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, United Kingdom, 1994.

Frank Smadja, Katheleen R. Mckeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22(1):1–38, 1996.

Frank A. Smadja and Kathleen R. McKeown. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics*, ACL '90, pages 252–259, Stroudsburg, Pennsylvania, 1990. Association for Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. The JRC-Acquis: a multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pages 2142–2147, Genova, Italy, 2006.

Jörg Tiedemann. News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume 5, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, 2009.

Kristina Toutanova and Christopher D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in Natural Language Processing and Very Large Corpora. Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, EMNLP '00, pages 63–70, Stroudsburg, Pennsylvania, 2000. Association for Computational Linguistics.

Peter D. Turney. The latent relation mapping engine: algorithm and experiments. *Journal of Artificial Intelligence Research (JAIR)*, 33:615–655, 2008.

Jeroen van Pottelberge. *Verbonominale Konstruktionen. Vom Sinn und Unsinn eines Untersuchungsgegenstandes.* Universitätsverlag Winter, Heidelberg, Germany, 2001.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *COLING '96: The 16th International Conference on Computational Linguistics*, pages 836–841, 1996.

Peter von Polenz. *Funktionsverben im heutigen Deutsch – Sprache in der rationalisierten Welt.* Schwann Verlag, Düsseldorf, Germany, 1963.

Joachim Wermter and Udo Hahn. Collocation extraction based on modifiability statistics. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, Pennsylvania, 2004. Association for Computational Linguistics.

Sina Zarrieß and Jonas Kuhn. 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, 2009. Association for Computational Linguistics.