Judith ECKLE-KOHLER, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart

Methods for quality assurance in semi-automatic lexicon acquisition from corpora

Abstract

This paper presents linguistics-based methods and engineering methods for quality assurance in semi-automatic acquisition of broad coverage lexicons from corpora. Automated linguistic tests are used to acquire candidates for particular subcategorization frames automatically; the regular use of metrics in the acquisition process contributes to a controlled development of these tests. The proposed methods are illustrated by the acquisition of a particular class of verbs taking *daß*-clauses in German, showing how the precision of the automatically acquired data can be maximized with only a slight decrease in recall.

Keywords: Quality assurance in lexicon acquisition, semi-automatic, corpus-based lexicon acquisition.

1. Introduction

1.1. Motivation

Quality assurance (QA in the following) has been recognized as an integral part of lexicon acquisition, not only in large-scale manual acquisition as in COMLEX (cf. Grishman et al. (1994)), but also in automatic acquisition from corpora as described in Briscoe, Carroll (1997) and Carroll, Rooth (1996). However, inaccuracies in a fully automatically acquired lexicon seem to be less problematic than those in a manually acquired lexicon, because the former provides relative frequencies of the different subcategorization frames of words, and these frequencies are exploited by probabilistic parsing systems: during parsing, inaccurate frames which are likely to have low frequencies are probably ranked out by accurate and more frequent frames. But since inaccuracies in the lexicon cause considerable problems to any purely symbolic parsing system, the use of fully automatically acquired lexicons is restricted. This calls for a semi-automatic approach to corpus-based lexicon acquisition, one which combines the advantages both of manual and fully automatic lexicon construction. These are on the one hand the construction of a lexicon of high quality and on the other hand the efficient construction of the lexicon.

The objectives of QA methods in semi-automatic acquisition then are firstly to maximize the quality – accuracy and completeness – of the automatically acquired data and secondly to minimize the effort needed by a lexicographer to assess and possibly subclassify these data. In our scenario, the lexicographer's task is merely to make yes/no-decisions about the accuracy of the automatically acquired data; thus high accuracy of these data would contribute immensely to a reduction of the necessary human effort. Therefore we decided to focus on high accuracy as opposed to completeness, trying to compensate for a possible lack of completeness by the use of very large corpora. In corpus-based acquisition, there is usually a trade-off between accuracy and completeness: often an increase in accuracy is accompanied by a decrease in completeness. Hence only a disciplined and systematic development of the

automatic acquisition procedures will result both in high accuracy and acceptable completeness.

1.2. Background

The methods for QA presented in the following are applied in the framework of the semi-automatic construction of a broad coverage syntactic lexicon of German to be used with a broad coverage Lexical Functional Grammar (LFG, cf. Dalrymple et al. (1995) and Kuhn, Rohrer (1997)) and possibly other lexicalist grammars. Our basic approach has been first to extract an initial lexicon from machine-readable dictionaries (from CELEX, cf. Baayen et al. (1995), and SADAW, cf. Dickmann et al. (1993)), and then to complement this lexicon by semi-automatically acquiring missing information types from corpora (for further details of the latter see also Eckle, Heid (1996a)). The information types provided by the lexicon for a given lemma are possible subcategorization frames which specify both the syntactic category and the syntactic function of the complements. Additionally, constraints on the usage of the lemma in certain syntactically defined contexts are given, for example information about mass nouns which may be used in the singular without article in German. Currently the lexicon contains information about 125 subcategorization frames (SC-frames) of 13328 verbs, 28 SC-frames of 8744 nouns and 36 SC-frames of 2274 adjectives. For 3673 verbs, 3974 nouns and 1894 adjectives, this information has been acquired semi-automatically from corpora.
We mainly use a 200 million word German news corpus which is tokenized (word and sentence boundaries), part-of-speech tagged with the STTS tagset and lemmatized. STTS stands for Stuttgart-Tübingen Tag Set; it is compatible with and trivially mappable onto the EAGLES morphosyntax specifications ELM-DE (cf. Teufel, Stöckert (1996)) and contains 54 tags with categorial, distributional and lexical distinctions (see Schiller et al. (1995)).
For the automatic extraction of lexicon data from corpora with this kind of annotations we use complex corpus queries which simulate partial chunk-parsing and which are processed by the CQP corpus query processor (see Christ, Schulze (1996)) which supports regular expressions of word forms and corpus annotations of any type.

2. The acquisition cycle

The following sections give an overview of the methods for QA which are applied in the above stated framework: these are both engineering methods, namely the use of a process model and the application of metrics, and linguistics-based methods. The effect of the QA methods is illustrated by an example of the semi-automatic acquisition of verbs taking a subject noun phrase and a *daß*-clause. For German, it is necessary to acquire these verbs from corpora, since existing dictionaries rarely contain details of possible clausal complements.

Our basic model of semi-automatic lexicon acquisition consists of two phases: firstly the automatic acquisition of raw material, see section 2.1, and secondly the human assessment of the raw material, see section 2.2. This basic process is repeated on the one hand for each lexical-syntactic property (e.g. for each possible subcategorization frame of a verb) and on the other hand for each modification of a particular automatic acquisition procedure. The former kind of iteration leads to a cyclic process and the latter to a spiral process. The spiral process in corpus-based lexicon building has been described in detail by Heid (1997); in our scenario, it results in maximum quality of a particular automatic acquisition procedure.

2.1. QA in automatic extraction of raw material

In the first phase of an acquisition cycle, lists of candidates for a certain lexical-syntactic property (LSP in the following) are automatically extracted from corpora. This requires the automatic identification of contexts which illustrate the chosen LSP. These contexts can be divided into two classes:
In one class, call it **A** for automatic, there are contexts which can correctly and reliably be recognized as illustrations of a LSP solely on the basis of the automatically accessible lexical knowledge, which comprises the available corpus annotations, as well as the results of partial chunk-parsing and already acquired lexical knowledge.
The correct identification of the contexts in the other class, call it **M** for manual, however, presupposes the existence of precisely that kind of lexical knowledge which is to be acquired. By automatic means, these contexts are correctly identified only by chance; a correct and reliable identification can be performed only by a human who has the necessary lexical knowledge.

Therefore our approach to maximize the accuracy of the automatically extracted candidate lists is to completely rely on contexts in class **A**, as far as these contexts exist. In other words, we use **automated linguistic tests** for LSPs.
We call the tests automated, because they are automatically carried out on corpus texts. An approach to automatically carry out linguistic tests on corpora has already been described by Hatzivassiloglou, McKeown (1995) for a particular class of adjectives in English.
Our automated linguistic tests are unambiguous patterns of surface realizations of predicate argument structures, very similar to the clause or sentence patterns used in traditional grammars to describe verb complementation (see for example Engel (1988) and for a discussion Helbig (1982)).
The patterns are unambiguous with respect to their predicate argument structure; this can be characterized as follows:
First, the predicate itself is unambiguously identifiable; this requires the presence of exactly one predicate of the type in question, thus excluding for example coordination of predicates.
Secondly, both the predicate's arguments and possible adjuncts have to be constrained in a way that makes their automatic identification possible. Obviously there are cases, where such constraints are not applicable, as for example in the case of verbs taking prepositional objects, where a surface cue for PPs functioning as arguments as opposed to adjuncts doesn't seem to exist.

To illustrate our understanding of automated linguistic tests, consider the example of verbs taking a subject noun phrase and a *daß*-clause. Table 1 shows an automated linguistic test for this class of verbs, specifying contexts of class **A**. The first column contains a specification of the unambiguous pattern: this is a sequence of morpho-syntactic chunks with brackets indicating optionality and the numbers in square brackets referring to constraints imposed on these chunks. In the second column, you find a short description of the chunks, including a description of the constraints which are necessary for the automatic identification of the predicate argument structure in question. The third column gives an example clause matching this pattern. Contexts which are covered by this pattern are verb-last clauses matching the specified sequence of chunks; since this covers only a part of the relevant verb-last clauses, there are a number of variants of this pattern accounting for the possible permutations of syntactic constituents in the German 'Mittelfeld'. While the pattern in table 1 matches only

clauses containing a noun phrase whose head is a common noun, there is also a variant of this pattern specifying a pronoun as subject.

Pattern	Description	Example
SubordConj	subordinating conjunction	*weil*
NC[1,2]	noun chunk: determiner as well as modifying adjectives and adverbs are optional [1]: no modifying adverb or adjective which can function as a correlative of an adverbial *daß*-clause [2]: noun does not subcategorize for a *daß*-clause	*der Besitzer*
(NCgen[3])	optional genitive noun chunk [3]: case of noun chunk is unambiguously genitive	*des Hauses*
(PC[4])	optional prepositional chunk: preposition followed by a noun chunk [4]: preposition is not subcategorized for by verbs	*trotz schwerer bedenken*
(AdvC[5,6])	optional adverb chunk [5]: no adverb which can function as a correlative of a *daß*-clause indicating its function as a prepositional object	*jetzt endlich*
VCactive[7]	verb chunk in the active voice, containing one lexical verb, including all possible constructions with auxiliaries and modals [7]: forms where the past tense is formed with the auxiliary *sein* must not occur	*zugegeben hat*
daß		*daß*

Table 1: Automated linguistic test for the identification of verbs taking a subject noun phrase and a dass-clause in verb-last clauses.

The constraints contribute to the automatic induction of the predicate argument structure in the following way: [3] is a cue for a syntactic structure where NC[1,2] and NCgen[3] form a single noun phrase as opposed to two noun phrases (genitive complements occur very rarely in newspaper texts). As we will see below, constraint [7] ensures that this noun phrase is the subject of the clause as opposed to its object. Both [1] and [5] are cues for a syntactic structure where the *daß*-clause functions as an argument of some predicate as opposed to an adjunct. [6] ensures that a construction with a pronominal adverb functioning as a correlative of the *daß*-clause doesn't occur, because constructions of this type are treated separately. [2] is a cue for a syntactic construction where the *daß*-clause is an argument of the verb, not of the head of the subject noun phrase, and [4] guarantees that the PC functions as an adjunct, not as an argument of the verb. Thus we end up with an illustration of the SC-frame we are looking for.

To see why the constraints are necessary for automatic identification, consider in more detail, what would happen if we would raise constraint [2]. Raising constraint [2] means to allow contexts where the head noun of the subject noun phrase subcategorizes for a *daß*-clause. An example of such a context matching an analoguous pattern as that in table 1 would be: *...weil die rechtzeitige Ankündigung sicherstellt, daß der Termin eingehalten wird.* Although this context is an illustration of the SC-frame we are looking for – the *daß*-clause depends on the verb *sicherstellen* –, this can reliably be determined only manually, not automatically: if we would try to extract this kind of context automatically, we would also get contexts *like ...weil*

der Eindruck besteht, daß der Termin nicht eingehalten werden kann, where the *daß*-clause is an argument of the noun *Eindruck*, since these contexts match exactly the same surface pattern. In other words, we would only get members of class M, whose correct and automatic identification would be pure chance.

Another example is to illustrate constraint [5]. By raising constraint [5], we would get contexts like *...weil er schon so oft geglaubt hat, daß er im Lotto gewonnen hat.* In this clause, *so* just modifies the adverb *oft*, i.e. the clause illustrates the SC-frame we are looking for. But we would also get contexts with the same surface pattern, where the *so* functions as a correlative of an adverbial *daß*-clause of result or outcome, like for example *...weil der Konflikt so schnell eskaliert, daß eine Lösung nicht möglich scheint.*
A last example is to be given for constraint [7]. [7] is necessary, because for German, there is no reliable information about auxiliary selection. If we would raise constraint [7], we would get contexts like *Sie sind übereingekommen, daß der Vertrag unterzeichnet wird*, which is a clause in the past tense and active voice, given the information that *übereingekommen* selects *sein* as auxiliary. But we would also get clauses like *Sie sind benachrichtigt, daß er sein Ziel erreicht*, which can be identified as a clause with a quasi-passive (called 'Zustandspassiv' in German) only by using the information that *benachrichtigen* selects *haben* as auxiliary.

With respect to completeness of the extraction results, the automated linguistic tests for each LSP have to be carefully designed in order not to systematically miss certain classes of lexemes. For example, verbs taking prepositional objects can be identified in sentences such as *Er rechnet damit, daß es regnet*, where the correlative of the *daß*-clause, here *damit*, indicates its function as a prepositional object. However, when the prepositional object refers to a human object, this construction is not possible. Therefore verbs which are restricted to human objects, such as *sprechen* in *Er spricht mit seinen Kollegen*, would be systematically missed, if we would use no other test.

2.1.1. Modular query templates

Each automated linguistic test is implemented by means of a complex corpus query which simulates a kind of partial chunk-parsing. The corpus queries contain macros for search patterns taken from a library of reusable search patterns. These macros correspond to the morpho-syntactic chunks in the automated linguistic tests. The macroprocessor for the CQP language, MP (see Schulze (1996)), expands the macros, before the query is processed by CQP.
Most of the constraints present in the automated tests can be put into the implementation of the macros, i.e. are implemented by means of corpus query. But as we will see in section 2.1.2, there are constraints which can more conveniently be implemented by an automatic filter which is used to postprocess the results of a corpus query.
The main results of a corpus query are first frequency distributions of candidates and, where useful, frequency distributions of candidates along with context partners. Secondly we obtain 'subcorpora' containing all sentences which have been matched by the query. The subcorpora are used to automatically extract sample sentences for the candidates.

2.1.2. Automatic linguistic filters

We use automatic linguistic filters to implement constraints which make use of already acquired lexical knowledge which is not annotated in the corpora. First there is a very simple morphological filter used to get rid of misspelled candidates and candidates which are wrongly tagged in the corpus; this filter checks the candidates against a morphology system.

Secondly there is a lexical-syntactic filter which checks the candidates or context partners in a frequency distribution against the already acquired syntactic lexicon. In our example, the lexical-syntactic filter is of particular importance, because it implements constraint [2] by using knowledge about nouns taking *daß*-clauses, which has been acquired from our news corpus in a previous acquisition cycle: The filter eliminates all those verb candidates whose context partner, being defined as the head of the subject noun phrase, takes a *daß*-clause. It is essential to note, that this knowledge about nouns can be acquired from class A contexts, where the *daß*-clause is unambiguously an argument of some noun. These are contexts of the *type Die Ankündigung, daß der Vortrag stattfindet, bewirkt* ..., where the noun and the *daß*-clause take the 'Vorfeld'-position of a verb-second sentence, a sentence position in German which is restricted to contain only one syntactic constituent. Although the thus acquired list of nouns might be incomplete, the use of this knowledge improves the accuracy of the verb candidate list in our example significantly, as can be seen in table 2.

2.2. QA in human assessment of raw material

2.2.1. Lexicon data with example sentences

In the second phase of an acquisition cycle, the automatically extracted and filtered candidate lists are assessed by a lexicographer. The candidate lists contain both true positives (TP), candidates correctly proposed by the automatic acquisition procedures, and false positives (FP), candidates incorrectly proposed. The lexicographer's task is to identify the FPs; this has to be done manually, because a fully accurate and complete reference lexicon, at least for German, does not exist (cf. Briscoe, Carroll (1997) who note this problem for English, too). The identification of the FPs serves two purposes: first, the quality of the automatic acquisition procedures can be measured, see section 2.2.2, and secondly, a preliminary lexicon version with a presumably small number of errors can be automatically generated.
To minimize the lexicographer's effort, we take the following steps: First, candidates present in the already acquired lexicon are automatically marked, as well as candidates which have been assessed as FPs in a previous part of the acquisition spiral. Then the remaining unmarked candidates are presented to the lexicographer via a WWW-interface where he can mark the FPs and look at the sample sentences which are dynamically extracted from the subcorpora on demand. The sample sentences especially simplify the identification of the FPs, since the lexicographer might not be sure, if a candidate should be marked as FP, or if he can't think of an example for the suggested LSP for that candidate. By looking at the sample sentences for FPs, the lexicographer also recognizes possible improvements of the automated tests, as well as automatic misclassifications which can not be avoided on the basis of the automatically accessible lexical knowledge. All those observations are documented in order to provide feedback to the designer of the automated linguistic tests. Equally important is the documentation of the principles or intuition underlying the assessment decisions. Such guidelines in combination with the sample sentences are most valuable for lexicon maintenance, because they help a human user of the lexicon to understand and possibly correct certain details which seem implausible to him.

In our example, FPs are mainly due to series of *daß*-clauses, e.g. as *in Die Tatsache, daß die Sonne scheint, daß die Vögel zwitschern, daß sogar einige Blumen blühen, wundert ihn.* In these cases, the test pattern matches a *daß*-clause which functions as an argument of some predicate itself, here the noun *Tatsache*. This problem could be avoided, if a similar

automated test for verb-second sentences would be used; as shown in table 2 in section 2.2.2, this in fact improves the accuracy of the candidate list considerably.

2.2.2. Metrics for quality control

After the human assessment step has been concluded, some metrics are automatically computed, based on the frequency counts of the manually identified true and false positives. The resulting measures provide feedback to the designer of the automated tests and therefore serve as a means of quality control.

We use one precision measure as indicator of accuracy and two recall measures as indicators of completeness, recall being defined by Salton, McGill (1983) as the ability of an extraction system to retrieve all relevant items, and precision being defined as the ability to retrieve only the relevant items. As a precision measure, we use the percentage of TPs to all automatically acquired candidates. To get a lower bound of the recall, we compute the ratio of the frequency of the different TPs (as given in the frequency distribution) to the overall frequency of the TP lemmas in the corpus (see Eckle, Heid (1996b)); besides that, we simply count how many TPs have automatically been acquired. Since it is important for quality control to measure the recall regularly, we prefer these approximate recall measures to a true recall measure, because the calculation of the latter requires the manual evaluation of large amounts of corpus data, which takes a lot of effort and is therefore typically done rarely and only for a small number of candidates.

Table 2 shows two of the measures we obtained for the verb candidate list in our example. We compared the measure results achieved by using all the constraints specified in table 1 with those obtained by systematically raising a single type of constraint. Thus the effect of the four constraints [1], [2], [5] and [7] becomes obvious: by using all constraints, we achieved an increase in precision of 23%, while the number of TPs is reduced only by one. Obviously the low precision value of 75% is mainly due to series of *daß*-clauses (see section 2.2.1), since the use of a similar test for verb-second sentences leads to an acceptable precision value of 94%, while the number of TPs is reduced only by three.

	Precision	Number of Tps
acquisition from verb-last clauses:		
all constraints, elimination of misspellings	75%	215
without constraints [7]	65%	216
without constraints [2]	65%	215
without constraints [1] and [5]	69%	215
without constraints [1], [2], [5] and [7]	52%	216
acquisition from verb-second sentences:		
all constraints, elimination of misspellings:	94%	213

Table 2: Precision and number of TPs obtained for the verb candidate list automatically acquired by the methods described in section 2.1 from a 200 million word news corpus.

3. Conclusion

The use of automated linguistic tests in semi-automatic, corpus-based lexicon acquisition as a means of automatically achieving lexicon data of high quality has been shown by an example of verbs taking *daß*-clauses. The quality of the automated linguistic tests themselves is maximized in a controlled way by the regular use of metrics during the acquisition process. The measures provided by the metrics indicate the accuracy and completeness of the automatically acquired lexicon data. In our example, we have shown how the precision of the automatic acquisistion procedures could be maximized, while the number of TPs was reduced only by three. The lexicographer's assessment and classification task is supported by automatically extracted sample sentences which are presented to him via a WWW-interface. The results of his assessment work provide the basis both of quality control and of the automatic generation of a lexicon with sample sentences, useful for linguists as well as for symbolic NLP-systems.

4. References

Baayen, R. H., R. Piepenbrock and L. Gulikers (1995). *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Briscoe, Ted and John Carroll (1997). Automatic Extraction of Subcategorization from Corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP-97)*, Washington, DC, USA.

Carroll, Glenn and Mats Rooth (1996). Valence induction with a head-lexicalized PCFG. ms., Stuttgart, IMS; http://www.ims.uni-stuttgart.de/~mats.

Christ, Oliver and Bruno M. Schulze (1996). Ein flexibles und modulares Anfragesystem für Textcorpora. In: Helmut Feldweg, Erhard W. Hinrichs (eds.): *Lexikon und Text*, Lexicographica: Series maior; 73; Niemeyer, Tübingen.

Dalrymple, M., R.M. Kaplan, J.T. Maxwell III and A. Zaenen (1995). *Formal Issues in Lexical-Functional Grammar*, CSLI Publications, Stanford, CA.

Dickmann, Ludwig, Julia Heine, Judith Klein, Fred Oberhauser, Hannes Pirker and Julia Simon (1993). Abschlussbericht der Arbeitsgruppe zur Bewertung und Kodierung des SADAW-Lexikons, ms., Saarbrücken.

Eckle, Judith and Ulrich Heid (1996a). Extracting raw material for a German subcategorization lexicon from newspaper text. In *Proceedings of the 4th International Conference on Computational Lexicography, COMPLEX '96*, Budapest.

Eckle, Judith and Ulrich Heid (1996b). Creating verb frequency lists for German, technical report, Stuttgart, IMS.

Engel, Ulrich (1988). *Deutsche Grammatik*, Julius Groos Verlag, Heidelberg.

Grishman, Ralph, Catherine Macleod and Adam Meyers (1994). Comlex Syntax: Building a Computational Lexicon. In *Proceedings of COLING '94*, Kyoto.

Hatzivassiloglou, Vasileios and Kathleen McKeown (1995). A Quantitative Evaluation of Linguistic Tests for the Automatic Prediction of Semantic Markedness. In *Proceedings of ACL '95*, Cambridge, Massachusetts.

Heid, Ulrich (1997). *Zur Strukturierung von einsprachigen und kontrastiven elektronischen Wörterbüchern*, Lexicographica: Series maior; 77; Niemeyer, Tübingen.

Helbig, Gerhard (1982). *Valenz - Satzglieder - semantische Kasus - Satzmodelle*, VEB Verlag Enzyklopädie Lepzig.

Kuhn, Jonas and Christian Rohrer (1997). Approaching ambiguity in real-life sentences - the application of an optimality theory-inspired constraint ranking in a large-scale lfg grammar. In *Proceedings of DGfS-CL*, Heidelberg 1997.

Salton, Gerard and Michael J. McGill (1983). *Introduction to Modern Information Retrieval, McGraw-Hill*, New York.

Schiller, Anne, Simone Teufel, Christine Stöckert and Christine Thielen (1995). Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS, ms., Stuttgart/Tübingen.

Schulze, Bruno Maximilian (1996). MP user manual, ms., Stuttgart, IMS.

Teufel, Simone and Christine Stöckert (1996). EAGLES specifications for German morpho-syntax, technical report, Stuttgart.