>>> revista tradumàtica
Traducció i Tecnologies de la Informació i la Comunicació
Número 06: Terminologia i traducció : desembre 2008 : http://www.fti.uab.cat/tradumatica/revista :
ISSN: 1578-7559

AUTOTERM: Term Candidate Extraction for Technical Documentation (Spanish/German)

Johann Haller, Universität des Saarlandes, Saarbrücken

Resum: En aquest article es presenten les característiques i les funcions més rellevants d'AUTOTERM, una eina híbrida d'extracció de terminologia desenvolupada pel IAI de Saarbrücken. Una anàlisi bàsica de textos en castellà i alemany és la base de partida per a una cerca lingüística de patrons i un rànquing estadístic de probables candidats a terme, tan de manera monolingüe com bilingüe. Els exemples han estat extrets de manuals de reparació d'automòbils en les seves versions castellana i alemanya.
Paraules clau: Terminologia, extracció terminological, apropament hybrid, anàlisi linguistico i estadística.

Resumen: En este artículo se presentan las características y las funciones más relevantes de AUTOTERM, una herramienta híbrida de extracción de terminología desarrollada por el IAI de Saarbrücken. Un análisis básico de textos en castellano i alemán es la base de partida para una búsqueda lingüística de patrones y un ranquin estadístico de probables candidatos a término, tnato de manera monolingüe como bilingüe. Los ejemplos han sido extraídos de manuales de reparación de automóviles en sus versiones castellana y alemana.
Palabras clave: Terminología, extracción terminológica, acercamiento híbrido, análisis lingüístico y estadístico.

Abstract: In this article, the most important steps and possibilities of AUTOTERM (a hybrid term extraction tool developed by IAI Saarbrücken) will be presented. A basic analysis of Spanish and German texts is the basis for a linguistic pattern search and a statistical ranking of probable term candidates, both mono- and bilingually. The examples come from car reparation manuals in their Spanish and German versions.
Key words: Terminology, term extraction, hybrid approach, linguistic and statistic analysis.

1. Terminology is important

Terminology is of utmost importance for many activities of economic interest. It is an unavoidable factor in scientific communication, international cooperation and legal regulation. Without a sound terminological basis, it is impossible to generate multilingual documentation in a correct and economically viable manner.

All attempts to rationalise this process, such as controlled language, translation memories and machine translation, have to be based on a correct and complete terminological data base.

Today, several tools are on the market which try to assist the technical author in his task, showing him possible problems in orthography, grammar, consistent terminology, abbreviations and stylistic deviations.

The IAI (Institute for Applied Information Research at the University of the Saarland) has carried out a series of projects (MULTILINT, MULTIDOC, TETRIS), which are described in detail on the web pages (iai.uni-sb.de).
From these projects originates the CLAT product family ("Controlled Authoring Tools") which checks, for German and English texts, the above mentioned features in technical literature. First and foremost, the German car manufacturers use these tools for their manuals in order to rationalize the process of multilingual text production. Other industrial companies have also joined the user group.
In a few cases, CLAT clients already have terminology collections at their disposal or are able to obtain them from translation memories. The interest for storing terminology in translation memories, however, consists mainly in collecting a maximum of variations for a term – in order to detect the term in a text and to provide a translation even for a misspelled variation.
In this case, IAI and the client use software tools for automatically detecting misspellings and variations in the lists, often manually compiled by students. Using these tools, the terminologists can set up new data bases faster and in a more reliable way.

In many cases, there are big (mono- and bilingual) corpora but still no terminology.

This is the point where term candidate extraction tools play a role, such as the AUTOTERM tool, developed by IAI, which is described in the following section.

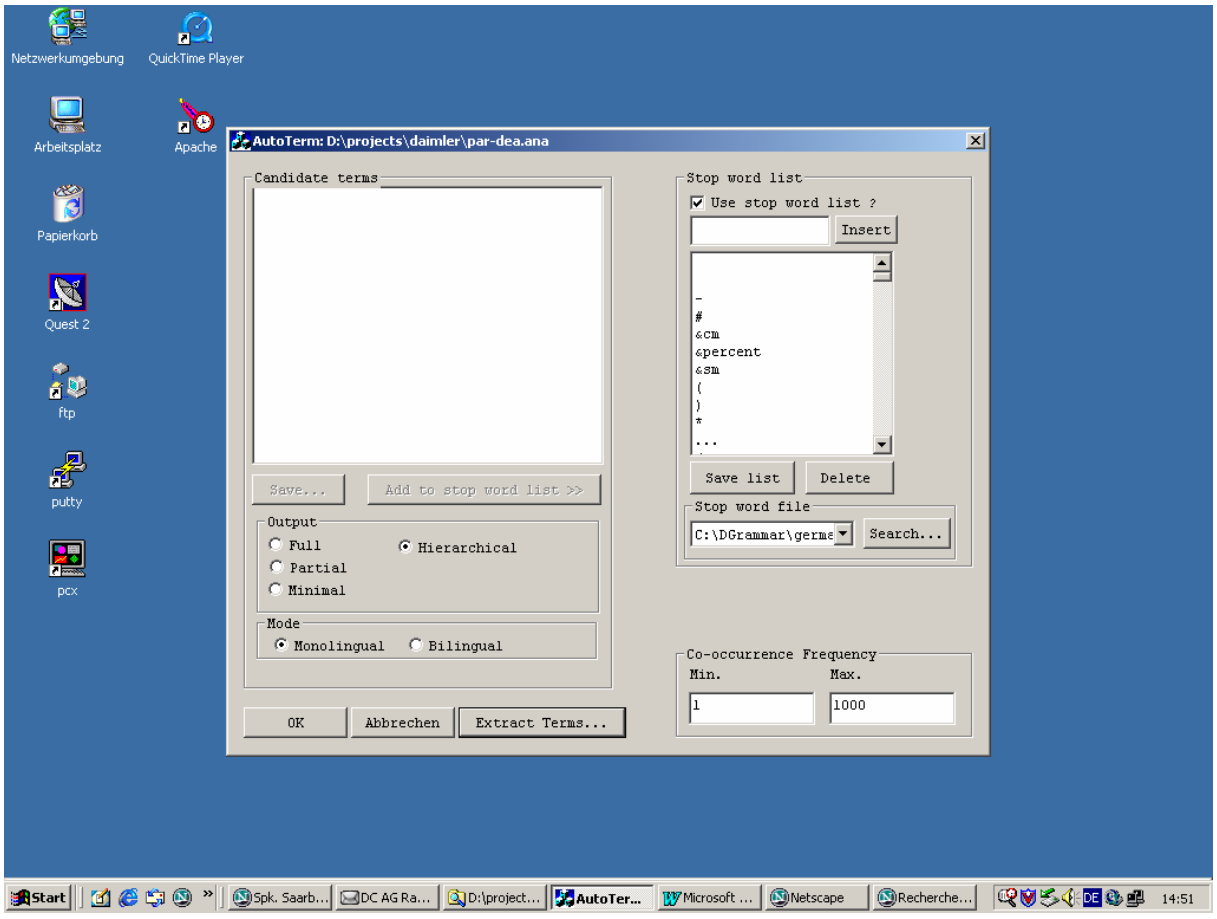
2. Monolingual Extraction

The general principles of term extraction, the main approaches and the basic principles of AUTOTERM are described, for example, in HONG (2001) and in BOURIGAULT (2001).

We start here with the hypothesis that linguistically analysed texts are better suited for such a term extraction than pure character-oriented methods like, for example, EXTRATERM from older TRADOS/SDL versions. It is not very clear if the recently launched SDL finder uses linguistic procedures (although it is said to function only for a restricted set of languages) - some descriptions mention this but practical experience does not confirm this hypothesis.
Experiments with different languages and different text volumes were carried out with AUTOTERM within seminar reports and diploma theses, and they confirm the efficiency of so-called hybrid methods.

It is clear that the number of languages which can be treated with such a tool is restricted because high quality linguistic analysis tools are in general available only for major languages in. One has to decide if the higher effort to be invested into these analyses can be justified in each specific case.

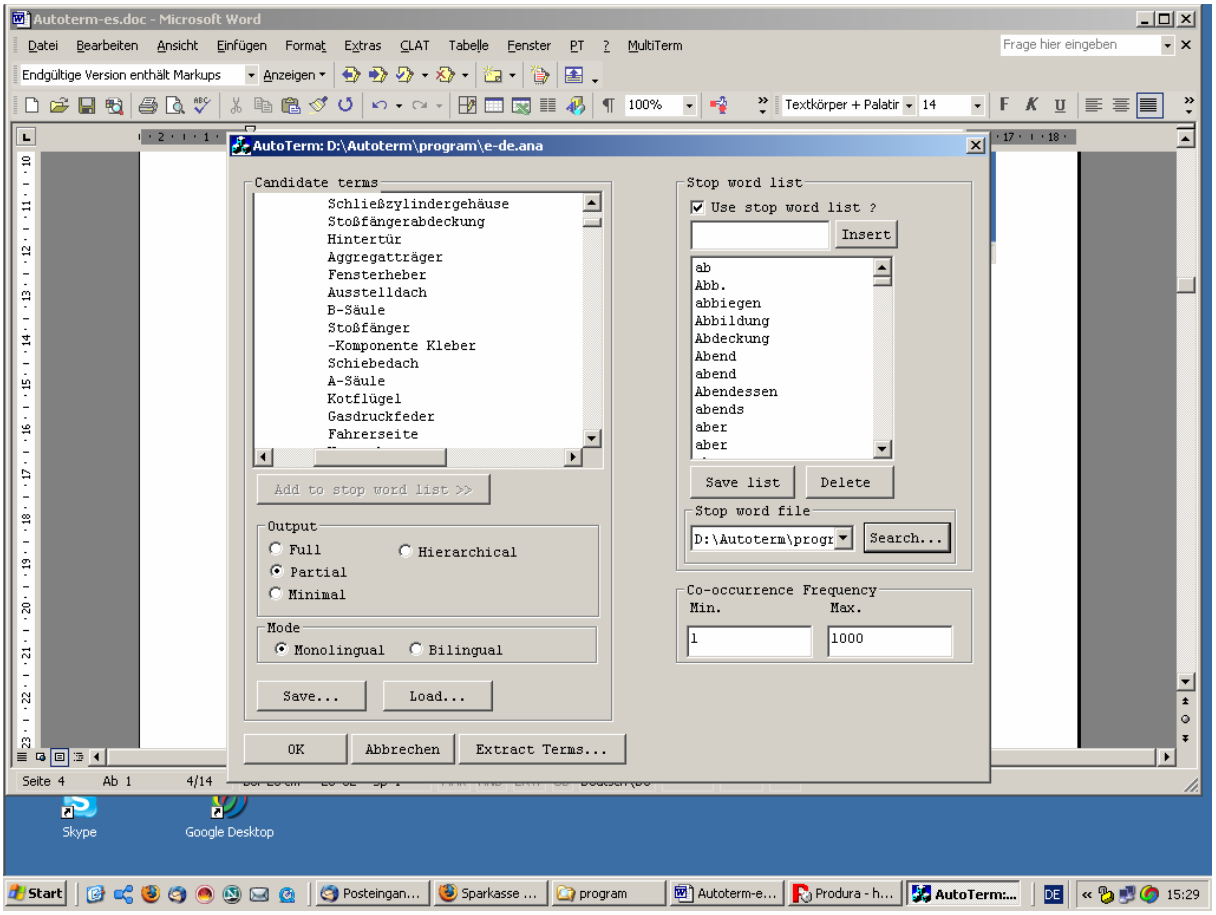
In the following section, the most important steps and possibilities of AUTOTERM will be presented; a basic analysis of Spanish and German texts as described in HONG (2001) is in any case necessary and will not be detailed herein. For the mono- and bilingual analysis, we will use extracts from car reparation manuals in their Spanish and German versions.



1: Initial AUTOTERM dialogue

There is a basic choice between mono- and bilingual operation mode. Moreover, the user can link a stop word list and regulate the maximal and minimal frequency of a term candidate.

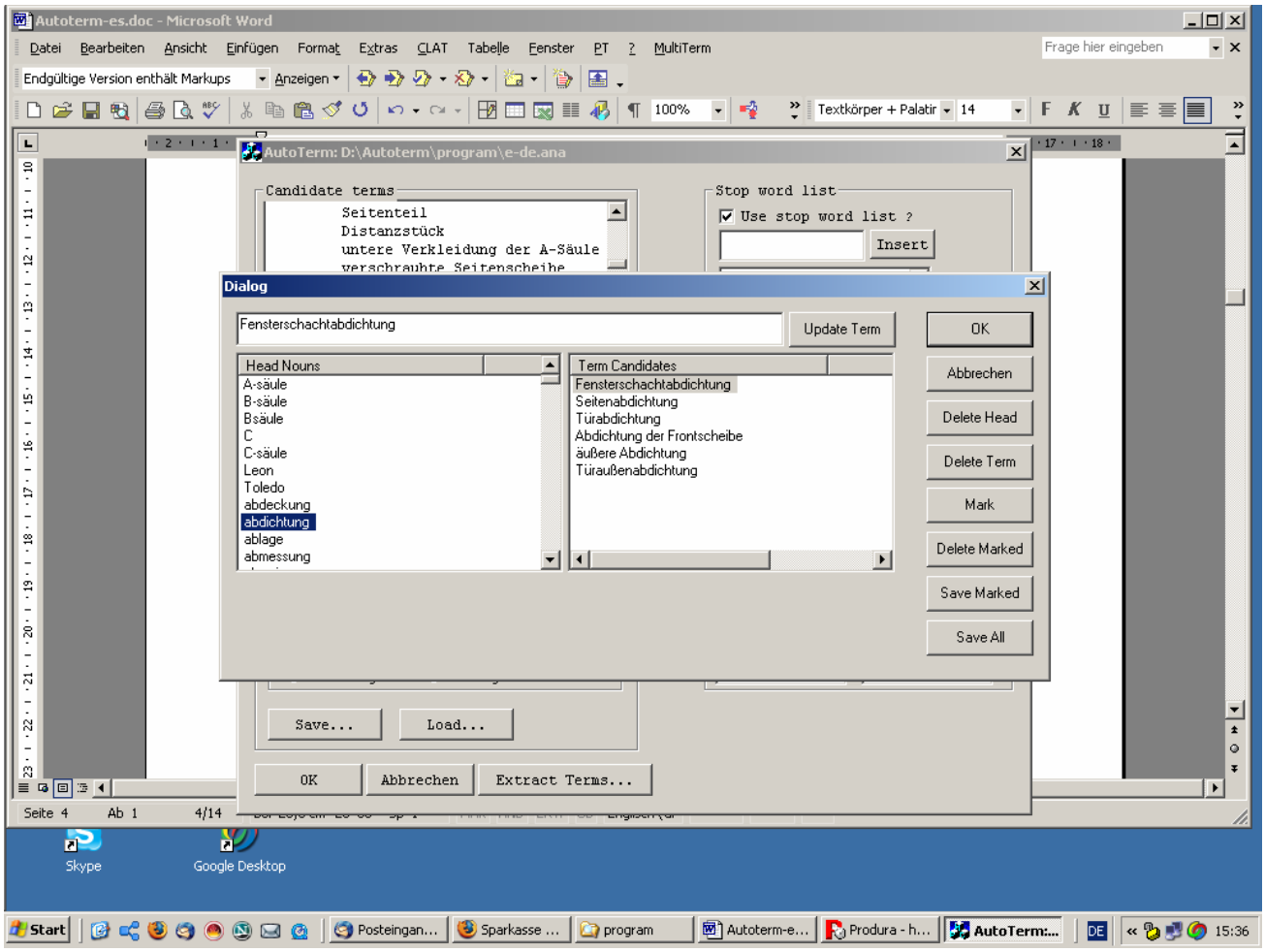
Different presentation versions (minimal, partial, maximal, hierarchical) are offered as options too. The user can sort the candidates alphabetically (minimal) or by the coherency value of the candidate (partial).
If the user chooses he can see also the frequency numbers which are the basic data for the coherency value ("likelihood") calculation (maximal).



2: Term candidates, sorted by ‘likelihood’

This format (which can easily be uploaded to, for example, a MULTITERM data base), presents the most valued candidates at the top of the list, in order to facilitate the work of the terminologist.

Another option is the hierarchical sort according to the head nouns: this is very useful for the immediate detection and elimination of variants.



3: Sorting according to Head Nouns

3. Bilingual Extraction

Using tmx files as input, AUTOTERM deals also with bilingual material. This will be illustrated with the Spanish text and its German equivalent.

In the annex, a sample of this tmx file can be found.

The first part of AUTOTERM consists of perl scripts which extract the two texts from the tmx file and initiate the respective linguistic analysis of the texts.

The linguistic procedures are the same as in the CLAT system for spelling, grammar and terminology checking as described in http://www.iai.uni-sb.de/CLAT/de/clat_home.htm; with this analysis, the text is additionally split into sentences and words. Words are morphologically analysed, tagged and lemmatized. There is an additional parsing module for grouping words into nominal, verbal and adverbial groups. A short description of the linguistic analysis can be found in MAAS (1998).

In order to illustrate the results of the linguistic analysis, a simplified example is presented here:

((Desenroscar (los tornillo-s)) y (extraer (el ((tubo flexible) de (conducción de aire))).

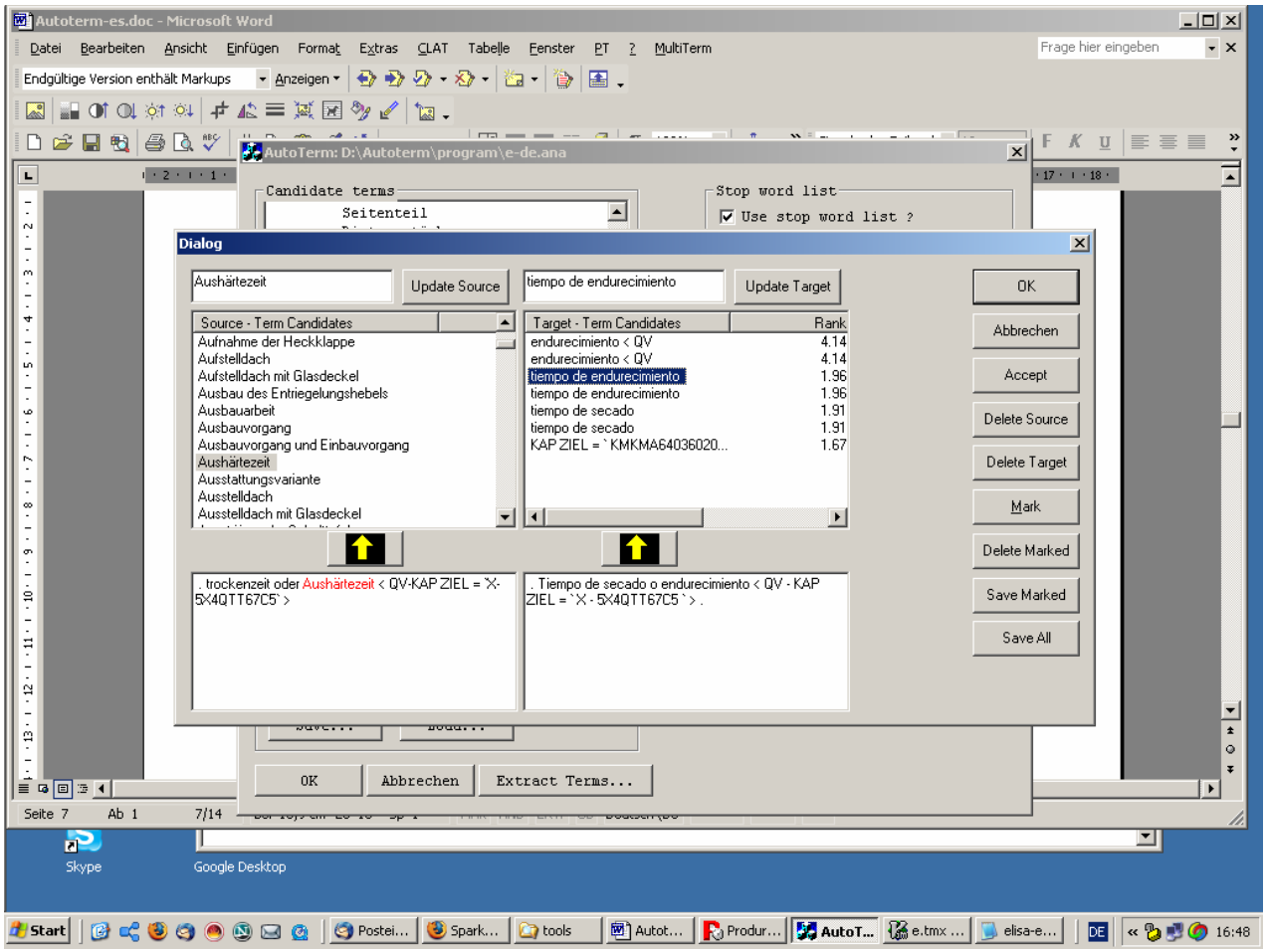
All information from the morphological analysis is available too, as for example the information about the plural form of ‚tornillos', the syntactic information about 'extraer V' etc.

AUTOTERM now uses (as in the monolingual extraction) a list of language specific syntactic patterns which are typical for term candidates (prepositional phrases for Spanish, compounds for German etc.). If these patterns have been identified and their frequencies counted, a special formula for 'likelihood' is applied: the more often the elements of such a pattern occur together (and the less often they occur together with other partners), the higher is the probability that they are a good term candidate. This value is calculated only on the basis of the text itself and does not currently use a 'general language model'. In the future, a comparison of these frequencies between the general model and the specific text group may lead to an even more precise calculation of this value.

In this process, lists of term candidates are created for the two languages. Co-occurrence in many sentences, in corresponding syntactic positions and syntactic correspondence rules (Spanish prepositional phrase can correspond to a German PP or compound etc.) are additional weighting factors for ranking the parallel candidates.

Sorted according to one of the two languages (here German), the candidates are then presented to the user who can confirm or delete the source entry and choose the right target entry from a ranked list.
In many cases, the corresponding target entry is in the first place.

In the following example, the user has to press Accept in order to confirm the translation of tiempo de endurecimiento into ‚Aushärtezeit'. He has the possibility to delete in a second step all other proposed equivalents by pressing the corresponding button “Delete all other possibilities”.



4. Bilingual term candidates

After accepting this proposal, he can also cut all other links backwards and reduce gradually the number of proposals in the following entries.

Consequently, the user sees in most cases only one correspondence proposal and can save a considerable amount of time, just confirming this proposal, arriving in a very short time at a bilingual term list.
In case of a lower quality of the proposals, the user can modify source and target entries and manually cut alternatives out of the sample sentences (which he sees below).

Another working mode is the ‚Marking' method if the number of correspondences seems to be rather low – the user can then only save the marked pairs together with the context.

The two main features are the lower number of 'noise proposals' and the fact that the terms are proposed in the basic form (and not in the inflected form as they occur in the text) and can immediately imported into the (MULTITERM or other) data base.

The AUTOTERM tool is being continuously tested and augmented; currently, it is available for the following languages:
German, English, French, Spanish, Italian, Portuguese. A few tests have also been carried out for other languages such as Russian and Lithuanian.

However, the tool can easily be created for other languages for which the IAI has linguistic resources (http://www.iai.uni-sb.de/de/res-mono.html), which can be updated and used for AUTOTERM in other application scenarios.

This article is an updated and enlarged version of (HALLER 2007).

Annex 1:

TMX file Spanish/German (extract)

```
<tu creationdate="20070620T122553Z" >
<tuv lang="ES-ES">
<seg>Manivela de la ventanilla</seg>
</tuv>
<tuv lang="DE-DE">
<seg>Scheibenkurbel</seg>
</tuv>
```

</tu>

<tu creationdate="20070620T122555Z" creationid">
<tuv lang="ES-ES">
<seg>Junta puerta exterior para regiones polvorientas</seg>
</tuv>
<tuv lang="DE-DE">
<seg>Türaußenabdichtung für staubige Gebiete</seg>
</tuv>
</tu>

Annex 2: Term candidates with ‚likely hood' (sample)

Monolingual German

55.55	Türfensterscheibe
55.08	Schlossträger
54.29	Abmessung Aufnahme
54.05	Tankklappeneinheit
54.05	Sprühdüse
54.05	Anzugsdrehmoment
49.50	Sicherungsknopf
47.97	mittlere Geräuschdämmung
47.97	Fußraum
47.15	Gitter der Stoßfängerabdeckung
46.69	Zentralsteuergerät
46.69	Scheibenwaschanlage
46.42	Vordertür Fahrerseite
46.02	einstellbarer Rückspiegel
46.02	Zusammenbau Schlosstaste
46.02	Sperrriegel des Lagerbügels
46.02	Sperrriegel
46.02	Sonnenblende
46.02	Lackschaden
46.02	Funkfernbedienung
45.60	seitliches Gitter

45.08	Scharnier der Heckklappe
44.50	Säulenverkleidung
44.50	Gepäckraumverkleidung
44.50	Dachverkleidung

Monolingual Spanish

147.27	portón posterior
132.75	revestimiento de paragolpes
131.73	puerta anterior
115.81	caja de agua
103.34	rejilla bajoparabrisas

100.37	puerta posterior
100.28	faro antiniebla
99.76	alzacristal eléctrico
92.09	lado izquierdo
89.20	operación de desmontaje
86.47	cristal de la puerta
85.02	techo abrible
83.39	lado derecho
77.66	puerta delantera
76.34	tablero de instrumento
74.54	unidad de la tapa
73.49	lugar de montaje
70.67	operación correspondiente
66.94	tapa de la caja
66.20	lado de conductor

Annex 3:

Bilingual terms (manually confirmed in less than 1 minute). The number at the end of the line corresponds to the sentence number of the bilingual context sentences.

Antriebsmotor	0.0	Motor de accionamiento<*>872
Anzugsdrehmoment	3.31	apriete de las tuercas<*>639
Ausbauarbeit	1.75	operación de desmontaje<*>642

Aushärtezeit	1.96	tiempo de endurecimiento<*>982
Außenspiegelgehäuse	4.94	carcasa de espejo<*>461
Außentemperatur	0.0	temperatura exterior<*>19
Befestigungsmutter	3.64	tuerca de fijación<*>134
Befestigungsmutter des Scharniers	4.44	tuerca de la bisagra<*>651
Befestigungsschraube	2.96	tornillo de sujeción<*>871
Beifahrerseite	2.88	lado acompañante<*>718

Bibliography:

Bourigault, Didier/ Christian Jacquemin/ Marie-Claude L'Homme, (Hrsg.) (2001): "Recent Advances in Computional Terminology" . In: Natural Language Processing, Volume 2. Amsterdam, Philadelphia

Haller, Johann (2007): „Autoterm: Automatische Terminologieextraktion Spanisch/Deutsch". In: AlbertoGil / Ursula Wienen (Hrsg.): Multiperspektivische Fragestellungen der Translation in der Romania. Frankfurt am Main

Hong, Munpyo/ Sisay Fissaha/Johann Haller (2001): "Hybrid Filtering for Extraction of Term Candidates from German Technical Texts", Proceedings of TIA'2001, Nancy – zu finden unter http://www.iai.uni-sb.de/de/pub.html]

Maas, Heinz Dieter (1998): „Multilinguale Textverarbeitung mit MPRO", in: Europäische Kommunikationskybernetik heute und morgen'98, Paderborn

Diploma Thesis:

* Lorenzini, Nathalie: Methoden der automatischen Termextraktion und -analyse: Aufbau und Organisation einer Terminologie im Bereich Wirtschaft für das Unternehmen SAP
* Schenk, Carsten : Neue Ansätze zur Terminologieverwaltung am Beispiel der Trados-Tools
* Kühn, Lina: Terminologie Völker/Europarecht Deutsch/Spanisch (Experimente mit automatischer Termextraktion)
* Wobken, Renate: Talentproblematik im Sport-Terminologie Französisch-Deutsch
* Donoso, Alexis: Aufbau einer Terminologie-Datenbank für eine mehrsprachige Pressemitteilung der Fa. Daimler-Chrysler

* Zielinski, Daniel: Computergestützte Termextraktion aus technischen Texten (Italienisch)

*Guerreiro, Carla: Bilinguale automatische Termextraktion Portugiesisch-Deutsch