



Europäisches
Patentamt
European
Patent Office
Office européen
des brevets

Date: 10 April 2013

Subject: Creating multi-lingual, domain-specific technical dictionaries using the public preparation ..DICT, and using these dictionaries to perform machine translation.

Author: Ian Chapple

Version	Date	Change history
0.1	29-11-12	1st draft.
0.2	30-11-12	added English, German and French stopword lists.
0.3	05-12-12	minor changes, typo corrected.
0.4	13-12-12	added an annexes section.
1.0	17-12-12	First official version.
1.1	17-12-12	1st revision.
1.2	10-04-13	2nd revision

Table of contents

1. Background	3
2. ..alpha\wordstat	4
2.1. Screenshots showing the results of running ..alpha\wordstat on the file SA1141957	6
2.2. ..Rose	7
2.3. Origins of ..Dict.....	8
3. ..Dict	9
3.1. Screenshots of the dictionary for the Class G01D11	12
3.2. Implications and Limitations.....	13
4. Creating Public Dictionaries	14
5. The Translation Tool	15
5.1. Using Keyboard Shortcuts	17
5.2. Calling the Translation Tool from other Programs and Preparations	21
6. Conclusions	22
7. Annexes	23
7.1. Current Status.....	23
7.2. Incorrect use of Machine-Translated Documents	23
7.2.1. Details	24
7.3. Calling the Translation Tool	25
7.4. Stopwords	26
7.4.1. English Stopwords	27
7.4.2. German Stopwords	29
7.4.3. French Stopwords.....	31
7.5. Acknowledgements	32

1. BACKGROUND

The aim of this document is to explain how the idea of creating multi-lingual technical dictionaries came about, as well as detailing the process that was followed.

Some of the potential uses, and limitations, of the approach used will also be outlined.

The use of the dictionaries produced in providing a statistical machine translation capability will also be discussed.

2. ..ALPHA\WORDSTAT

The preparation **..alpha\wordstat** was first made public in March 2011. The purpose of this preparation was to analyse the text of a patent application, to:

1. Identify important phrases in the text;
2. Identify the most interesting paragraphs in the description;
3. Create a "Table of Contents";
4. Insert reference signs into the text of the claims;
5. Extract the most interesting concepts.

..alpha\wordstat works by first processing the text of the description and claims, to convert it into phrases; in this context, a phrase is a text "snippet" delimited by punctuation symbols and/or language-specific stopwords. Once converted into phrases, the text can be analysed, largely statistically, in a variety of ways.

The invention comprises, in one form thereof, a method of operating a wind turbine, including providing a wind turbine having a plurality of blades. A respective sensor is attached to each of the blades. First measurements of a structural characteristic of each of the blades are repeatedly taken by use of the sensors. A tolerance band is established for the measurements. Signals indicative of the first measurements are wirelessly transmitted only if the first measurements are outside of the tolerance band. The transmitted signals are received at a controller. An actuator signal is sent from the controller to at least one actuator associated with the blades. The actuator signal is sent in response to the receiving of the transmitted signals. At least one of the blades is actuated dependent upon the actuator signal. The actuating is performed by the at least one actuator. Second measurements of the structural characteristic of each of the blades are repeatedly taken by use of the sensors after the actuating step. The wirelessly transmitting, receiving, sending and actuating steps are repeated for the second measurements.

Original text of a paragraph (taken from SA1141957)

<phrase> FORM </phrase> <phrase> OPERATING </phrase> <phrase> WIND TURBINE
</phrase> <phrase> WIND TURBINE </phrase> <phrase> BLADES </phrase> <phrase>
SENSOR </phrase> <phrase> ATTACHED </phrase> <phrase> BLADES </phrase>
<phrase> MEASUREMENTS </phrase> <phrase> STRUCTURAL CHARACTERISTIC </phrase>
<phrase> BLADES </phrase> <phrase> REPEATEDLY TAKEN </phrase> <phrase>
SENSORS </phrase> <phrase> TOLERANCE BAND </phrase> <phrase> ESTABLISHED
</phrase> <phrase> MEASUREMENTS </phrase> <phrase> SIGNALS INDICATIVE
</phrase> <phrase> MEASUREMENTS </phrase> <phrase> WIRELESSLY TRANSMITTED
</phrase> <phrase> MEASUREMENTS </phrase> <phrase> OUTSIDE </phrase> <phrase>
TOLERANCE BAND </phrase> <phrase> TRANSMITTED SIGNALS </phrase> <phrase>
RECEIVED </phrase> <phrase> CONTROLLER </phrase> <phrase> ACTUATOR SIGNAL
</phrase> <phrase> SENT </phrase> <phrase> CONTROLLER </phrase> <phrase>
ACTUATOR ASSOCIATED </phrase> <phrase> BLADES </phrase> <phrase> ACTUATOR
SIGNAL </phrase> <phrase> SENT </phrase> <phrase> RESPONSE </phrase> <phrase>
RECEIVING </phrase> <phrase> TRANSMITTED SIGNALS </phrase> <phrase> BLADES
</phrase> <phrase> ACTUATED DEPENDENT </phrase> <phrase> ACTUATOR SIGNAL
</phrase> <phrase> ACTUATING </phrase> <phrase> PERFORMED </phrase> <phrase>
ACTUATOR </phrase> <phrase> MEASUREMENTS </phrase> <phrase> STRUCTURAL
CHARACTERISTIC </phrase> <phrase> BLADES </phrase> <phrase> REPEATEDLY TAKEN
</phrase> <phrase> SENSORS </phrase> <phrase> ACTUATING STEP </phrase>
<phrase> WIRELESSLY TRANSMITTING RECEIVING SENDING </phrase> <phrase>
ACTUATING STEPS </phrase> <phrase> REPEATED </phrase> <phrase> MEASUREMENTS
</phrase>

The same paragraph after having been converted to phrases by ..alpha\wordstat

One restriction that was applied was to limit the processing of phrases to those ending with what appeared to be a reference sign, as this enabled “noisy” terms/phrases to be excluded.

2.1. SCREENSHOTS SHOWING THE RESULTS OF RUNNING ..ALPHAWORDSTAT ON THE FILE SA1141957

Reference Sign	Item
[12]	BLADE HITTING TOWER, TOWER
[14]	NACELLE
[16]	ROTATION AROUND ROTOR, ROTOR
[18]	HUB
[22]	DIRECT SENSORS, INSTRUCT SENSORS, INSTRUCTING SENSORS, SENSOR NODES
[24]	AXES
[26]	ACCEPTABLE STEADY STATE RANGE CONTROLLER, CENTRAL CONTROLLER, FACE CONTROLLER, SPECIFICALLY CONTROLLER, STARTUP MODE CONTROLLER
[28]	ACTUATORS, BLADE ACTUATOR, COMMAND ACTUATOR
[34]	WIND DIRECTION SENSOR
[122]	DESIRABLE VALUE SENSOR, DIRECT SENSOR, INSTRUCT SENSOR, INSTRUCTING SENSOR
[126]	OPERATE CONTROLLER

Table of Contents

1. A method of operating a wind turbine, said method comprising the steps of: providing a wind turbine including a plurality of blades (10A,20A,20B,20); attaching a respective sensor (122,22A,22) to each of the blades (10A,20A,20B,20); repeatedly taking first measurements of a structural characteristic of each of the blades (10A,20A,20B,20), the measurements being taken by use of the sensors (22A,22B,22); establishing a tolerance band for the measurements; wirelessly transmitting signals (130,30A) indicative of the first measurements only if the first measurements are outside of the tolerance band; receiving the transmitted signals (130,30A) at a controller (126,26); sending an actuator signal (132) from the controller (126,26) to at least one actuator (128,28) associated with the turbine, the sending being in response to the receiving of the transmitted signals (130,30A); performing an action (136) associated with the turbine dependent upon the actuator signal (132), the action (136) being performed by the at least one actuator (128,28); repeatedly taking second measurements of the structural characteristic of each of the blades (10A,20A,20B,20), the measurements being taken by use of the sensors (22A,22B,22) after the actuating step (316); and repeating the wirelessly transmitting, receiving, sending and performing steps for the second measurements.

Claim 1, containing automatically inserted reference signs

SA1141957

[View the index file](#)

11. The invention comprises, in yet another form thereof, a method of controlling a machine, including installing a sensor, a controller and an actuator associated with the machine. Upon startup of the ...

9. The invention comprises, in one form thereof, a method of operating a wind turbine, including providing a wind turbine having a plurality of blades. A respective sensor is attached to each of the blades. ...

27. As soon as the sensor data goes beyond the limits of the predefined tolerance band, sensor 22 may initiate a communication between the sensor node and the base station (e.g., controller 26) in order ...

31. As wind turbine arrangement 10 begins to operate in the startup mode, controller 26 may adjust its command data to actuator 28 in order to reach optimal, or at least acceptable, values for electrical ...

10. The invention comprises, in another form thereof, a method of controlling a machine, including providing a sensor associated with the machine. First measurements of a parameter associated with the ...

56. As machine 120 begins to operate, controller 126 may adjust its command data to actuator 128 in order to reach optimal, or at least acceptable, values for the output parameters of machine 120. Controller ...

20. Disposed within nacelle 14 is a central controller 26 electrically connected to a blade actuator 28 within hub 18. Central controller 26 can be disposed anywhere within the wireless range of sensors ...

The present invention provides an energy-efficient communication scheme for wind turbines. However, the invention may also be applied to other types of machines, appliances and utilities.

More particularly, the invention may provide a wireless connection between the sensors on the rotor blades and the central unit in the base station, thus making all wiring, including slip rings, obsolete.

The invention comprises, in one form thereof, a method of operating a wind turbine, including providing a wind turbine having a plurality of blades. A respective sensor is attached to each of the blades. First measurements of a structural characteristic of each of the blades are repeatedly taken by use of the sensors. A tolerance band is established for the measurements. Signals indicative of the first measurements are wirelessly transmitted only if the first measurements are outside of the tolerance band. The transmitted signals are received at a controller. An actuator signal is sent from the controller to at least one actuator associated with the blades. The actuator signal is sent in response to the receiving of the transmitted signals. At least one of the blades is actuated dependent upon the actuator signal. The actuating is performed by the at least one actuator. Second measurements of the structural characteristic of each of the blades are repeatedly taken by use of the sensors after the actuating step. The wirelessly transmitting, receiving, sending and actuating steps are repeated for the second measurements.

The invention comprises, in another form thereof, a method of controlling a machine, including providing a sensor associated with the machine. First measurements of a parameter associated with the machine are repeatedly taken by use of the sensor. A tolerance band is established for the measurements. Signals indicative of the first measurements are wirelessly transmitted only if the first measurements are outside of the tolerance band. The transmitted signals are received at a controller. An actuator signal from the controller is sent to at least one actuator associated with the machine. The actuator signal is sent in response to the receiving of the transmitted signals. At least one part of the machine is actuated dependent upon the actuator signal. The actuating is performed by the at least one actuator. Second measurements of the parameter of the machine are repeatedly taken by use of the sensor after the actuating step. The wirelessly transmitting, receiving, sending and actuating steps are repeated for the second measurements.

The invention comprises, in yet another form thereof, a method of controlling a machine, including installing a sensor, a controller and an actuator associated with the machine. Upon startup of the machine, the sensor, controller and actuator are operated in a closed loop fashion. The closed loop operation includes taking a first startup measurement of a parameter associated with the machine. The measurement is taken by use of the sensor. A startup measurement signal indicative of the startup measurement is wirelessly transmitted. The transmitted startup measurement signal is received at a controller. A startup actuator signal is from the controller to the actuator in response to the receiving of the transmitted startup signal. At least one part of the machine is actuated dependent upon the startup actuator signal. The actuating is performed by the actuator. A second startup measurement is taken of the parameter of the machine. The second startup measurement is taken by use of the sensor after the actuating step. The wirelessly transmitting, receiving, sending and actuating steps are repeated for the second startup measurements. A tolerance band is calculated for steady state measurements of the parameter. The calculating is dependent upon the received startup measurement signals. The calculated tolerance band is wirelessly transmitted from the controller to the sensor. The machine is operated in steady state in an open loop fashion, including repeatedly taking first steady state measurements of the parameter associated with the machine. The first steady state measurements are taken by use of the sensor. Steady state measurement signals indicative of the first steady state measurements are wirelessly transmitted only if the first steady state measurements are outside of the tolerance band. The transmitted first steady state measurement signals are received at a controller. A steady state actuator signal is sent from the controller to the

Results of the "Paragraphs" analysis

2.2. ..ROSE

Due to improvements in the public preparation **..Rose**, which enables a full concept-based search to be carried out, it became possible to "send" the most interesting concepts directly to **..Rose** for searching.

To improve the quality and completeness of these concepts, a simple translation mechanism was introduced. This relied on the family principle, and worked by linking phrases found in one family member to the phrases found in another language in another family member, using the reference signs. Essentially, any phrases found in two or more languages and ending in the same reference sign were considered as being equivalent; this enabled the most relevant concepts to be translated before being passed on to **..Rose**.

This approach was based on the assumption that the same reference signs would be used to indicate the same features in different language family members.

In the meantime, the interaction between **..Rose** and **..alpha\wordstat** has been changed. Instead of **..alpha\wordstat** sending concepts to **..Rose**, **..Rose** now requests **..alpha\wordstat** to analyse the contents of an application using a special, **Rose**-specific, analysis mode, which makes additional use of the text of the title and abstract.

1	-	975	WIND	TURBINE
2	-	960	CONTROLLER	
3	-	920	SENSOR	
4	-	810	TOLERANCE	BAND
5	-	650	MEASUREMENTS	
6	-	428	BLADE	
7	-	390	ACTUATOR	
8	-	160	RESPONSE	
9	-	140	OUTSIDE	
10	-	80	ACTUATOR	SIGNAL
11	-	60	PARAMETER	
12	-	60	ACTUATING	STEP
13	-	21	STEP	
14	-	10	ACTION	
15	-	3	LONGITUDINAL	AXIS

The results of the "Rose" analysis of SA1141957

2.3. ORIGINS OF ..DICT

During a presentation of ..alpha\wordstat to directorate 1.2.36 in 2011, the question of whether or not the technique used for translation could be applied to whole classes rather than single families was raised. Although this had not previously been considered, it was an interesting suggestion, one which would eventually lead to the creation of the public preparation **..Dict**.

3. ..DICT

..Dict works in much the same way as ..alpha\wordstat, except that some or all of the families contained in an ECLA class are processed.

..Dict works by carrying out the following steps:

1. A class query is executed in EPODOC (e.g. **G01D1/08/EC**); this produces a list of documents;
2. The results are transferred to TCPAT (using **..xt/*xt**); this produces a list of families;
3. These families are then combined with queries containing fulltext database names; this enables those families having fulltext available in two (bi-lingual family) or all three (tri-lingual family) official languages of the EPO to be identified;

TCPAT	SS	Status	Results	Query
> 1			1.150	*XT
2			20.397.286	OR TXTAU1, TXTCNT, TXTEP1, TXTGB1, TXTJPR, TXTJPS, TXTJPT, TXTKRT, TXTLN1, TXTUS0, TXTUS1, TXTUS2, TXTUS3, TXTUS4, TXTUS5, TXTW01, TXTW0T
3			3.612.136	OR TXTAT1, TXTBEG, TXTCHG, TXTDE1, TXTEPG, TXTWOG
4			1.922.185	OR TXTBEF, TXTCHF, TXTEPF, TXTFR1, TXTLNF, TXTWOF
5			52.707	AND 2, 3, 4
> 6			7	1 AND 5
7			799.223	OR (2 AND 3), (2 AND 4), (3 AND 4)
8			746.516	7 NOT 5
> 9			65	1 AND 8

Callouts in the image:

- Transfer from EPODOC (points to row 1)
- English fulltext (points to row 1)
- German fulltext (points to row 3)
- French fulltext (points to row 4)
- Fulltext available in all 3 languages (points to row 7)
- Fulltext available in 2 out of 3 languages (points to row 8)
- Tri-lingual families (points to row 9)
- Bi-lingual families (points to row 6)

4. The list of bi-

and tri-lingual families is stored in a file (family file);

```

G01D1.txt - Notepad
File Edit Format View Help
GB2195776 TXTGB1 DE3632672 TXTDE1 FR2604517 TXTFR1
US4858153 TXTUS1 DE3611772 TXTDE1 FR2596869 TXTFR1
GB2283547 TXTGB1 DE4338069 TXTDE1 FR2712243 TXTFR1
GB2280718 TXTGB1 DE4325940 TXTDE1 FR2708546 TXTFR1
GB2470262 TXTGB1 DE202010005371U TXTDE1 FR2945353 TXTFR1
GB2291204 TXTGB1 DE19524884 TXTDE1 FR2722287 TXTFR1
JP2002366225 TXTJPT DE10123625 TXTDE1 FR2824908 TXTFR1
GB2332077 TXTGB1 DE19633550 TXTDE1 # #
US5253173 TXTUS1 DE3841089 TXTDE1 # #
US2011304345 TXTUS4 DE102009022992 TXTDE1 # #
  
```

5. The fulltext descriptions of each family member are checked, and if they have not previously been downloaded, they are downloaded via Epoque Internal;
6. Each description is saved as a zipped text file, to save space.

Once all of the fulltext descriptions have been downloaded and stored, control passes from the preparation ..Dict to the REXX program **Dict.rexx**. This program is identical to the preparation ..Dict, but is able to run outside of Epoque; this prevents the Epoque Internal session from being blocked for longer than is absolutely necessary.

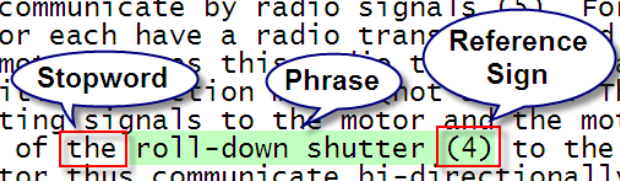
Dict.rexx executes the following steps:

1. The family file is read in;
2. Each family is processed, one family at a time;
3. The description for each family member is read in, and the text is scanned, word-by-word, to identify words which appear to be reference signs;
4. Once a reference sign is identified, the text is processed backwards, until a punctuation symbol or stopword is found; the text delimited by the stopword and the reference sign forms a phrase, which is then stored in a list;

[0076] Fig. 1 shows a graphic representation of a remote control (1) according to an embodiment of the invention, in which the real-time position of a roll-down shutter (4) operated by a motor (3) is displayed on the screen (2) of the remote control. The remote control (1) and motor (3) communicate by radio signals (5). For this, the remote control and motor each have a radio transmitter and receiver (not shown). The motor shares this radio transmitter and receiver with internal position detection means (not shown). The remote control sends operating signals to the motor and the motor sends data on the position of the roll-down shutter (4) to the remote control. The remote and motor thus communicate bi-directionally.

Extract from EP2533115













[0076] Fig. 1 shows a graphic representation of a remote control (1) according to an embodiment of the invention, in which the real-time position of a roll-down shutter (4) operated by a motor (3) is displayed on the screen (2) of the remote control. The remote control (1) and motor (3) communicate by radio signals (5). For this, the remote control and motor each have a radio transmitter and receiver (not shown). The motor shares this radio transmitter and receiver with internal position detection means (not shown). The remote control sends operating signals to the motor and the motor sends data on the position of the roll-down shutter (4) to the remote control. The remote and motor thus communicate bi-directionally.



The same extract, showing a phrase identified by reference sign (4)

5. Once the two or three documents in the family have been processed, the lists of phrases ending in reference signs are compared;

6. Any phrases in the different languages having the same reference signs are considered as being equivalent, and are saved as a single dictionary entry;
7. Once all of the families have been processed, the dictionary entries are de-duplicated and sorted in English (or in French or German, if preferred), and are then saved in a variety of formats:
 - HTML (for viewing in a browser)
 - text version (this is the “master” copy)
 - CSV (for importing into a database application)
 - index (this is used for retrieval)
 - nGrams (bigrams, trigrams & quadrigrams [groups of 2, 3 and 4 words])

Name ▲	Size
 dict.js	1 KB
 G01D7.bigram	183 KB
 G01D7.csv	151 KB
 G01D7.htm	760 KB
 G01D7.index	233 KB
 G01D7.quadrigram	45 KB
 G01D7.trigram	96 KB
 G01D7.txt	143 KB
 G01D7.unsorted	207 KB
 print.css	1 KB
 screen.css	1 KB
 sort.png	1 KB

The list of files corresponding to the dictionary for the class G01D7

3.1. SCREENSHOTS OF THE DICTIONARY FOR THE CLASS G01D11

51.	<ul style="list-style-type: none"> • ABUTMENT 	<ul style="list-style-type: none"> • GEGENLAGER 	
52.	<ul style="list-style-type: none"> • ABUTMENTS • INSIDE AGAINST ABUTMENTS 	<ul style="list-style-type: none"> • WIDERLAGER 	<ul style="list-style-type: none"> • BUTÉES
53.	<ul style="list-style-type: none"> • ABUTMENT SURFACE • CONTACTING ABUTMENT SURFACES • FITTING AREA • FITTING SURFACE • IDEALLY ABUTMENT SURFACES 	<ul style="list-style-type: none"> • ANSCHLAGFLÄCHE • KOMPLEMENTÄRE ANSCHLAGFLÄCHE • PASSFLÄCHE 	

HTML version

```

50 ABS_(ANTI_LOCK_BRAKE_SYSTEM)_ALARM_LAMP ABS_ALARM_LAMP # ABS_WARNLAMPE
51 ABUTMENT # GEGENLAGER #
52 ABUTMENTS INSIDE_AGAINST_ABUTMENTS # WIDERLAGER # BUTÉES
53 ABUTMENT_SURFACE CONTACTING_ABUTMENT_SURFACES FITTING_AREA FITTING_SURFACE
54 ABUTMENT_WALL STUDS # # ERGOTS ERGOTS_LATÉRAUX PAROIS_DE_BUTÉE
55 ABUTMENT_WEBS # RAGENDE_WIDERLAGERSTEGE WIDERLAGERSTEGEN #
56 ABUTS_CONTACT CONTACT ELECTRIC_CONTACT # ABDECKELEMENT ELEKTRISCHE_KONTAKT

```

Text version

```

50 "ABSORB BODY MEMBER; BODY MEMBER; FASTENING BODY MEMBER; MOMENT"
51 "ABS (ANTI LOCK BRAKE SYSTEM) ALARM LAMP; ABS ALARM LAMP", "ABS_WARNLAMPE"
52 "ABUTMENT", "GEGENLAGER", ""
53 "ABUTMENTS; INSIDE AGAINST ABUTMENTS", "WIDERLAGER", "BUTÉES"
54 "ABUTMENT SURFACE; CONTACTING ABUTMENT SURFACES; FITTING AREA; FITTING SURFACE"
55 "ABUTMENT WALL; STUDS", "", "ERGOTS; ERGOTS LATÉRAUX; PAROIS DE BUTÉE"
56 "ABUTMENT WEBS", "RAGENDE WIDERLAGERSTEGE; WIDERLAGERSTEGEN", ""

```

CSV version

```

330 ABTRIEBSZAHNRAD : 1807 DE 1
331 ABUT : 60 EN 1
332 ABUTMENT : 51 EN 1 : 53 EN 1 : 53 EN 4 : 53 EN 11 : 54 EN 1 :
333 ABUTMENTS : 52 EN 1 : 52 EN 4
334 ABUTS : 56 EN 1
335 ABUTTING : 57 EN 1 : 58 EN 1 : 59 EN 1
336 ABWEICHENDE : 762 DE 3

```

Index

```

515 ABS_WARNLAMPE : 50 DE 1
516 ABTÄSTENDE_ABSCHNITT : 2056 DE 2
517 ABUTMENT_AGAINST : 2658 EN 2
518 ABUTMENT_SURFACE : 53 EN 1
519 ABUTMENT_SURFACES : 53 EN 2 : 53 EN 5
520 ABUTMENT_WALL : 54 EN 1
521 ABUTMENT_WEBS : 55 EN 1
522 ABUTS_CONTACT : 56 EN 1

```

Bigrams

3.2. IMPLICATIONS AND LIMITATIONS

There are various implications resulting from this method of processing the descriptions:

1. Any phrases not ending in what can be identified as a reference sign are ignored.
2. Any "feature" that is not represented or clearly identified in the figures will be ignored.
3. Items in the figures indicated by multiple reference signs (e.g. *Weighbeams 13 and 15*) are generally only linked to the first reference sign found.
4. The terms extracted are largely represented by nouns, with some associated adjectives. Any verbs, pronouns, adverbs and adjectives not directly associated with a phrase identified as ending in a reference sign will not be present in the dictionary.
5. Each dictionary record is usually based on a single reference sign from a single family, although there is some merging/grouping of records.
6. The quality of the fulltext plays a large role. OCR errors, formatting errors and missing or incorrect reference signs all have a negative impact on the quality of what can be extracted.
7. The translations are not necessarily **linguistically** correct; as it was the original intention to store phrases which would lend themselves to being used as search queries in Epoque, the decision was taken to suppress certain parts of the source text, such as **l'** and **d'** in French.
8. As the phrase creation relies heavily on the detection of stopwords, the maintenance of the stopword lists is extremely important. However, this is essentially a never-ending task and, as such, the list of stopwords in each of the three official languages is never 100% complete. This means that certain stopword-like words (mainly adjectives and adverbs) may be present in the dictionaries.
9. It is also important to note that the preparation ..Dict and the REXX program Dict.rex have absolutely no linguistic awareness of any kind. If a phrase in one language appears to match a phrase in another language (based purely having a common reference sign), even if it would be immediately obvious to a person that these phrases are totally unrelated, then they will be considered as equivalents and stored as a dictionary entry.

4. CREATING PUBLIC DICTIONARIES

Once the dictionaries produced by ..Dict were seen to be of a reasonably high quality, it became clear that to make it truly useful, it would be necessary to process all classes and technical domains present in ECLA.

1. A list of all **7201** top-level ECLA classes (those ending in /00) was created;
2. Dictionaries for each of these classes were created, based purely on tri-lingual families;
3. All of the dictionaries created in a single technical domain were combined, to produce a single domain-specific dictionary (e.g. the dictionaries G01D1, G01D3, G01D4, G01D5, G01D7, G01D9, G01D11, G01D13, G01D15, G01D18 and G01D21 were combined to form the dictionary G01D);
4. These domain-specific dictionaries were then uploaded to a public share (**W:\PublicDictionaries**), thereby giving all examiners access to the dictionaries.

At this stage, the coverage was relatively poor, but at least a certain number of dictionaries were made available in a short space of time.

The decision was then taken to expand the coverage of each dictionary, by processing some bi-lingual as well as all tri-lingual families. By maintaining a master-list of the classes processed, it was possible to create new lists of classes to re-process, based on factors such as the number of families present in the class and what proportion had already been processed. In every case, each list lead to an increased coverage for those classes contained in the list.

In November 2012, a new master list of top-level ECLA classes was compiled; this revealed that 114 new top-level classes had been created, and 116 had been deleted. As a lot of work had already been done on those classes that had been deleted, it was decided to add the new classes to the master-list but not to remove the deleted classes; this resulted in a new list containing **7315** top-level classes. It also transpired that six technical domains (B29H, B29K, B29L, C10N, F21M and F21Q) had been deleted and three new ones (C12Y, G04R and H02S) had been created.

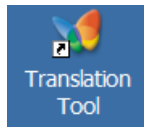
At this moment, it is not clear whether the introduction of CPC will have any significant effect on the work already completed, or whether extra processing will be needed. Approximately 1800 new top-level CPC classes have been created, which correspond to ICO codes which were **not** extensions to ECLA classes (these are known as orthogonal ICO codes). These classes are identified by having four digits before the slash; eg. G01K**2201**/00.

It may be necessary to process the families classified using such class symbols, but as most of the families will already have been classified with ECLA classes, it is unclear how much of an improvement this will bring.

5. THE TRANSLATION TOOL

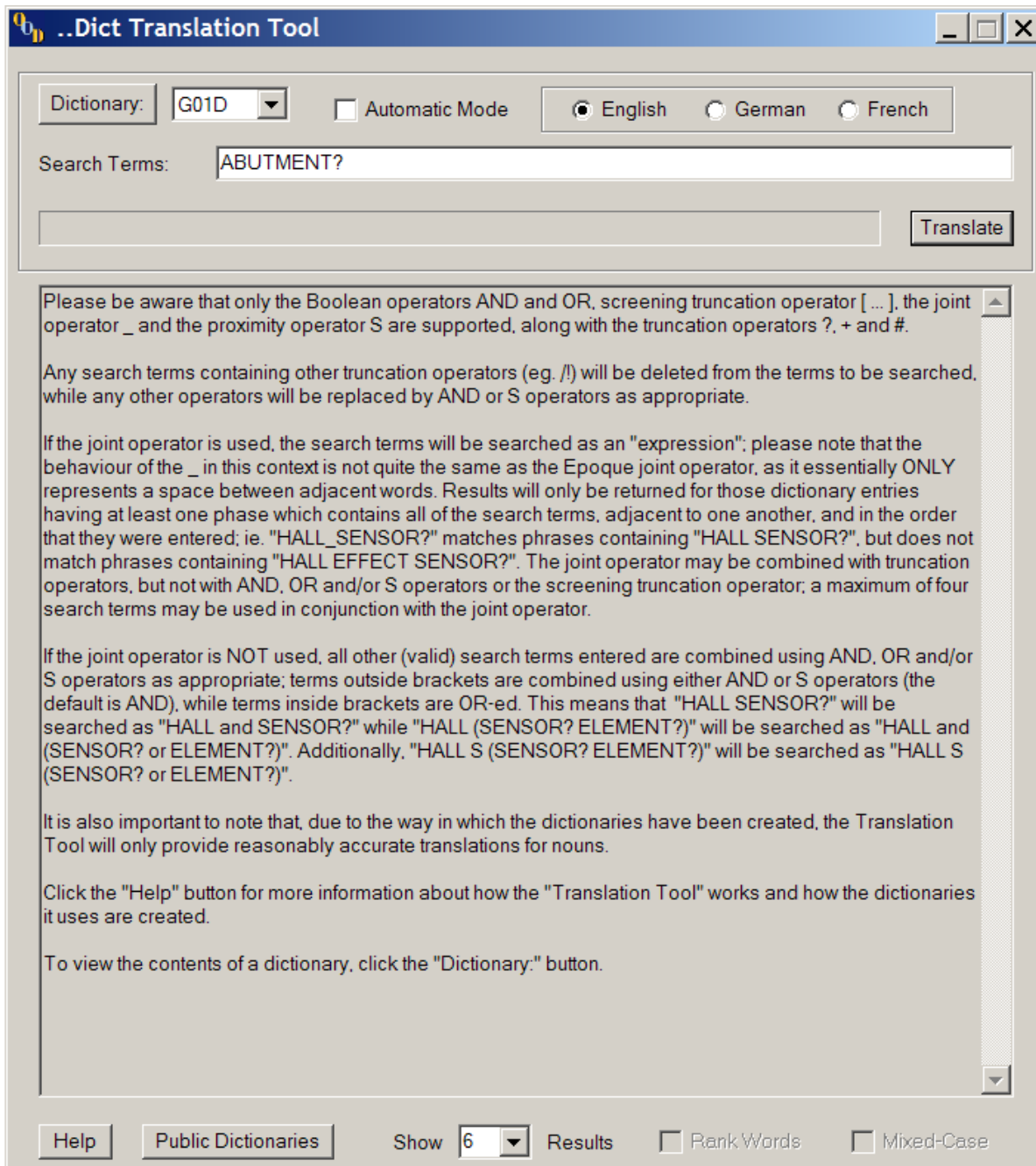
A later addition to the process was the creation of a **Translation Tool**, to provide translations and/or synonyms, based on the content of the dictionaries.

The Translation Tool can be started by typing `..dict translate` in Epoque Internal, or by clicking this icon



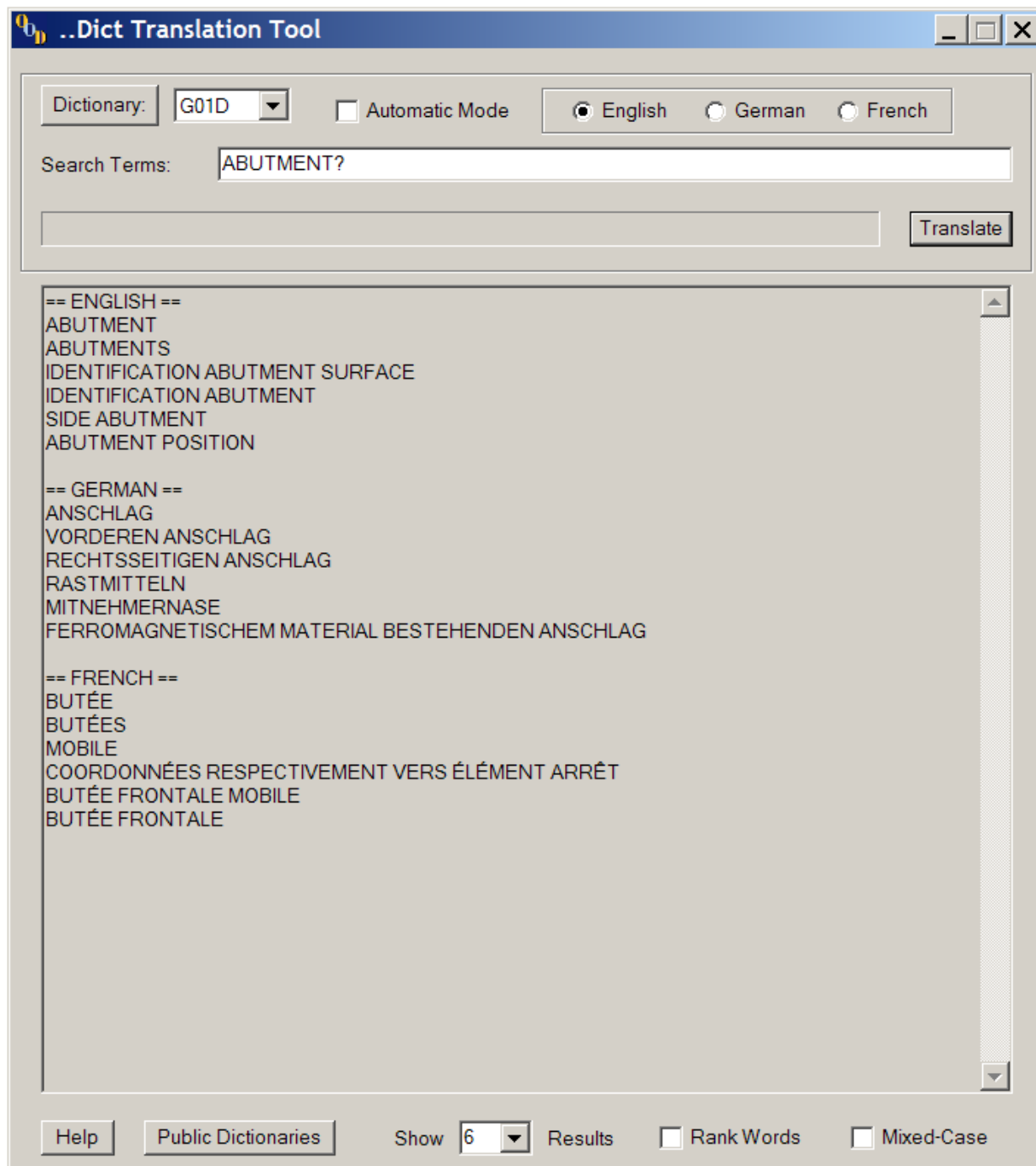
on the desktop.

The following window then opens:



Here it is possible to select the dictionary to be used, whether the **Automatic Mode** should be used, the language of the search terms being entered, the number of results to display, and the search terms themselves.


Clicking on the **Translate** button causes the selected dictionary to be loaded, after which the search terms entered are searched in the index. The results are presented in the same window, and represent the most statistically-relevant phrases found in all three official languages.



Because the Translation Tool works statistically, the results of the translation are *usually* meaningful, as the less interesting (noisier) phrases present in the dictionaries are in general less statistically significant, and therefore should either be excluded entirely from the results, or should appear sufficiently far down that they may be ignored.

If the **Automatic Mode** is selected, the Translation Tool translates the search terms using the specified language, in the dictionary that it identifies as being the most relevant.

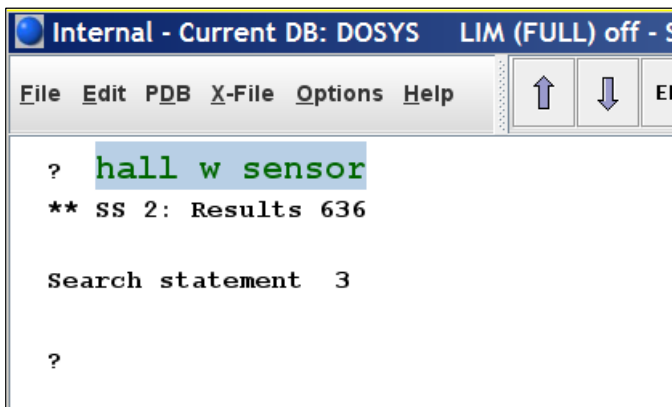
5.1. USING KEYBOARD SHORTCUTS

If this icon  is visible in the Windows SysTray, then the Translation Tool may be accessed using keyboard shortcuts.

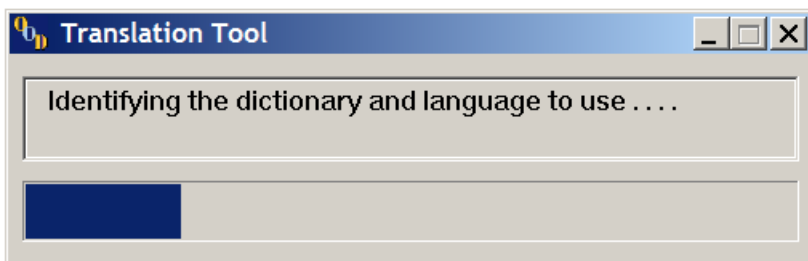
Pressing the key combination **WinKey+T** causes the Translation Tool window to open, with any marked text being copied to the Search Terms entry field.

Pressing the key combination **WinKey+S** will cause any marked text to be automatically translated. In this case, a special automatic mode is used, in which the indexes of all dictionaries are checked, to identify both the dictionary and the language most likely to produce meaningful results.

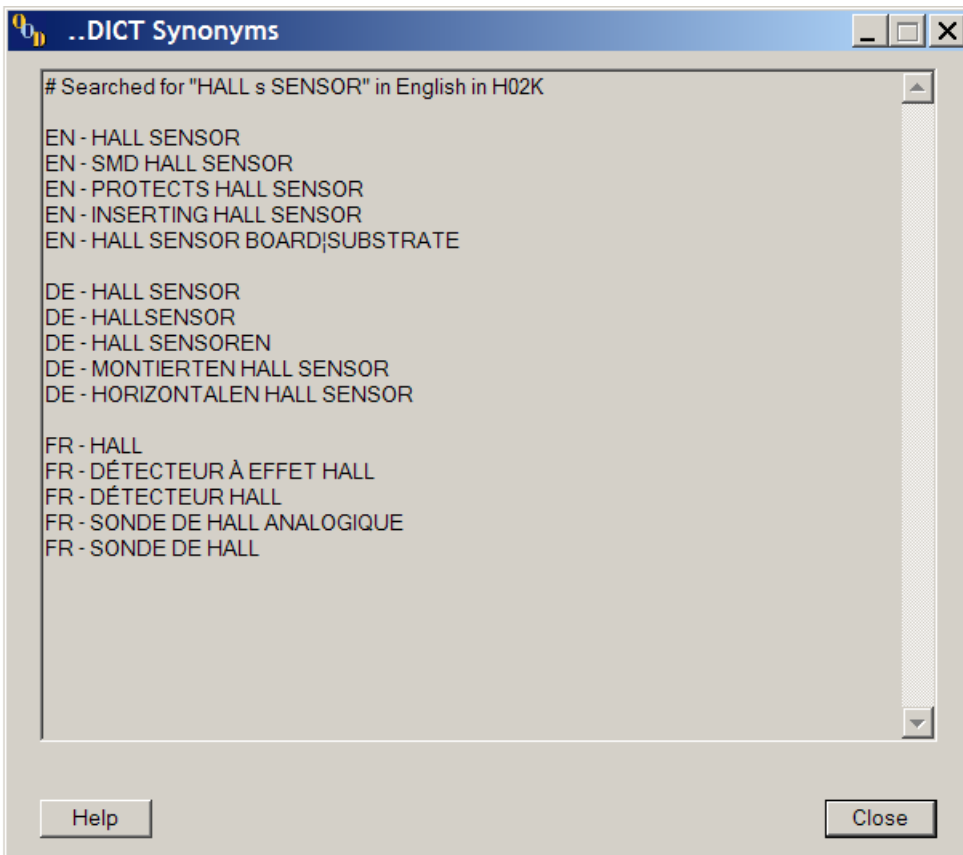
For example, if the phrase **hall w sensor** is marked in Epoque Internal



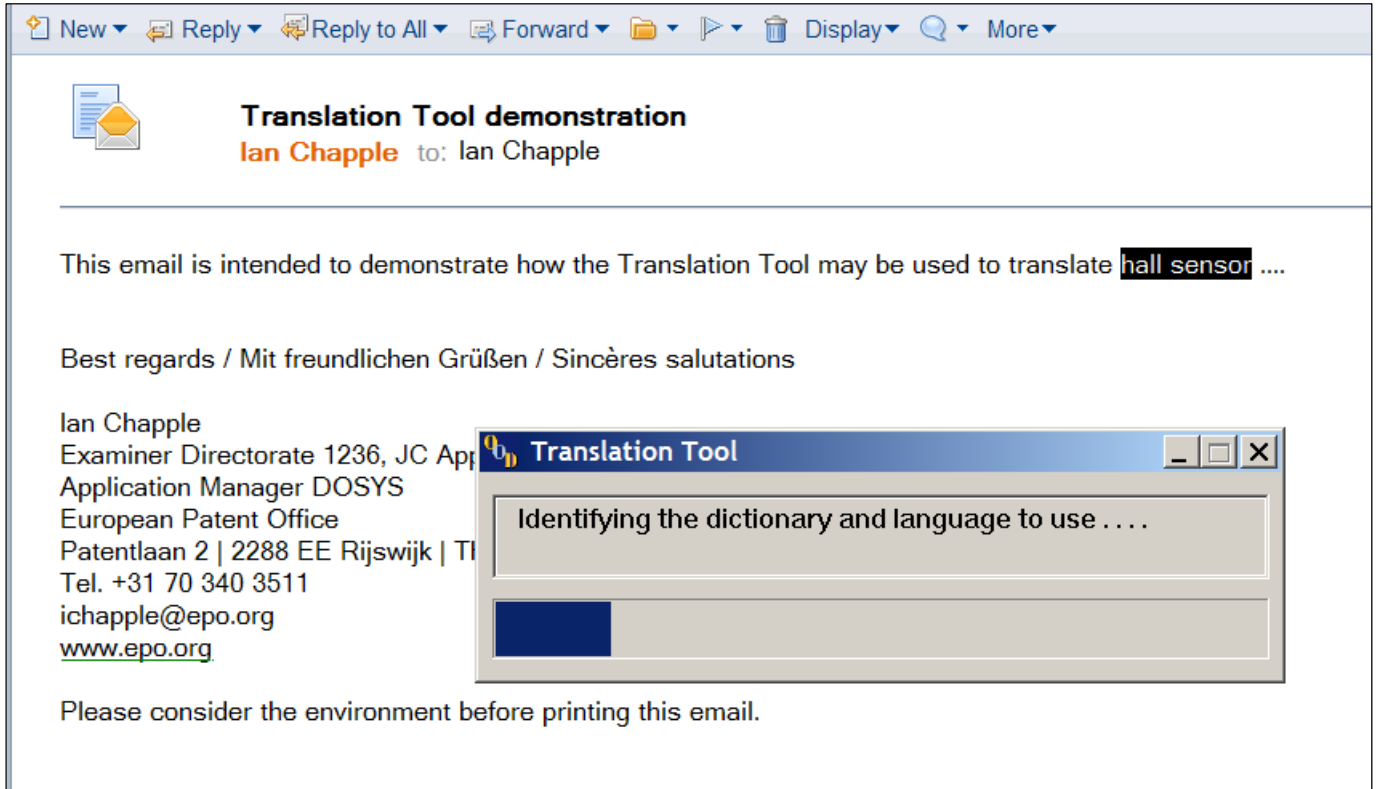
and **WinKey+S** is pressed, the following window appears:



Once the target dictionary and language have been identified, translation takes place and the results are presented in a small pop-up window:

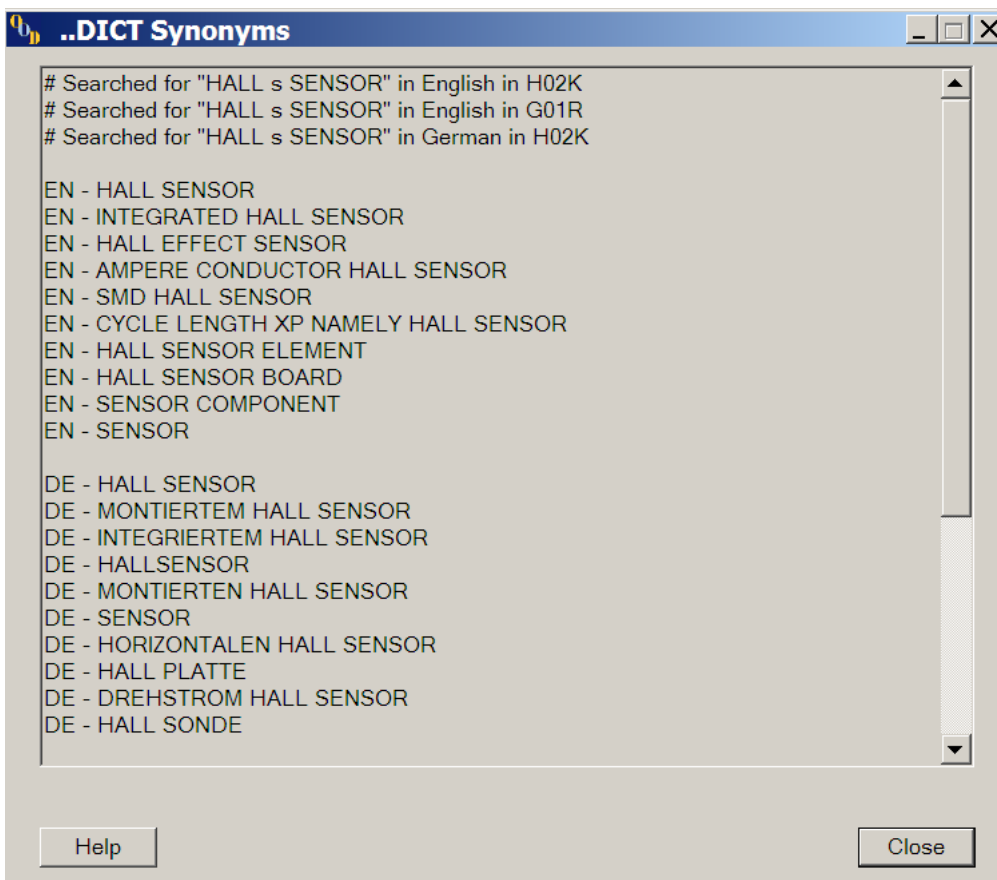
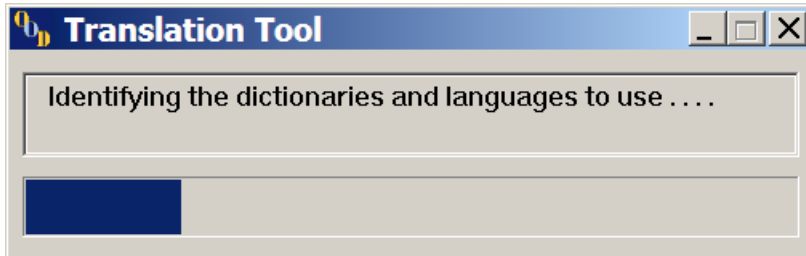


The same approach may be used for translating text in eg. an email:

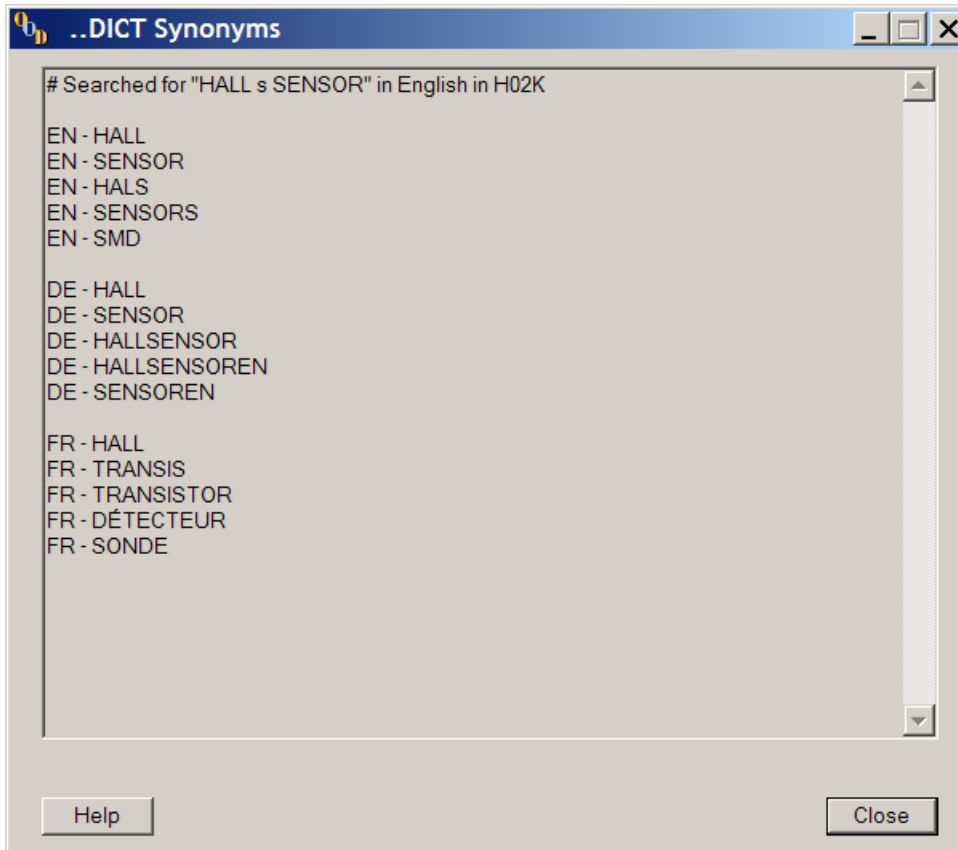


The shortcut **Alt+WinKey+S** is similar to **WinKey+S**. However, instead of basing the results on only the first dictionary/language pairing identified as being likely to produce meaningful results, multiple translations are performed, using different dictionary/language pairings; the results of these multiple translations are then combined.

If the term **hall w sensor** is marked and then **WinKey+Alt+S** is pressed, the popup window shows that the results are based on 3 separate translations:



The shortcut **WinKey+W** works in a similar way to **WinKey+S**, except that the results of the translation are individual words rather than phrases.

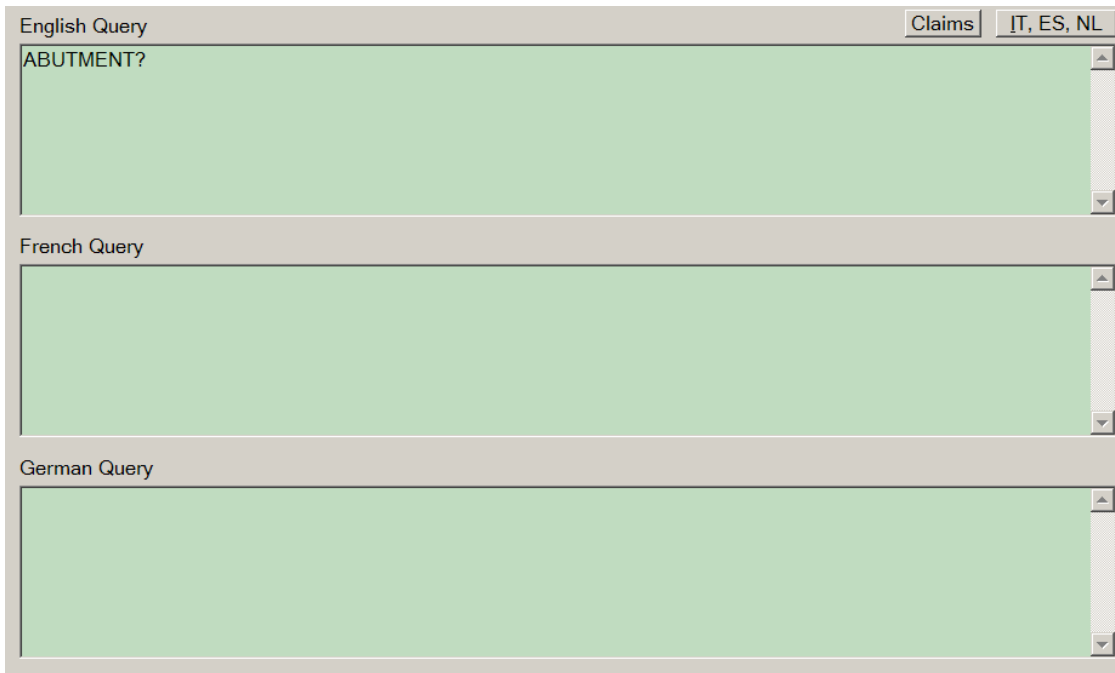


5.2. CALLING THE TRANSLATION TOOL FROM OTHER PROGRAMS AND PREPARATIONS

As the Translation Tool is a standard REXX program, it may also be called by other programs or preparations.

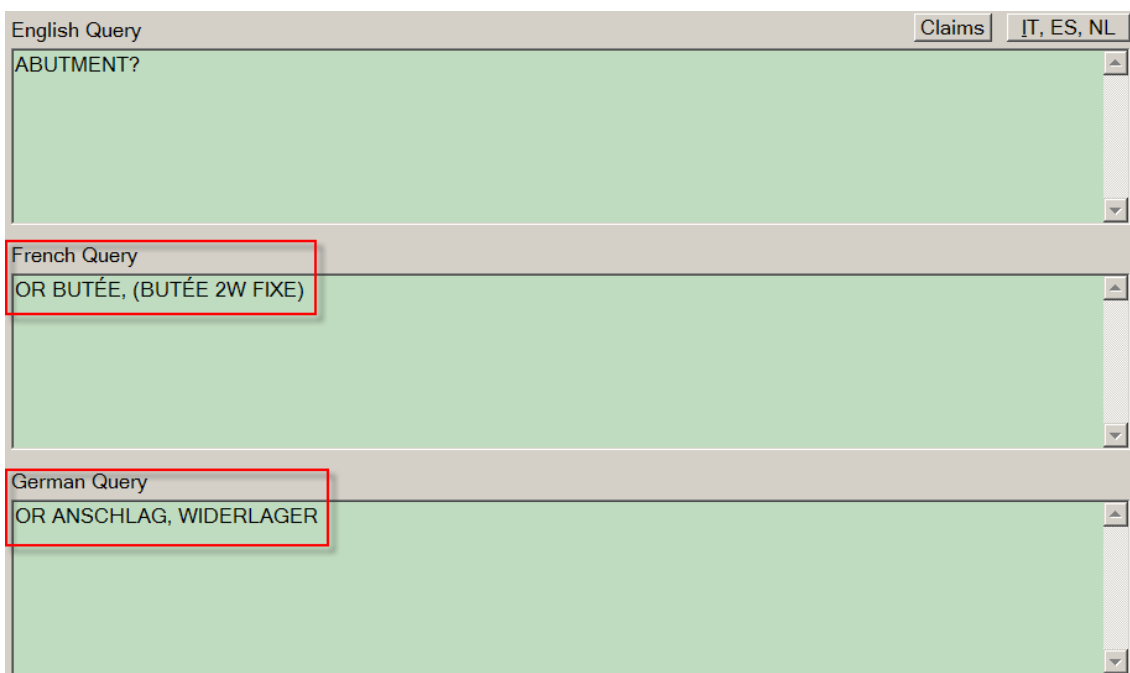
Currently, only the public preparation ..Rose makes use of this functionality, but it has so far proved to be quite reliable.

As an example, if the English query **abutment?** is entered in the ..Rose editor,



The screenshot shows a window titled 'English NL Query' with three tabs: 'Claims', 'IT, ES, NL', and 'NL, ES, IT'. The 'English Query' field contains the text 'ABUTMENT?'. Below it, the 'French Query' and 'German Query' fields are empty.

and the **Auto-Translate** button is clicked, then the French and German queries are automatically populated with the **2** best results of the automatic translation of **abutment?**.



The screenshot shows the same window as above, but now the 'French Query' field contains the text 'OR BUTÉE, (BUTÉE 2W FIXE)' and the 'German Query' field contains the text 'OR ANSCHLAG, WIDERLAGER'. Both fields are highlighted with red boxes.

6. CONCLUSIONS

The approach used by ..Dict and its Translation Tool offers some advantages when compared to other translation/synonym tools. It also appears to be an approach which has not been used on a large scale before.

The Epoque standard application **QBA** works in a similar way, but relies primarily on search queries entered in the search tool XFull. The main limitation is that if a particular expression has not been used during a search, it will not be found by the QBA.

Online dictionaries and translation services have been available for a while; good examples of these are the **leo.org** dictionary (popular amongst EPO examiners) and **Google Translate**.

The main problem with these online services is that it is often difficult, or impossible, to make use of them from within a program or preparation; most are not usually directly accessible programmatically, as they either do not have an API (an interface usable by a computer program) or if they do, it is only for commercial use. However, the overall quality of the content is probably higher in most cases.

Another problem is one of confidentiality, as it is currently not recommended to use Google Translate for unpublished patent applications.

The domain-specific dictionaries produced by ..Dict rely on the fact that family members in different languages are usually reasonably faithful translations of one another. Therefore, unlike the QBA, they do not rely on terms or expressions which have been used during a search.

Similarly, because the Translation Tool is a standard REXX program, it is potentially available to any EPO application or preparation which could benefit from some kind of translation capability.

It is hoped that the content of the dictionaries will eventually be made available via an Epoque database, as this will make the data available to a much wider range of tools and preparations, including the QBA.

7. ANNEXES

7.1. CURRENT STATUS

The creation of the domain-specific dictionaries has so far lasted 21 months (August 2011 until April 2013) which amounts to approximately 17 months actual processing time. All of the work has been carried out on a single examiner PC, which has been running for an average of 20 hours per day (including weekends); when required, additional use was made of the author's personal laptop. It is estimated that the total processing time is in the region of 11,000 hours.

The processing has been carried out in a number of distinct phases (currently 56 phases), where each phase is based on a list of classes to re-process; each list is derived from the contents of the master-list, and is intended to improve the coverage for those classes meeting certain criteria (e.g. the total number of families is above a certain threshold, the percentage of the total number of families processed is below a certain threshold etc.).

Extensive use has been made of the Epoque framework (mainly Epoque Internal and a wide variety of fulltext databases), as well as a number of REXX programs to compile the lists of classes to re-process and carry out the uploading and backing-up of the data produced.

7.2. INCORRECT USE OF MACHINE-TRANSLATED DOCUMENTS

In January 2013, it was discovered that many English-language documents being used to create the dictionaries were in fact themselves machine-translations, from either Chinese, Japanese or Korean. This was caused by processing the list of databases in alphabetical order when identifying which documents to download.

This necessitated the carrying out of a major re-processing task, to reduce the degree of reliance on machine-translated documents; it was also decided to increase the degree of coverage for many classes containing large numbers of families, by processing additional families. This involved re-processing approximately 5000 classes and the downloading of approximately 600,000 new fulltext descriptions.

7.2.1. Details

Valid on 10th April 2013

Zippped volume of fulltext processed: 25.7GB
Total volume of fulltext processed: **90.1GB**

Total number of families processed: **1774597**
Number of de-duplicated families: 1028471
Number of unique families: **1024068**
Number of tri-lingual families: 60635
Number of bi-lingual families: 963433
Total number of docs processed: 3664117
Number of unique docs processed: **2120204**

Total number of records: **18275640**
Total number of indexed terms: 17272884
Total number of phrases: **88860896**
Total number of bigrams: 37326984
Total number of trigrams: 23207510
Total number of quadrigrams: 10310035

Number of unique indexed terms: **3001018**

7277 / 7315 classes (99%) have been processed to at least 100%.
7290 / 7315 classes (99%) have been processed to at least 90%.
7308 / 7315 classes (99%) have been processed to at least 75%.
7313 / 7315 classes (99%) have been processed to at least 50%.
7315 / 7315 classes (100%) have been processed to at least 25%.

TCPAT statistics

=====

Number of multi-lingual families: **1202859**
Number of tri-lingual families: 68129
Number of bi-lingual families: 1134730

7.3. CALLING THE TRANSLATION TOOL

To call the Translation Tool from other programs or preparations, the call must pass via the Windows command line; the syntax is as follows:

```
W:\PublicDictionaries\@Support\TranslationTool.rex {field} {language}  
{search terms} -nogui
```

The syntax for the automatic mode is similar:

```
W:\PublicDictionaries\@Support\TranslationTool.rex {search terms} -auto
```

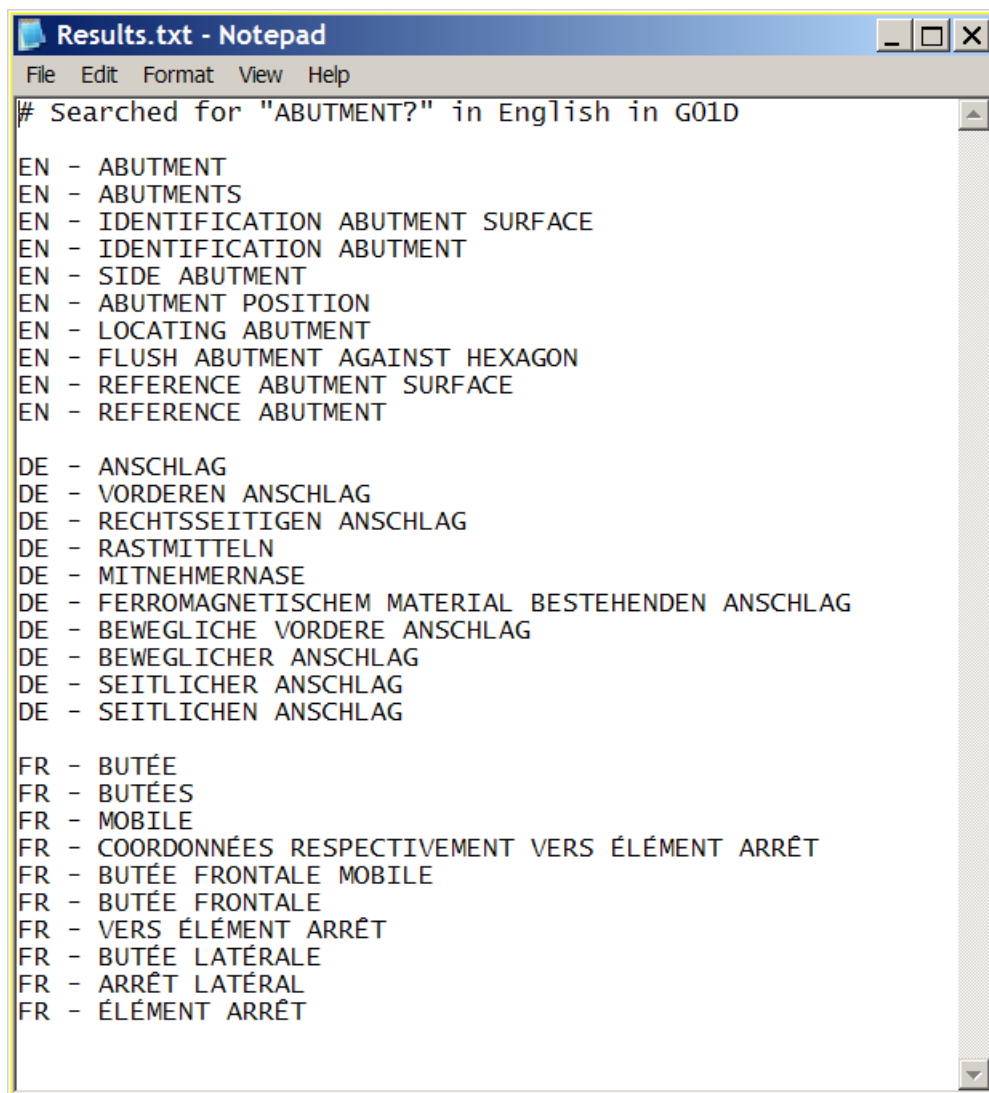
If no popup window containing the translations is needed:

```
W:\PublicDictionaries\@Support\TranslationTool.rex {search terms} -auto -  
quiet
```

The following call would translate the **English** term **abutment?** in the dictionary **G01D**:

```
W:\PublicDictionaries\@Support\TranslationTool.rex g01d en abutment?  
-nogui
```

The results are saved in the form of a local text file, which can be read in by the calling program:
C:\UserData\PublicDictionaries\@Support\TranslationResults\Results.txt.



```
Results.txt - Notepad  
File Edit Format View Help  
# Searched for "ABUTMENT?" in English in G01D  
  
EN - ABUTMENT  
EN - ABUTMENTS  
EN - IDENTIFICATION ABUTMENT SURFACE  
EN - IDENTIFICATION ABUTMENT  
EN - SIDE ABUTMENT  
EN - ABUTMENT POSITION  
EN - LOCATING ABUTMENT  
EN - FLUSH ABUTMENT AGAINST HEXAGON  
EN - REFERENCE ABUTMENT SURFACE  
EN - REFERENCE ABUTMENT  
  
DE - ANSCHLAG  
DE - VORDEREN ANSCHLAG  
DE - RECHTSSEITIGEN ANSCHLAG  
DE - RASTMITTELN  
DE - MITNEHMERNASE  
DE - FERROMAGNETISCHEM MATERIAL BESTEHENDEN ANSCHLAG  
DE - BEWEGLICHE VORDERE ANSCHLAG  
DE - BEWEGLICHER ANSCHLAG  
DE - SEITLICHER ANSCHLAG  
DE - SEITLICHEN ANSCHLAG  
  
FR - BUTÉE  
FR - BUTÉES  
FR - MOBILE  
FR - COORDONNÉES RESPECTIVEMENT VERS ÉLÉMENT ARRÊT  
FR - BUTÉE FRONTALE MOBILE  
FR - BUTÉE FRONTALE  
FR - VERS ÉLÉMENT ARRÊT  
FR - BUTÉE LATÉRALE  
FR - ARRÊT LATÉRAL  
FR - ÉLÉMENT ARRÊT
```

7.4. STOPWORDS

Due to the sheer number of stopwords as well as the large number of variants (mainly in German and French), some stopwords end with truncation characters (either ? or ??). This simplifies the maintenance of the stopword lists, but does rely on a stopword identification routine which supports the use of such truncation characters.

The following stopword lists were valid on 7th December 2012.

7.4.1. English Stopwords

a about above abovementioned accordance according accordingly across actual actually adapted advantage advantageously advantages aforementioned aforesaid after again all along already also alternative alternatively alternatives although among an and and/or another any apparatus applicant applicants application applications approximate approximately april are arranged arrangement arrangements as at august

b based be because becomes been being between both brought but by

c can characterise characterised characterises characterize characterized characterizes claim claimed claims compose composed composes comprise comprised comprises comprising configured consist consisting consists could

d december deg describe described describes desire desired desires device devices different dispose disposed disposes div do does due during

e each effective effectively eg e.g. e.g. either embodiment embodiments especially even eventual eventually ever every example examples exemplary

f fabrication february fifth fig figs fig. figure figures first follow follows following for fourth friday from ft further furthermore

g general generally give given gmbh good great greater greatest

h had has have having here hereby herein hereupon how however

i ie ie. i.e. if in inc include included includes including into invention inventions is it its it's

j january july june just

k kabushiki kaisha kk known

l least like likewise ltd

m made make making many march may means meanwhile mention mentioned mentioning mentions method methods monday more moreover most much machine

n na no nohe norm not november now

o object objects obtain obtains obtained obtaining obtd october of on one ones only onto or other otherwise over

p part parts particular particularly patent pct pd plurality pn possible possibly preferable preferably preferred prescribe prescribed prescribes present process processes processed provide provided provides providing purpose purposes publish published publishes publishing

q

r refer refers referred referring relate relates relating relatively remain remains remaining require requires requiring respective respectively

s said same saturday search second select selected september should similar since so special specification specifications stated statement statements substantially sub such suitable suitably sunday supp system systems

t tb than that the their them then thence there thereafter thereby therefore therein thereof thereon thereto therewith these they third this those three through thursday thus title titled to tuesday two

u until up upon use used useful using

v various

w was wednesday were what when whenever where whereby wherein whether which
while who whose will with within without would

x

y

z

1st 2nd 3rd 4th 5th

7.4.2. German Stopwords

a ab aber ag ähnlich?? al alle allen alles als also ander?? angeordnet??
anlage anordnung anspr anspruch?? ansprüche? april auch august am an auf aus
ausgebildet?? ausserdem ausführungsbeispiel?? ausführungsform
ausführungsformen ausführungsforms abb abbildung abbildungen

b bei beid?? beim beispiel beispiele beispielhaft?? beispielsweise
beschrieben?? beziehungsweise bezugszahlen bezugszahlenlist?? bezugszeichen
bezugszeichenlist?? bezugsziffer bis bzw

c

d da dabei dadurch dafür daher damit dann daran dargestellt?? darauf darum
darunter das dasjenige dass dasselbe davon dazu dazugehörig?? de dem demjenige?
demselbe? den denen denjenige? dennoch denselbe? der derartig?? deren
dergleichen derjenige? derselbe? des deshalb desselbe? dessen deswegen
dezember die diejenige? diensttag dies diese diese? dieselbe? doch donnerstag
dr drei dritt?? du durch durchführen durchführung

e ebenfalls eigen?? eigentlich ein eine eine? einander?? einrichtung??
einzelnen?? einzig?? entweder erfindung erfindung?? erfindungsgemäss
erfindungsgemäss?? er erst?? erstreckend?? etwa etwas

f februar fig fig. figs figur?? freitag fünf für fur fuer

g gebildet?? gebrauchsmuster? gegenstand gegenstands gegenüber gekennzeichnet
gemeinsam?? genannt?? gerät?? gezeigt?? gibt gm gt

h haben hat herstellen herstellung herstellungen hier hierbei höchstens

i ihr?? in im insbesondere ist

j januar je jede? jedenfalls jedoch jene? jeweilig?? jeweils juli juni

k kabushiki kaisha kk kann kein?? kennzeichen können können

l

m märz mai mehrer?? mindestens mit miteinander mittel mitteln mittwoch montag

n nach neu?? nicht noch november null nur

o ob oben obgleich obschon obwohl oder ohne oktober

p patent? patentanspruch?? patentansprüche ps

q

r

s samstag sehr sein?? september sich sind sie so sobald solange solch?? soll
somit sonntag sowie system??

t trotzdem

u über überdies um und und/oder ungeachtet ungefähr unten unter unteranspr
unteranspruch?? unteransprüche?

v variant?? verbessert?? verbesserung?? verfahren? verlag versehen?? verwenden
verwendung?? vier viert?? vom von voneinander?? vorgesehen?? vorrichtung??
vorzugsweise

w während wahrend waehrend was wass weder weiter?? welch?? wenn werden
wesentlich?? weshalb wie wird werden wobei wurde wurden würde würden

x

y

z ziemlich zu zueinander?? zugehörig?? zugeordnet?? zugewandt?? zum zumindest
zur zulässig?? zwar zwei zweit?? zwischen

7.4.3. French Stopwords

a ainsi alors amélioration? amélioré?? août appareil? après au auquel autre?
aux aussi avec avoir avril ayant

b bien brevet?

c ce capable? caractérisé?? ceci cela celle? celui cependant ces cet cette?
ceux chaque? cinq comme comprenant

d dans décembre des desquel??? décrit?? des deux deuxième? dimanche dispositif?
divers?? donc dont du duquel

e elle? en entre environ est et et/ou etant été être exemple?

f fabrication faire fait?? février font fig fig. figs figure?

g

h

i

j janvier jeudi juin juillet

k kabushiki kaisha kk

l la ladite laquelle le ledit lequel les lesdit?? lequel??? leur? lundi

m machine mai mais mardi mars mercredi méthode? mettant mettre mode? moyen?

n na no nohe norm novembre ne ni notamment nouvel??? nouveau nouveaux

o obtenu?? octobre oeuvre? ont ou

p par parmi particulier patent perfection??? perfectionnement? peut peuvent
plus possible? pour préférence? première? procédé? puis précédent??

q qu quatre quatrième? que quel? quelconque quelle? qui

r relatif? relative? relativement revendication? réalisation? réalisé??

s sa samedi se septembre sera servant selon seront son ses si soit sont
supérieur?? sur système?

t tel? telle? tres troisième? tous tout?? trois

u un une utilisant utilisation utilis??

v va vendredi vient viennent vont

w

x

y

z

7.5. ACKNOWLEDGEMENTS

I would like to thank the following people for their support and input:

M.Keita, R. Feldhoff, W. de Paepe, C. Felicetti, C. Albrecht, S. Heising and Directorate 1.5.57

I would also like to thank DG2 for their assistance in providing a central location to host the dictionaries produced during this process.