AutoLearn<word>

Kurt Eberle
Lingenio GmbH
Karlsruher Straße 10
69126 Heidelberg
Germany
k.eberle@lingenio.de

Abstract

AutoLearn<word> extracts new translation relations for words and multiword expressions (MWE) of any category from bilingual texts of any size in high quality and prepares the information found as a conventional dictionary entry - with morpho-syntactic and semantic classifications and contextual use conditions.
The function uses Lingenio's MT-system and analysis components as knowledge source, integrates its results into these and, by this bootstrapping approach, adapts dictionary and MT to the needs of the customer. Manual intervention is restricted to a very reduced number of difficult cases and can be carried out easily in an ergonomic graphical user interface, without need of effortful training. This is enabled by the underlying MT-architecture with rule-based core and additional statistical features.
The use conditions connected to the new dictionary entries are derived from the local representation the considered word or expression is part of in the considered reference(s). They restrict the corresponding translation to similar cases so that interferences with other translations in the dictionary are avoided.
A basic version of the function is already available in the current version of Lingenio's *translate*.

1 Motivation

Unknown words and lack of information about translations of words and expressions are notorious problems of Machine Translation (MT). The problems differ however depending on the type of MT system considered and the coverage of the lexicon aimed at: In a statisticial machine translation system (SMT), 'unknown' means that the form is not listed in the frequency lexicon of the source language and the translation model doesn't show corresponding elements in the phrase tables. In a so-called rule based Machine Translation system (RBMT), 'unknown' means that for a form detected in the source text information about lemmatization, morphological classification and about syntactic and semantic properties is missing. With respect to translation, lack of information may mean that besides the pure (weighted) transfer relations use conditions are missing or unsufficient, such that a contextually justified selection of the target form cannot be made. For hybrid MT systems the problems may vary depending on the type of system considered.

Vocabulary and translation of expressions may change significantly when considering different text topics and domains. No system can provide constant translation quality for any text. Therefore, terminology extraction from (monolingual and bilingual) texts and corpora is a very important prerequisite of customization and effective professional translation. Corresponding research and development have a rather long tradition.
As the purposes of extraction are many and the environments an extraction tool shall be part of different (consisting of, or integrating, different CAT-tools and MT-systems respectively), the research area covers a wide range of subtasks, questions and approaches to solutions. Main topics and distinctions are: extraction of references for search terms in monolingual versus bilingual texts, corresponding monolingual versus bilingual search of new terms, where 'term'

may refer to single words or, in contrast or additionally, to multiword expressions and to specific categories (mostly noun) or alternatively to expressions of any category. Search may be carried out on the basis of 'knowledge-rich' or 'knowledge-poor systems' where 'knowledge' may be statistical or 'rule-based' or hybrid. Well known commercial tools in the area of bilingual extraction for CAT and/or MT respectively are, or are included in: *Systran Business Translator*, *ProMT Professional*, *Across Personal Edition* (*crossTerm*), *SDL Trados* (*MultiTerm*). Examples of (free) web-based monolingual term recognition and extraction tools are *FiveFilters* and TermMine. 'Concordancer' like AntConc find reference sentence pairs for glossary items.

An older but still valid overview of the architectural attempts made in the field is (Thurmair 2003). A good selection of research papers can be found at the homepage of the EU FP7 project *Terminology Extraction, Translation Tools and Comparable Corpora* (*TTC*) which terminated recently.

Some of the systems available combine linguistic analysis and statistical filtering, in particular if they don't only extract references, but suggest dictionary entries. However, if so they typically confine to subtypes of NPs only (in English typically: adjective-noun, noun-noun, noun-PP-combinations). Suggestion of words and MWEs for any category and adding transfer conditions is seldom. AutoLearn<word> is such a system.

2 Background

The background of AutoLearn<word> is a hybrid MT architecture with rule-based core and statistical features (cf. Babych et al. 2012). The rule-based core traces back to the longtime IBM R&D project *Logic Based Machine Translation* (*LMT*, McCord 1989b) which had been initiated in the late 80s at IBM research. The systems resulting from this have been developed further and extended by Lingenio GmbH, which is an IBM spin-off from 1999.

The architecture follows the *transfer* approach: It segments the text, assigns analyses to the segments (sentences), translates these and generates the target from the target structures. The analysis system is built on slot grammar (McCord 1989a), a unification-based dependency grammar. Slot grammar is also used in deep linguistic processing of IBM Watson (McCord et al 2012).

From the slot grammar analysis of a sentence the system constructs a shallow semantic representation, called a *dependency tree*. These trees are compact encodings of so-called *flat underspecified discourse representation structures* (*FUDRSs*). FUDRSs are the representation items of flat underspecified discourse representation theory (FUDRT; cf. Eberle 2002, 2004), which is an extension of (underspecified) *Discourse Representation Theory* (*DRT* / UDRT; cf. Kamp 1981, Kamp/Reyle 1993, Reyle 1993). The trees represent the predicate argument structure(s) of the sentence as can be derived from the syntactic structure and from ontological knowledge about the words. They are compact as they don't spell out the semantics of the lexical elements, but only point to them, i.e. they aren't the FUDRS of a sentence but determine its construction. FUDRSs allow to include additional information, in particular pragmatic information from the discourse history and from inter-sentential links, and to use this for stepwise disambiguation of lexical and structural ambiguities, if needed. The tree encoding can account for this.

The dependency trees are translated recursively into corresponding structures of the target language using the knowledge of the bilingual dictionary. From these 'deep structures' of the target side slot grammar structures are generated and from these, finally, the target sentence. (In the older LMT systems the shallow semantic level was missing and transfer happened at the level of slot grammar analysis).

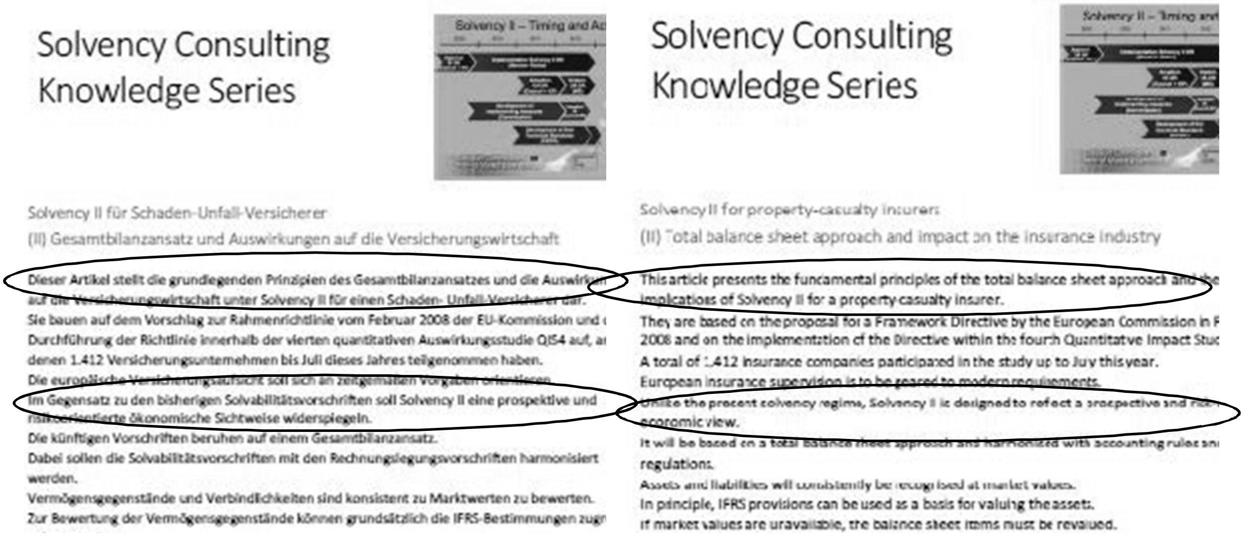
The advantage of the representation for the purpose of AutoLearn<word> is that it abstracts from the syntactic details of the surface structure and concentrates on the representation of the events and states of the sentence and the corresponding roles without going too much into detail with semantic specification. As on this level source representation and target representation are typically more similar to each other than on the syntactic level, it is more appropriate for the extraction of lexical source-target correspondences than this.

Note however that the details of the more 'lower' level analyses are not lost; they can be used by transfer and generation as they are connected to the 'higher' levels through corresponding links (so that, for instance, calculation of the surface order of the target sentence may use the shaping of the source sentence as an additional knowledge resource).

3 Algorithm

After aligning the sentences of a bilingual text, AutoLearn<word> carries out structural analyses of the corresponding source-target sentence pairs and then uses the translation knowledge of the MT system to relate words and expressions of the respective source structure and the target structure to each other.

The example that we will use is from the *solvency consulting knowledge series* (cf. Munich Re), which is about European solvency law.



Solvency Consulting
Knowledge Series

Solvency II für Schaden-Unfall-Versicherer
(II) Gesamtbilanzansatz und Auswirkungen auf die Versicherungswirtschaft

Dieser Artikel stellt die grundlegenden Prinzipien des Gesamtbilanzansatzes und die Auswirkun
auf die Versicherungswirtschaft unter Solvency II für einen Schaden- Unfall-Versicherer dar.
Sie bauen auf dem Vorschlag zur Rahmenrichtlinie vom Februar 2008 der EU-Kommission und d
Durchführung der Richtlinie innerhalb der vierten quantitativen Auswirkungsstudie QIS4 auf, an
denen 1.412 Versicherungsunternehmen bis Juli dieses Jahres teilgenommen haben.
Die europäische Versicherungsaufsicht soll sich an zeitgemäßen Vorgaben orientieren.
Im Gegensatz zu den bisherigen Solvabilitätsvorschriften soll Solvency II eine prospektive und
risikoorientierte ökonomische Sichtweise widerspiegeln.
Die künftigen Vorschriften beruhen auf einem Gesamtbilanzansatz.
Dabei sollen die Solvabilitätsvorschriften mit den Rechnungslegungsvorschriften harmonisiert
werden.
Vermögensgegenstände und Verbindlichkeiten sind konsistent zu Marktwerten zu bewerten.
Zur Bewertung der Vermögensgegenstände können grundsätzlich die IFRS-Bestimmungen zugr

Solvency Consulting
Knowledge Series

Solvency II for property-casualty insurers
(II) Total balance sheet approach and impact on the insurance industry

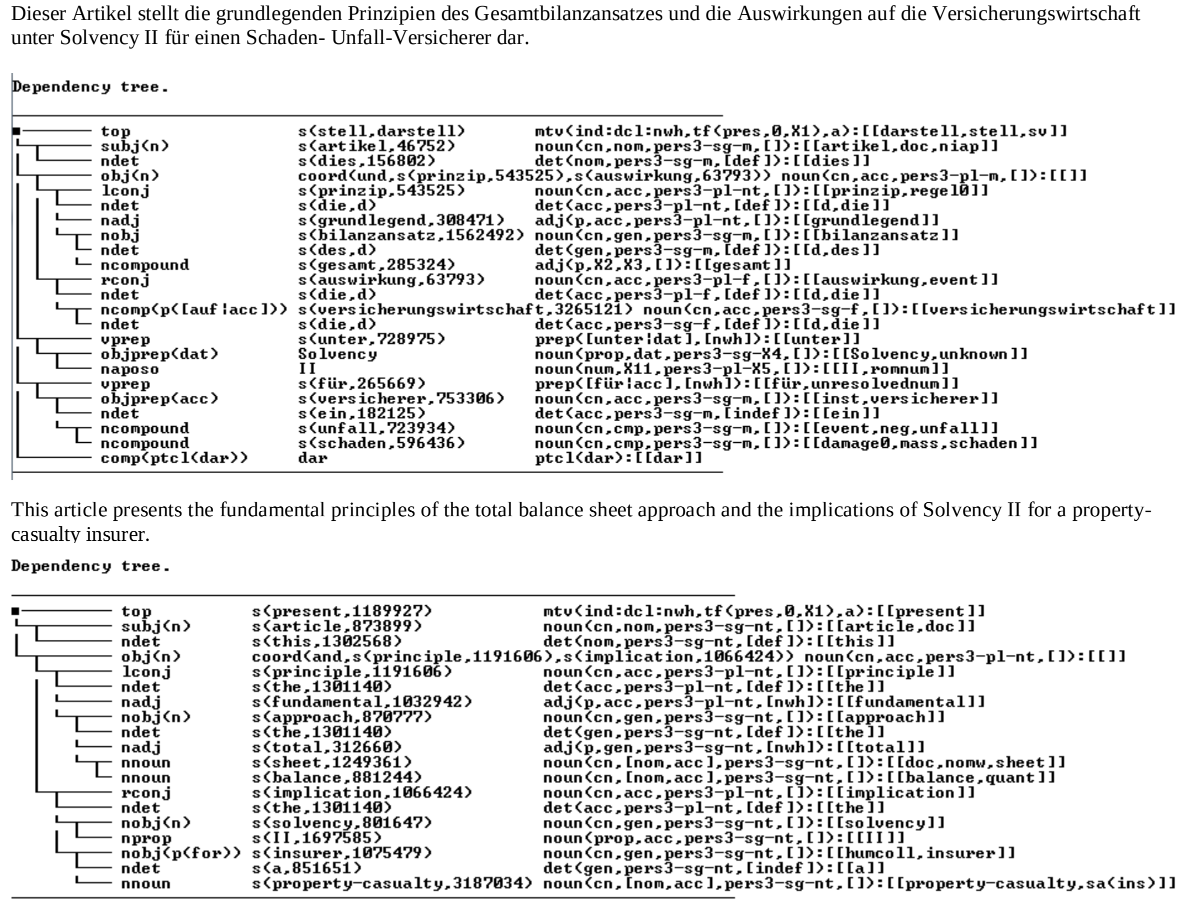
This article presents the funcamental principles of the total balance sheet approach and the
implications of Solvency II for a property-casualty insurer.
They are based on the proposal for a Framework Directive by the European Commission in F
2008 and on the implementation of the Directive within the fourth Quantitatve Impact Stud
A total of 1.412 insurance companies participated in the study up to July this year.
European insurance supervision is to be geared to modern requirements.
Unlike the present solvency regime, Solvency II is designed to reflect a prospective and risk-
economic view.
It will be based on a total balance sheet approach and harmonized with accounting rules and
regulations.
Assets and liabilities will consistently be recognised at market values.
In principle, IFRS provisions can be used as a basis for valuing the assets.
If market values are unavailable, the balance sheet items must be revalued.

Fig. 1 Lingenio sentence aligner applied to bilingual text

Fig. 1) shows parts of the texts and highlights two items of the output of the sentence aligner.

3.1 Dependency tree representations

Fig2. shows the dependency trees for source sentence and target sentence of the first pair of sentences in fig.1 marked in red. The picture rotates the dependency trees by 90 degrees to the left, such that the top is on left and the leaf information on right, and the different information types are structured in a column-like form where the first column represents the dependency relations between the lexical items and the second and the third the corresponding semantic and morpho-syntactic reading of these items. The trees neglect the order of the lexical items in the sentence. (As said before, this information can be accessed, if necessary). Expressions of the form *s(Stem,ID)* in the second column are called *senses*. The first argument represents the homonym stem and the second the lexical identifier of the corresponding reading via which the respective semantic interpretation in the sense of DRT or predicate logic can be constructed, if needed (for instance in order to represent temporal and referential links of the discourse history). For the purposes of AutoLearn<word>, mostly the shallow ontological classifications represented in the dependency tree will do: *s(artikel,46752)* in the second column of the second row of the first tree in fig. 2 stands for the reading of the word *Artikel* as an *article of a document* like a *newspaper article* - in contrast to the reading as an *item of a product portfolio* - and consequently it is classified as a *doc*, i.e. a document, in the third *column*. This column combines the morpho-syntactic features as assigned to the item in the sentence representation selected and the ontological classification available for it (if any) in the form *Feas*:*Types*, where 'Types' arranges the available semantic supertypes of the word reading in disjunctive normal form as a list of lists. (Typically, a word reading is classified by one disjunct only).



Dieser Artikel stellt die grundlegenden Prinzipien des Gesamtbilanzansatzes und die Auswirkungen auf die Versicherungswirtschaft unter Solvency II für einen Schaden- Unfall-Versicherer dar.

```
Dependency tree.

■──────── top                   s(stell,darstell)       mtv(ind:dcl:nwh,tf(pres,0,X1),a):[[darstell,stell,sv]]
└─┬────── subj(n)               s(artikel,46752)        noun(cn,nom,pers3-sg-m,[]):[[artikel,doc,niap]]
│ └────── ndet                  s(dies,156802)          det(nom,pers3-sg-m,[def]):[[dies]]
└─┬────── obj(n)                coord(und,s(prinzip,543525),s(auswirkung,63793)) noun(cn,acc,pers3-pl-m,[]):[[]]
│ └─┬──── lconj                 s(prinzip,543525)       noun(cn,acc,pers3-pl-nt,[]):[[prinzip,regel0]]
│ │ └──── ndet                  s(die,d)                det(acc,pers3-pl-nt,[def]):[[d,die]]
│ │ └──── nadj                  s(grundlegend,308471)   adj(p,acc,pers3-pl-nt,[]):[[grundlegend]]
│ │ └─┬── nobj                  s(bilanzansatz,1562492) noun(cn,gen,pers3-sg-m,[]):[[bilanzansatz]]
│ │   └── ndet                  s(des,d)                det(gen,pers3-sg-m,[def]):[[d,des]]
│ │   └── ncompound             s(gesamt,285324)        adj(p,X2,X3,[]):[[gesamt]]
│ └─┬──── rconj                 s(auswirkung,63793)     noun(cn,acc,pers3-pl-f,[]):[[auswirkung,event]]
│   └──── ndet                  s(die,d)                det(acc,pers3-pl-f,[def]):[[d,die]]
│   └─┬── ncomp(p([auf|acc]))   s(versicherungswirtschaft,3265121) noun(cn,acc,pers3-sg-f,[]):[[versicherungswirtschaft]]
│     └── ndet                  s(die,d)                det(acc,pers3-sg-f,[def]):[[d,die]]
└─┬────── vprep                 s(unter,728975)         prep([unter|dat],[nwh]):[[unter]]
│ └─┬──── objprep(dat)          Solvency                noun(prop,dat,pers3-sg-X4,[]):[[Solvency,unknown]]
│   └──── naposo                II                      noun(num,X11,pers3-pl-X5,[]):[[II,romnum]]
└─┬────── vprep                 s(für,265669)           prep([für|acc],[nwh]):[[für,unresolvednum]]
│ └─┬──── objprep(acc)          s(versicherer,753306)   noun(cn,acc,pers3-sg-m,[]):[[inst,versicherer]]
│   └──── ndet                  s(ein,182125)           det(acc,pers3-sg-m,[indef]):[[ein]]
│   └─┬── ncompound             s(unfall,723934)        noun(cn,cmp,pers3-sg-m,[]):[[event,neg,unfall]]
│     └── ncompound             s(schaden,596436)       noun(cn,cmp,pers3-sg-m,[]):[[damage0,mass,schaden]]
└──────── comp(ptcl(dar))       dar                     ptcl(dar):[[dar]]
```

This article presents the fundamental principles of the total balance sheet approach and the implications of Solvency II for a property-casualty insurer.

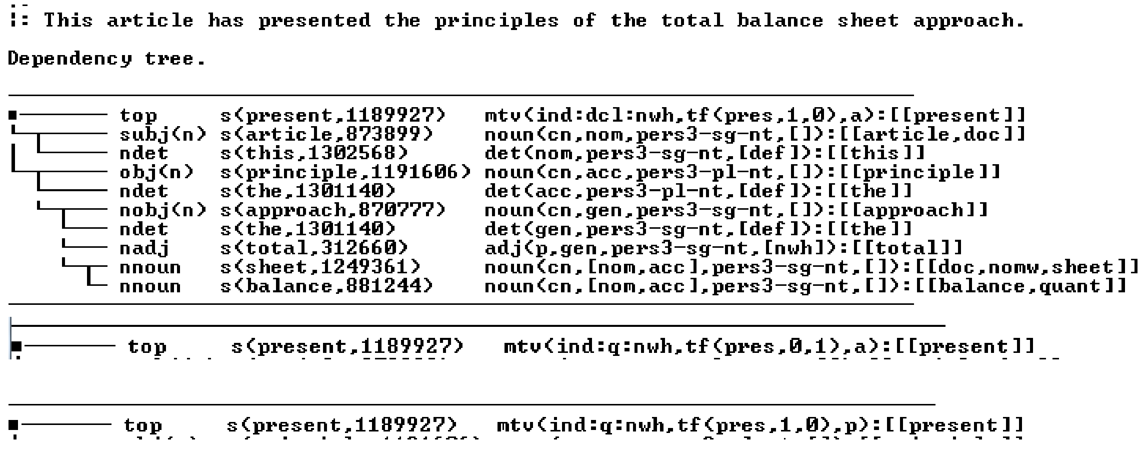
```
Dependency tree.

■────────── top            s(present,1189927)      mtv(ind:dcl:nwh,tf(pres,0,X1),a):[[present]]
└─┬──────── subj(n)        s(article,873899)       noun(cn,nom,pers3-sg-nt,[]):[[article,doc]]
│ └──────── ndet           s(this,1302568)         det(nom,pers3-sg-nt,[def]):[[this]]
└─┬──────── obj(n)         coord(and,s(principle,1191606),s(implication,1066424)) noun(cn,acc,pers3-pl-nt,[]):[[]]
  └─┬────── lconj          s(principle,1191606)    noun(cn,acc,pers3-pl-nt,[]):[[principle]]
  │ └────── ndet           s(the,1301140)          det(acc,pers3-pl-nt,[def]):[[the]]
  │ └────── nadj           s(fundamental,1032942)  adj(p,acc,pers3-pl-nt,[nwh]):[[fundamental]]
  │ └─┬──── nobj(n)        s(approach,870777)      noun(cn,gen,pers3-sg-nt,[]):[[approach]]
  │   └──── ndet           s(the,1301140)          det(gen,pers3-sg-nt,[def]):[[the]]
  │   └──── nadj           s(total,312660)         adj(p,gen,pers3-sg-nt,[nwh]):[[total]]
  │   └─┬── nnoun          s(sheet,1249361)        noun(cn,[nom,acc],pers3-sg-nt,[]):[[doc,nomw,sheet]]
  │     └── nnoun          s(balance,881244)       noun(cn,[nom,acc],pers3-sg-nt,[]):[[balance,quant]]
  └─┬────── rconj          s(implication,1066424)  noun(cn,acc,pers3-pl-nt,[]):[[implication]]
    └────── ndet           s(the,1301140)          det(acc,pers3-pl-nt,[def]):[[the]]
    └─┬──── nobj(n)        s(solvency,801647)      noun(cn,gen,pers3-sg-nt,[]):[[solvency]]
    │ └──── nprop          s(II,1697585)           noun(prop,acc,pers3-sg-nt,[]):[[II]]
    └─┬──── nobj(p(for))   s(insurer,1075479)      noun(cn,gen,pers3-sg-nt,[]):[[humcoll,insurer]]
      └──── ndet           s(a,851651)             det(gen,pers3-sg-nt,[indef]):[[a]]
      └──── nnoun          s(property-casualty,3187034) noun(cn,[nom,acc],pers3-sg-nt,[]):[[property-casualty,sa(ins)]]
```

Fig.2 Dependency tree analyses of a pair of sentences from the bilingual text

Though the dependency trees of fig. 2 resemble very much slot grammar representations of the syntactic level because the compact representation format chosen 'hides' the proper semantic contribution of the lexical items, they nevertheless significantly abstract from syntactic details. For instance, consider the core sentence of the (first conjunct of the) English sentence of the example, (1.a) below. (1.b) - (1.d) are variants which develop from (1.a) by using other mood and/or other tense forms respectively, including forms with auxiliary constructions.

(1.a) *This article presents the principles of the total balance sheet approach.*
(1.b) *This article has presented the principles of the total balance sheet approach.*
(1.c) *Does this article present the principles of the total balance sheet approach?*
(1.d) *Have the principles of the total balance sheet approach been presented by this article?*

Though syntactically different, all sentences are represented by structurally identical trees; the differences are represented by different features assigned to the verb node only. There is no other difference with respect to dependency structure.



```
: This article has presented the principles of the total balance sheet approach.

Dependency tree.

■──────── top      s(present,1189927)    mtv(ind:dcl:nwh,tf(pres,1,0),a):[[present]]
└─┬────── subj(n)  s(article,873899)     noun(cn,nom,pers3-sg-nt,[]):[[article,doc]]
│ └────── ndet     s(this,1302568)       det(nom,pers3-sg-nt,[def]):[[this]]
└─┬────── obj(n)   s(principle,1191606)  noun(cn,acc,pers3-pl-nt,[]):[[principle]]
  └────── ndet     s(the,1301140)        det(acc,pers3-pl-nt,[def]):[[the]]
  └─┬──── nobj(n)  s(approach,870777)    noun(cn,gen,pers3-sg-nt,[]):[[approach]]
    └──── ndet     s(the,1301140)        det(gen,pers3-sg-nt,[def]):[[the]]
    └──── nadj     s(total,312660)       adj(p,gen,pers3-sg-nt,[nwh]):[[total]]
    └─┬── nnoun    s(sheet,1249361)      noun(cn,[nom,acc],pers3-sg-nt,[]):[[doc,nomw,sheet]]
      └── nnoun    s(balance,881244)     noun(cn,[nom,acc],pers3-sg-nt,[]):[[balance,quant]]

■──────── top      s(present,1189927)    mtv(ind:q:nwh,tf(pres,0,1),a):[[present]]

■──────── top      s(present,1189927)    mtv(ind:q:nwh,tf(pres,1,0),p):[[present]]
```

Fig 3. Dependency tree of sentence (1.b) and verb node representations of (1.c) and (1.d)

Fig. 3 gives the representation of (1.b) and screenshots of the verb rows of the representations of (1.c) and (1.d). The features of (1.a) are as in fig. 2.

The Features of verb nodes in dependency trees represent mood, tense and voice (*mtv*). *ind:dcl:nwh* means that the node is not dependent on something else (*ind*)- i.e. it stems from the main clause, is introduced by a declarative sentence (*dcl*) which doesn't come with a *wh*-element (as relative clauses and wh-questions do). In contrast: *q* indicates question (*q:nwh* yes/no-questions, *q:wh* wh-question). The representation of the meaning of tenses used here follows the three-dimensional analysis suggested in (Kamp/Rohrer 1985), also used in (Kamp/Reyle 1993) and in many contributions to tense meaning since then. According to this, the tense forms are analyzed into *tense level* (*pres, past, fut*), *perfectivity* (*yes/no* (*1/0*)), and *progressivity* (*yes/no* (*1/0*)). For the voice we use: *active* (*a*), *passive* (*p*) and *resultative passive* (*rp*).
This is sufficient to describe the differences between (1.a)-(1.d) on the semantic level. When constructing a DRS or FUDRS, this information would be expanded to different temporal relations between the event, the speech time and contextual perspective time and reference time.

Dependency trees don't always avoid 'unfolding' of the representation as they do here. Parts of the representation may be 'multiplied out', if this helps to make the source tree and the target tree more similar and supports structure-preserving translation therefore. An example is a sentence pair like '*the experiment was difficult and unsuccessful in the end – das Experiment war schwierig und missglückte am Ende*'. Here, '*be*' in the English sentence is used as a copula in a predicative structure and obtains different translations depending on the respective conjunct of the coordination which fills the predicative argument of the structure. The factored form with the coordination raised to VP eases translation a lot, because now there are two occurrences of predicate structures with '*be*' (… *was difficult and was unsuccessful* …) that can easily be translated separately. Other types of coordination and (some types of) ellipsis present further examples. As, nevertheless, these phenomena are not very frequent, 'unfolding' is carried out only if triggered by needs in transfer. (For a more detailed description of this compare Eberle 2003). Another feature of dependency trees is that they can 'underspecify' dependency relations, if the context doesn't provide enough information about how some expression shall be related to the context. Frequently this is the case with prepositional phrases: In the example of fig. 2, *for a property-casualty insurer* may relate to the second conjunct of the coordination only or, in contrast, may relate to the entire coordination (i.e. to both conjuncts). Depending on the options selected in the analysis system, the representation may disambiguate such ambiguities (as is the case in fig. 2) or not. We don't go into detail with this too (see Eberle et al. 2009 for a more elaborate description).

3.2 Relations between dependency trees

AutoLearn<word> uses different types of relational knowledge of the system in order to find correspondences between nodes of the source dependency tree and nodes of the target tree. The strategy is as follows: First, correspondences are searched that provide 'anchors' of the transfer relation, i.e. correspondences that are safe knowledge; then, on the basis of this, additional correspondences are searched that represent likely correspondences. This is done using several levels of reliability. On the basis of the resulting bi-structural knowledge, the system extracts those relations found which are not yet covered by the dictionary of the system. The set of these is suggested to the user for (automatic) integration in user dictionaries.

The relational knowledge used by the system in the first step consists of:
a) the translation relations between the lexical elements of the system dictionary (and already existing user dictionaries respectively) and
b) formal similarity between source and target expressions if the corresponding source or target item is unknown to the system (names for instance).

As words (and translations) may be used several times in the pair of segments considered, the anchors of the first step must satisfy to a number of uniqueness constraints to be sure that they represent fixed points of a translation.

Starting out from the fixed points the subsequent steps of the algorithm use similarity considerations about the dependency relations connected to the anchor node pairs of the trees in order to recursively apply the search procedure of the first step to the remaining nodes. As the search spaces shrink step by step in this procedure, uniqueness constraints may be satisfied in subsequent steps that are not in preceding steps. At the same time the reliability conditions may be reduced step by step. The algorithm uses a number of parameters for regulating the behavior of the algorithm in this respect.

The options include to trigger computation of factored representations for parts of the dependency tree or to underspecify the trees further, if seems reasonable. Another option is to integrate information from statistical word alignment as an additional knowledge resource of the computation of the bi-structural relations.

3.3 Candidates for new dictionary items

From the cascaded procedure in 3.2 we obtain node relations and with this a coverage of source- and target sentence by pairs of partial structures whose quality reflects the information as available to the system. As, through this, the pairing doesn't relate to nodes only but is extended to the local substructures connected to the nodes, the suggestion of new lexical relations can go beyond single words and can specify multiword relations also and even larger phrasal relations if desired. The new relations include appropriate linguistic annotations as they can reuse the morpho-syntactic and semantic properties of the senses they refer to in the trees. In addition, through the cross-language links morpho-syntactic and even semantic information can be inferred for words which, before, have been unknown to the system.

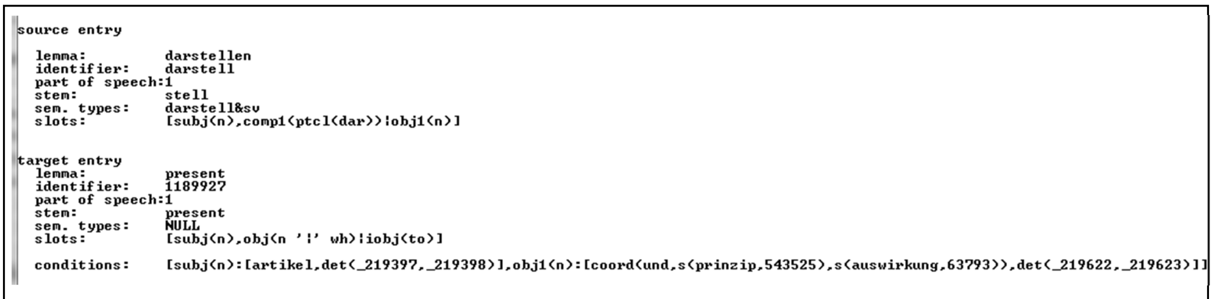
On the basis of the system dictionary used in the Lingenio product *translate pro* version 12.1 (the current version) the system suggests 6 new relations for the sentence pair of fig. 4:

	German	English
S1	*darstellen*	*to present*
S2	*Gesamtbilanzansatz*	*total balance sheet approach*
S3	*Solvency*	*solvency*
S4	*II*	*II*
S5	*Schaden-Unfall*	*property-casuality*
S6	*Schaden-Unfall-Versicherer*	*property-casuality insurer*

S1 relates single words to each other where both words are known to the system, but not the relation between them (a gap in the system dictionary). S2, S5 and S6 relate MWEs where neither the source expression nor the target expression is known to the system, but only (the) leafs of the corresponding structures. Note that the respective internal structures are not pairwise homomorphic (see fig. 2) and therefore illustrate different subcases of the multiword suggestion type of AutoLearn<word>. S3 and S4 again relate single words to each other, but here the source word is unknown to the system in the one case and the target in the same category in the other (cf. fig. 2). When including the suggestions into the dictionary the derivable information is added: DE-Solvency is treated as a common noun like EN-solvency and the MWEs obtain standardized representations concerning the parts they consist of. Of course, the user can edit the entries and correct or add supplementary information.

4 Extensions

S1 is specific as it relates expressions to each other which are familiar to the system and therefore already obtain translations in the dictionary. This means that the new relation is unsatisfactory as, until now, it doesn't define under which circumstances the new relation shall be used in contrast to those already known by the system. The availability of local context helps to define corresponding conditions: The system knows the respective subcategorization frames and can infer from the dependency trees how the slots have been filled. We can obtain the more complete suggestion of fig. 4 for S1 therefore:



```
source entry

  lemma:          darstellen
  identifier:     darstell
  part of speech:1
  stem:           stell
  sem. types:     darstell&sv
  slots:          [subj(n),comp1(ptcl(dar))¦obj1(n)]


target entry
  lemma:          present
  identifier:     1189927
  part of speech:1
  stem:           present
  sem. types:     NULL
  slots:          [subj(n),obj(n '¦' wh)¦iobj(to)]

  conditions:     [subj(n):[artikel,det(_219397,_219398)],obj1(n):[coord(und,s(prinzip,543525),s(auswirkung,63793)),det(_219622,_219623)]]
```

Fig 4. Suggestion of a new dictionary entry extended by use conditions

In fig. 4 the translation *to present* is restricted to the case where the subject is headed by *Artikel* and the direct object is filled by a conjunction made up by NPs headed by *Prinzip* and *Auswirkung*, as in the dependency tree of fig.2.

The restriction of fig. 4 omits details of the complements in the dependency tree: neither does the determiner condition, *det(Var1,Var2)*, specify whether the NPs should be *sg, pl, definite* or *indefinite*, nor are there included representations of modifiers (of the complements or the verb). It depends on the parameter setting whether a more restrictive extraction of conditions is selected or a more general one or a mixture of both. Typically, the system assigns conditions that relate to the ontological classifications of the words instead of the words themselves. In the example of fig. 2 this means to weaken the conditions to stipulating that the subject be of type *doc* (*document*) and the direct object of type *regel0* (*rule*) or *event*.

Currently, the system is trained on large corpora in this respect, in order to optimize the conditions of the dictionary. This is joint work of the university of Leeds and Lingenio funded by the EU project HyghTra (http://www.hyghtra.eu). Quantitative results are on the way, but couldn't made available yet. As the method is of type 'knowledge-rich' its reliability is very high, if validity and coverage of the underlying resources are, as is the case for the Lingenio analysis systems.

5 Conclusion

We described an algorithm for extracting new dictionary entries from bilingual corpora. It considers words and multiword expressions of any category, where 'multiword' is any linguistically closed structure whose consideration is justified by the administrator of the system - who may include phrasal templates too as well as statistical collocation information, if desired.

The new relations are annotated by morpho-syntactic and semantic classifications and by use conditions as can be derived from the system knowledge and the context in the sentence. The use conditions specify the circumstances under which the translation is triggered.

A basic version of AutoLearn<word> is already available in the current version of Lingenio's MT-series *translate*. A more comprehensive version with features as described here will be integrated in the next product version, which is currently worked out.

Acknowledgement

This research is supported by a Marie Curie IAPP project taking place within the 7th European Community Framework Programme (Grant agreement no.: 251534), cf. http://www.hyghtra.eu.

Literature

Bogdan, Babych, Kurt Eberle, Johanna Geiß, Mireia Ginestí-Rosell, Anthony Hartley, Reinhard Rapp, Serge Sharoff and Martin Thomas (2012): *Design of a Hybrid High Quality Machine Translation System.* Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), EACL 2012, Avignon

Kurt Eberle (2002): *Tense and Aspect Information in a FUDR-based German French Machine Translation System.* In: Hans Kamp, Uwe Reyle (Hrsg.), How we say WHEN it happens. Contributions to the theory of temporal reference in natural language, S. 97–148. Niemeyer, Tübingen, Ling. Arbeiten, Band 455.

Kurt Eberle (2003): C*oordination, incorporation and dynamic semantic representation in transfer.* In: Proceedings of the EAMT-CLAW conference, Dublin.

Kurt Eberle (2004): *Flat underspecified representation and its meaning for a fragment of German* (Habilitation) Universität Stuttgart.

Kurt Eberle, Gertrud Faaß, Ulrich Heid (2009): *Corpus-based identification and disambiguation of reading indicators for German nominalizations*. In: Proceedings of the fifth Corpus Linguistics Conference, Liverpool.

Hans Kamp (1981): *A Theory of Truth and Semantic Representation*. In: J.A.G. Groenendijk, T.M.V. Janssen and M.B.J. Stokhof: Formal Methods in the Study of Language, Mathematical Centre Tract, Amsterdam.

Hans Kamp and Uwe Reyle (1993): *From Discourse to Logic*, Kluwer Academic Publishers, Dordrecht.

Hans Kamp and Christian Rohrer (1985): *Temporal Reference in French*, (ms), University of Stuttgart.

Michael C. McCord (1989a): Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars. Natural Language and Logic 1989: 118-145

Michael C. McCord (1989b): *Design of LMT: A Prolog-Based Machine Translation System*. Computational Linguistics 15(1): 33-52 (1989)

Michael C. McCord, J. William Murdock, Branimir Boguraev (2012): Deep parsing in Watson. IBM Journal of Research and Development 56(3): 3

Munich Re: http://www.munichre.com/de/group/focus/solvency-ii/knowledge-series/index.html

Uwe Reyle (1993): *Dealing with ambiguities by underspecification: Construction, representation, and deduction*, Journal of Semantics 10(2), pp. 123-179

Gregor Thurmair (2003): *Making Term Extraction Tools Usable*, Proceedings of the EAMT-CLAW conference, Dublin.

TTC: http://www.ttc-project.eu/