ASS.I.TERM
ASSOCIAZIONE ITALIANA PER LA TERMINOLOGIA

XXX Convegno Associazione Italiana per la Terminologia
Risorse e strumenti per l'elaborazione
e la diffusione della terminologia

15-16 ottobre 2020

Terminologia basata su frame: la creazione di un *tagset* per l'annotazione semantica di corpora in ambito tecnico

Laura Giacomini

Università di Hildesheim
Istituto di Scienze
dell'informazione
e Linguistica computazionale

Università di Heidelberg
Istituto per Interpreti
e Traduttori

1. Un progetto terminologico basato su frame
2. La variazione terminologica
3. La definizione di un frame e dei suoi elementi
4. L’annotazione semantica basata su frame

1. Un progetto terminologico basato su frame

- Progetto terminologico in ambito tecnico (tesi di abilitazione all'Università di Hildesheim)

- Sviluppo di un metodo frame-based (v. Faber 2015) al fine di:
- modellizzare la variazione terminologica in un ambito tecnico,
- **estrarre varianti terminologiche da corpora annotati semanticamente,**
- rappresentare varianti in una banca dati terminologica / una risorsa lessicografica.

- Effetti del processo di validazione del metodo:
  - lingua tedesca → prospettiva multilingue
  - dominio dell'isolamento termico → ulteriori domini tecnici

2. La variazione terminologica

- Nozione di variazione:
  - le varianti sono coreferenti (sinonimi assoluti o parziali)
  - la variazione avviene allo stesso livello sistemico (ad es. non varianti diatopiche o diastratiche)

Tipo di variazione	Aspetti linguistici coinvolti	Esempi
Variazione morfologica (totale/ parziale)	Morfemi lessicali	pannello isolante – lastra isolante *ETICS – External Thermal Insulation Systems*
Variazione sintattica	Parte del discorso, ordine degli elementi di un composto, struttura sintagmatica	coibentare – coibentante termoisolante – isolante termico
Variazione ortografica	Uso di maiuscole/minuscole e trattini, allografi, variazione ad hoc	XPS – xps

Tipo di termini	Esempi
Termini semplici	*poliuretanico* *isolare* *massetto*
Termini complessi	*termoacustico* *vetroresina* *controsoffitto* *sottotetto*
Frasemi specialistici	*pannello isolante* *fibra di vetro* *cappotto termico* *isolamento termico a cappotto*

- Esempio di variazione in un testo tecnico informativo:

Cos'è una schiuma poliuretanica isolante espansa e come si ottiene

La **schiuma riempitiva poliuretanica** trova la sua origine proprio dal poliuretano; questo, altro non è che un polimero, ossia una macromolecola composta da più gruppi molecolari uniti tra di loro da un legame ripetuto (detto anche legame covalente). […]

La **schiuma espansa**, o **schiuma poliuretanica isolante**, deriva dal cosiddetto poliuretano espanso rigido, un polimero reticolato termoindurente. Questo, oltre che per la produzione di **schiume coibentanti**, viene sfruttato anche per l'isolamento di attrezzature ed impianti coinvolti nella catena del freddo (sia domestico che industriale/commerciale), oltre che per i mezzi di trasporto in regime di temperatura controllata e per la coibentazione delle lamine esterne usate per i canali di ventilazione. [...]

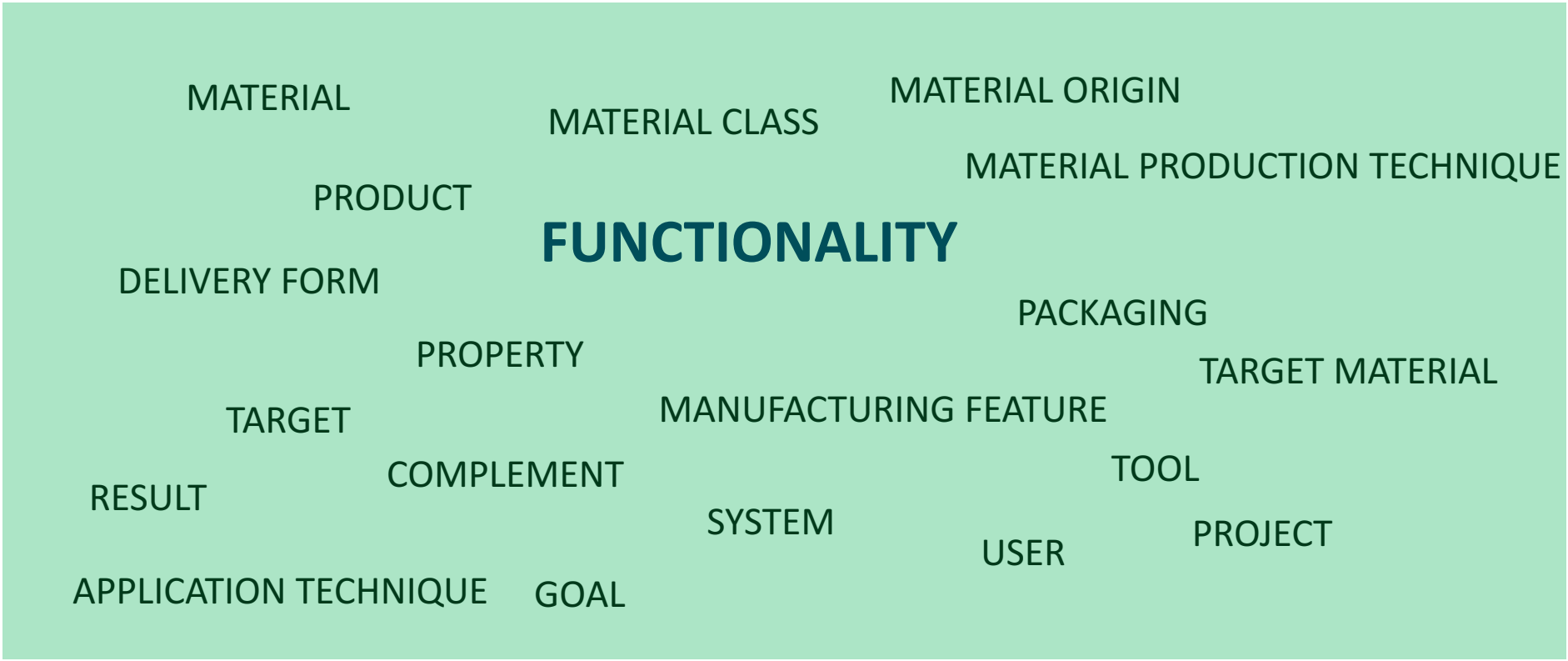
Non tutte le **schiume isolanti** sono uguali e consigliamo di affidarsi a persone competenti in materia per scegliere la schiuma isolante giusta per la coibentazione voluta. La **schiuma espansa poliuretanica** è un ottimo prodotto da preferire per isolare solai e sottotetti delle abitazioni in quanto è un prodotto prevalentemente a celle chiuse quindi ad alta densità che consente di essere calpestato e risulta impermeabile. [...]

https://www.coibentarecasa.it

3. La definizione di un frame e dei suoi elementi

- Frame (v. Fillmore & Baker 2010): struttura cognitiva che identifica specifici significati e strutture argomentali di una parola

- Esempio di frame riguardante il dominio degli isolanti termici:



- Analisi sintattica e semantica tramite elementi del frame:

pannello in polistirene estruso
N_{FORM} *in* N_{MAT} V_{MAT_TECH}

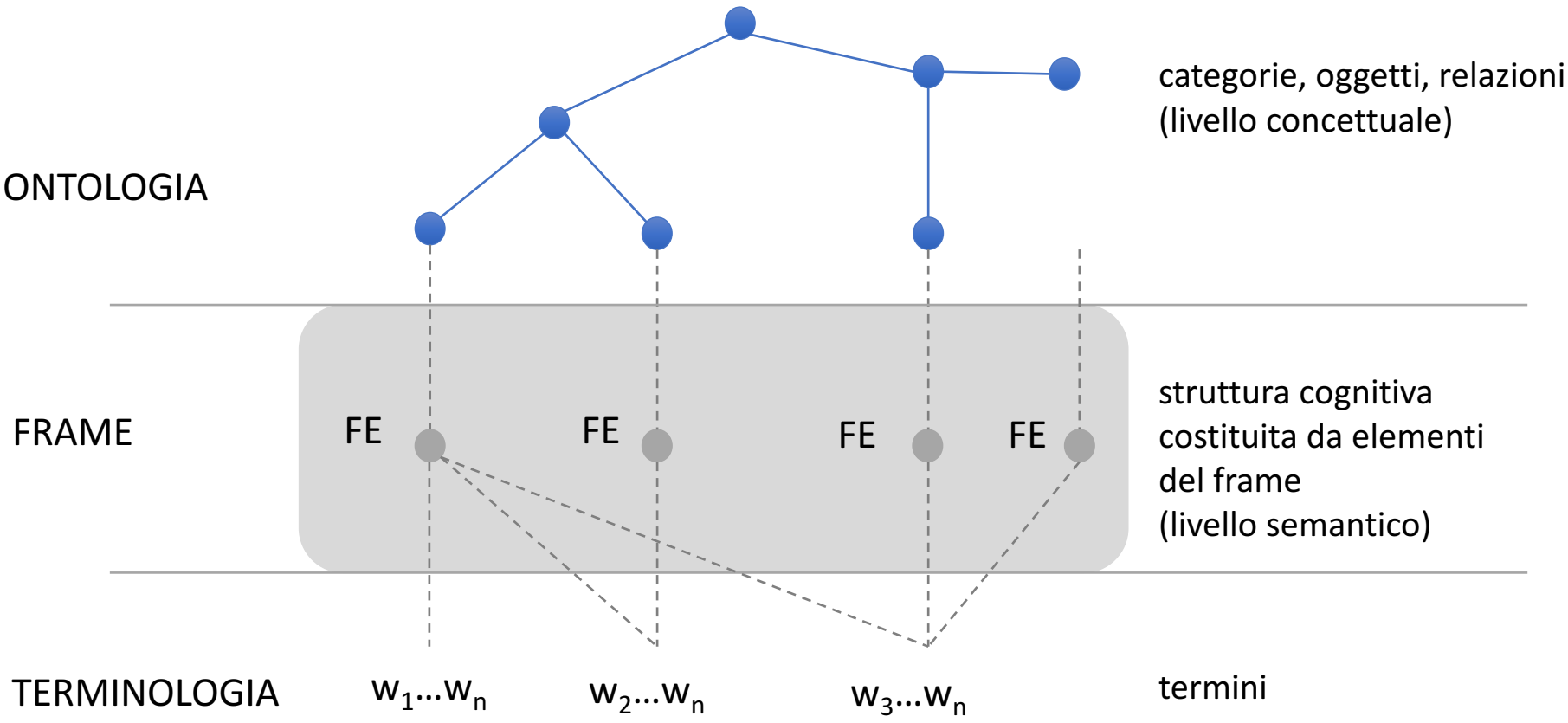
pannello isolante in polistirene estruso
N_{FORM} V_{GOAL} *in* N_{MAT} V_{MAT_TECH}

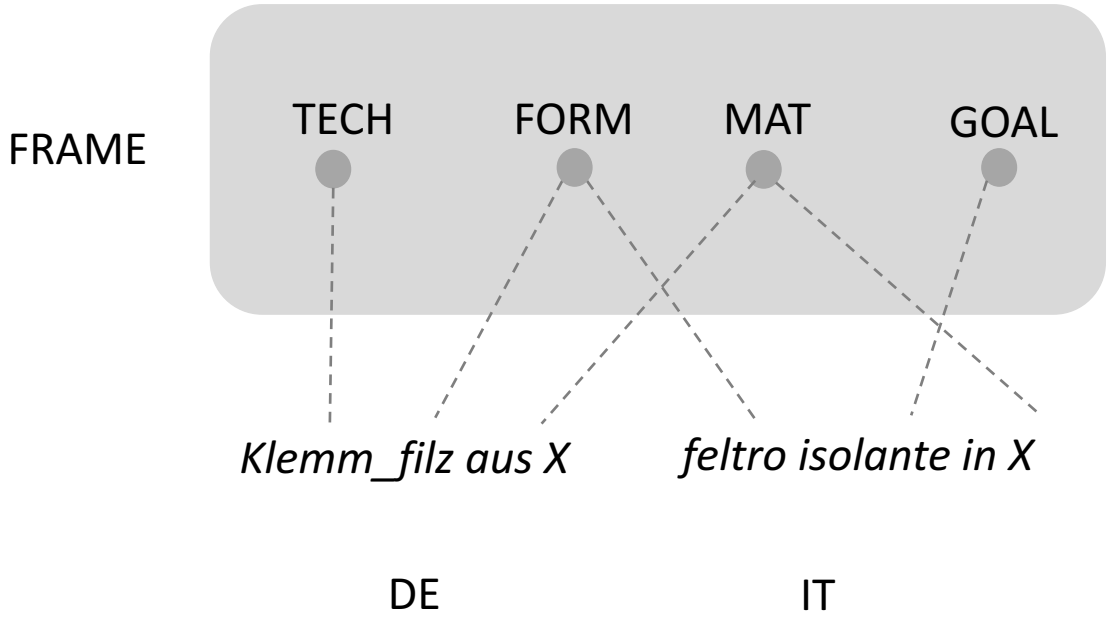
lastra isolante in polistirene estruso
N_{FORM} V_{GOAL} *in* N_{MAT} V_{MAT_TECH}

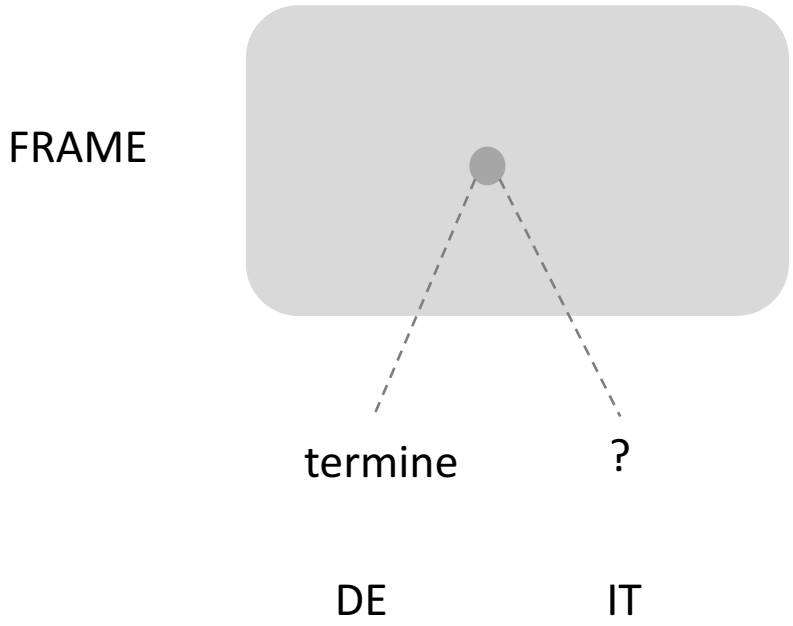
pannello isolante in polistirene espanso estruso
N_{FORM} V_{GOAL} *in* N_{MAT} V_{MAT_TECH} V_{MAT_TECH}

pannello isolante in schiuma di polistirene estruso
N_{FORM} V_{GOAL} *in* N_{MAT_CLASS} *di* N_{MAT} V_{MAT_TECH}

*pannello XPS** *(*extruded polystyrene)*
N_{FORM} (V_{MAT_TECH} N_{MAT})


ONTOLOGIA
categorie, oggetti, relazioni
(livello concettuale)
FRAME
FE
FE
FE
FE
struttura cognitiva
costituita da elementi
del frame
(livello semantico)
TERMINOLOGIA
w1…wn
w2…wn
w3…wn
termini


FRAME
TECH
FORM
MAT
GOAL
Klemm_filz aus X
feltro isolante in X
DE
IT

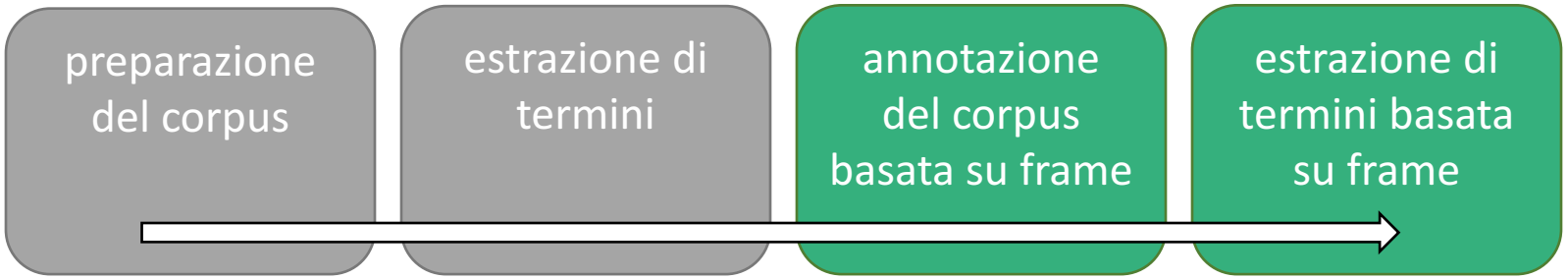

FRAME
termine
?
DE
IT

4. L’annotazione semantica basata su frame

4.1 Corpora utilizzati:

- Corpus originario: Insulation Corpus, testi online in lingua tedesca, 5.2 Mio di parole
- Corpus utilizzato per la validazione: testi comparabili in lingua italiana, 2 Mio di parole
- Composizione del corpus: testi tecnici o semi-tecnici da fonti specifiche (ad es. produttori, rivenditori specializzati, siti di informazione) e di generi specifici (riviste specializzate, manuali, descrizioni di prodotti, schede tecniche)

4.2 Procedimento:



1. Preparazione del corpus:

   include POS-tagging con RFTagger (Schmid & Laws, 2008) e codifica su Corpus Workbench (Evert & Hardie, 2011)

2. Estrazione di termini candidati (Schäfer et al. 2015)

3. Selezione di stringhe terminologiche (parole, radici e temi) rilevanti all'interno della lista dei candidati

4. Tagset iniziale: compilazione di un seed lexicon che associa stringhe terminologiche ad elementi del frame, ad es.

   MATERIAL: *legno, sughero, fibr-, …*

   MATERIAL ORIGIN: *natural-, vegetal-, origine, …*

   MATERIAL PRODUCTION TECHNIQUE: *espan-, estrus-, …*

   APPLICATION TECHNIQUE: *applic-, install-, insuffl-…*

   GOAL: *isol-, coibent-, …*

5. Annotazione automatica dei token che contengono stringhe del *seed lexicon*. Una stringa può essere potenzialmente associata a più di un tag, ad es. *legno* (elementi del frame: MATERIAL e TARGET MATERIAL)

6. Dato il set s di stringhe individuate in un termine, calcoliamo tutti i possibili profili di variazione, ad es.

pannello isolante sottovuoto

s={*pannell-*, *isol-*, *sottovuoto*}

con quattro differenti profili di variazione:

s_v1 = {{*pannell-*}, {*isol-*}, {*sottovuoto*}},
s_v2 = {{*pannell-*, *isol-*}, {*sottovuoto*}},
s_v3 = {{*sottovuoto*}, {*isol-*, *pannell-*}} e
s_v4 = {{*sottovuoto*, *pannell-*}, {*isol-*}}.

La variazione viene considerata nei limiti di un periodo.

4.3 Metodi utilizzati per identificare ulteriori varianti:

1. Restrizioni sintattiche
2. Nel calcolare i profili di variazione, lasciare una stringa come variabile, ad es.

s={*pannell-*, *isol-*, *sottovuoto*}

s_1={**FORM**, *isol-*, *sottovuoto*}

{*pannell-*, *lastr-*, *materassin-*, *rotol-*, …} ∈ seed lexicon

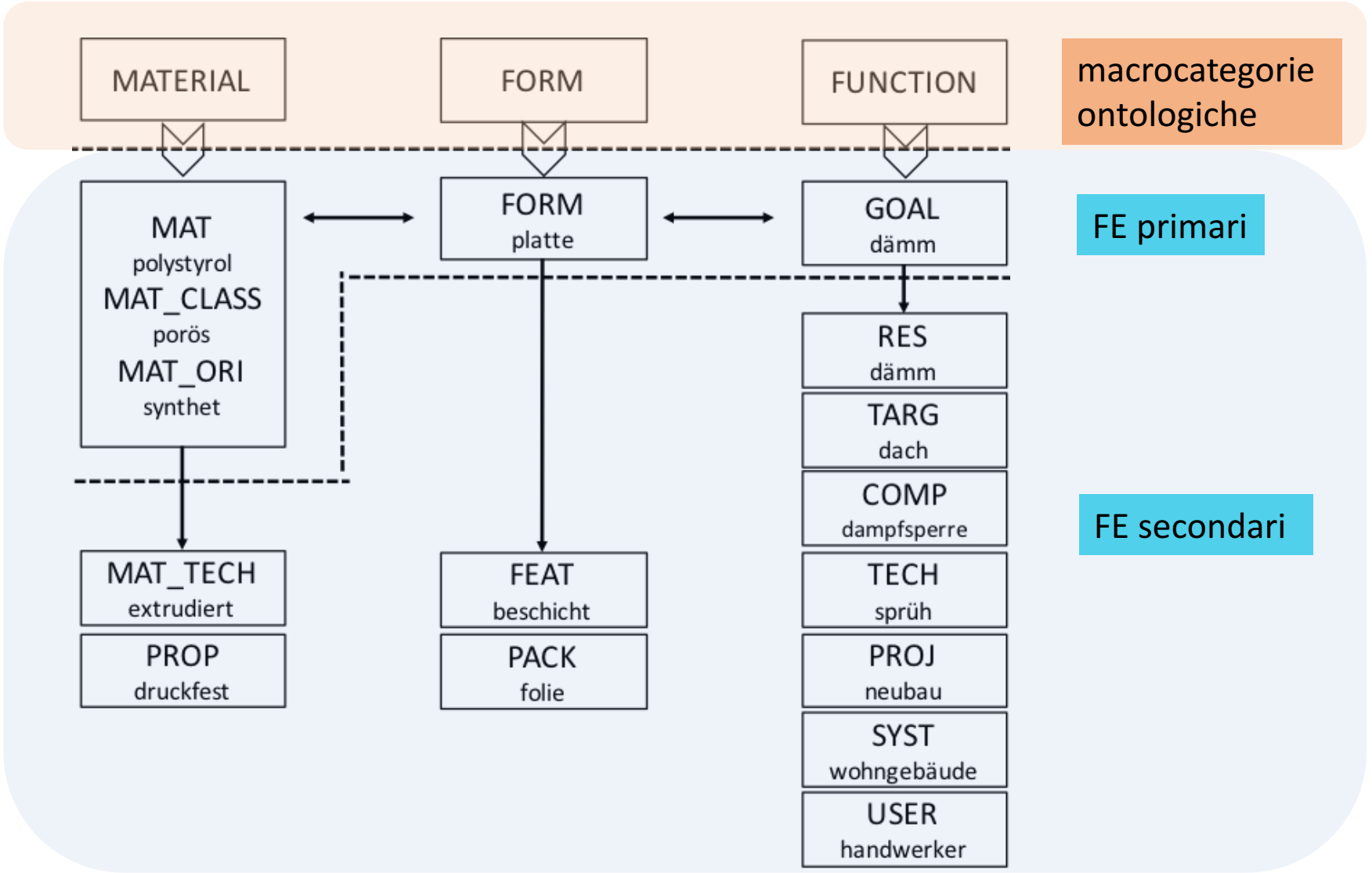
3. Restrizioni ontologiche:

   per evitare l'estrazione di termini non rilevanti per un determinato frame, ad es.

   TECH/adesiv- MATERIAL ORIGIN/natural- → *adesivo naturale*

   si possono applicare restrizioni ontologiche alle combinazioni degli elementi del frame.

   Le restrizioni ontologiche sono regole che stabiliscono la possibile combinazione degli elementi del frame (primari e secondari) tra di loro considerando le relazioni tra elementi del frame e categorie dell'ontologia di dominio.


MATERIAL
FORM
FUNCTION
macrocategorie ontologiche
MAT
polystyrol
MAT_CLASS
porös
MAT_ORI
synthet
FORM
platte
GOAL
dämm
FE primari
RES
dämm
TARG
dach
COMP
dampfsperre
FE secondari
MAT_TECH
extrudiert
PROP
druckfest
FEAT
beschicht
PACK
folie
TECH
sprüh
PROJ
neubau
SYST
wohngebäude
USER
handwerker

5. Considerazioni finali

- L'annotazione semantica tramite elementi del frame consente di impiegare diverse strategie (ad es. la permutazione di elementi lessicali e l'introduzione di variabili) per l'estrazione di termini semplici e polilessicali assieme alle corrispondenti varianti.
- Nel costituire il tagset iniziale, gli elementi del frame predefinito possono corrispondere a lemmi ma anche a specifiche forme lessicali, radici e temi.
- Il metodo consente di estrarre nuovi profili sintagmatici di un termine e nuovi termini riconducibili agli elementi del frame.
- Gli esperimenti finora svolti indicano che il metodo di annotazione basato su frame è applicabile ad altri domini e ad altre lingue.

Fonti

Evert, S. and A. Hardie (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. http://cwb.sourceforge.net/index.php

Faber, P. (2015). Frames as a framework for terminology. Handbook of terminology 1, 14–33.

Fillmore, C. J. and C. Baker (2010). A frames approach to semantic analysis. In B. Heine and H. Narrog (Eds.), The Oxford handbook of linguistic analysis, pp. 313–339. Oxford University Press.

Schäfer, J., I. Rösiger, U. Heid, and M. Dorna (2015). Evaluating noise reduction strategies for terminology extraction. In TIA, pp. 123–131.

Schmid, H. and F. Laws (2008). Estimation of conditional probabilities with decision trees and anapplication to fine-grained pos tagging. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, pp. 777–784. Association for Computational Linguistics.