Bilingual Term Extraction with Big Data.

Research project with the Zurich University of Applied Sciences (Switzerland).
Conducted by Mark Cieliebak.

Rémy Blättler
Chief of the System

In Zurich, Berlin and Los Angeles.
In English, Spanish, Portuguese, German,
Chinese, Japanese and 80 other languages.

Naïve approaches

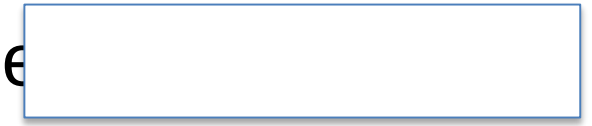
- Simple word frequencies
- Using dictionaries to find “phrases”

GENSIM word2vec

- Superficial, two-layer neural networks trained to reconstruct linguistic contexts of words
- Free & open-source

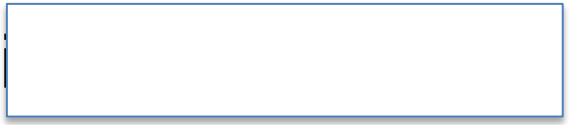
GENSIM word2vec

breakfast cereal dinner lunch => Ce



man => boy

woman => gi

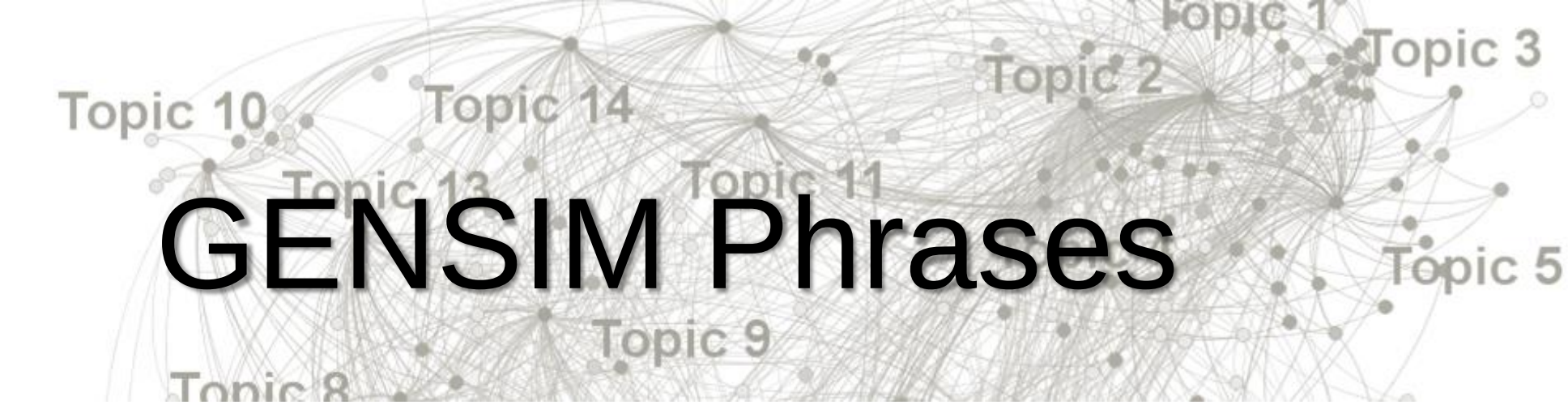


Sweden

⇒ Norway 0.76

⇒ Finland 0.71

⇒ Estonia 0.54



GENSIM Phrases

New York City

Terms and Conditions

Never Follow (Audi)

Just do it (Nike)

Term extraction 1

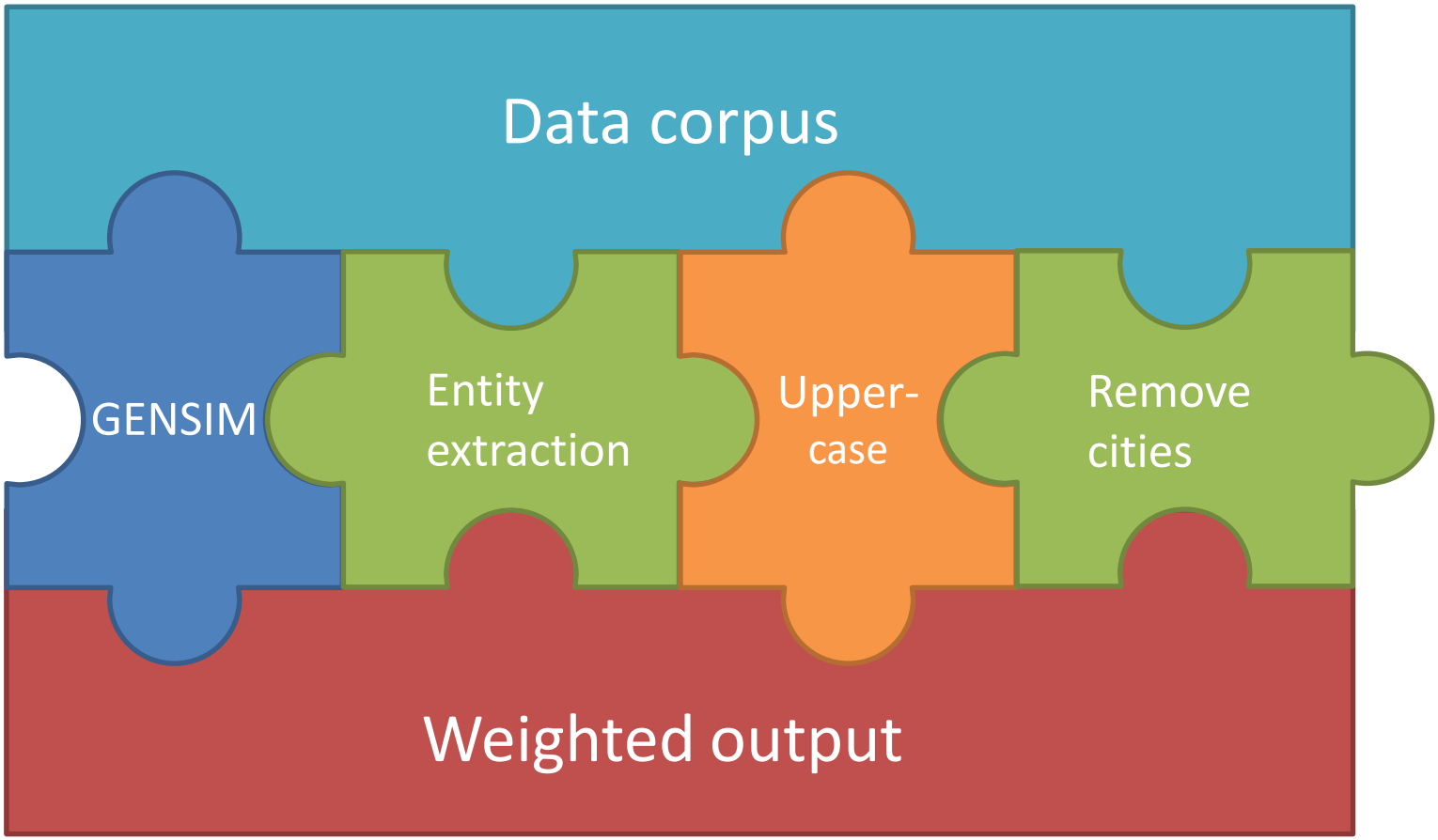
1. Average occurrence of a term over all corpora
2. Average occurrence of a term for one client
3. Same for the other language

Term extraction 2

Detect the specific phrase in the source & target:

Geänderte Segmentberichterstattung erhöht Aussagekraft.

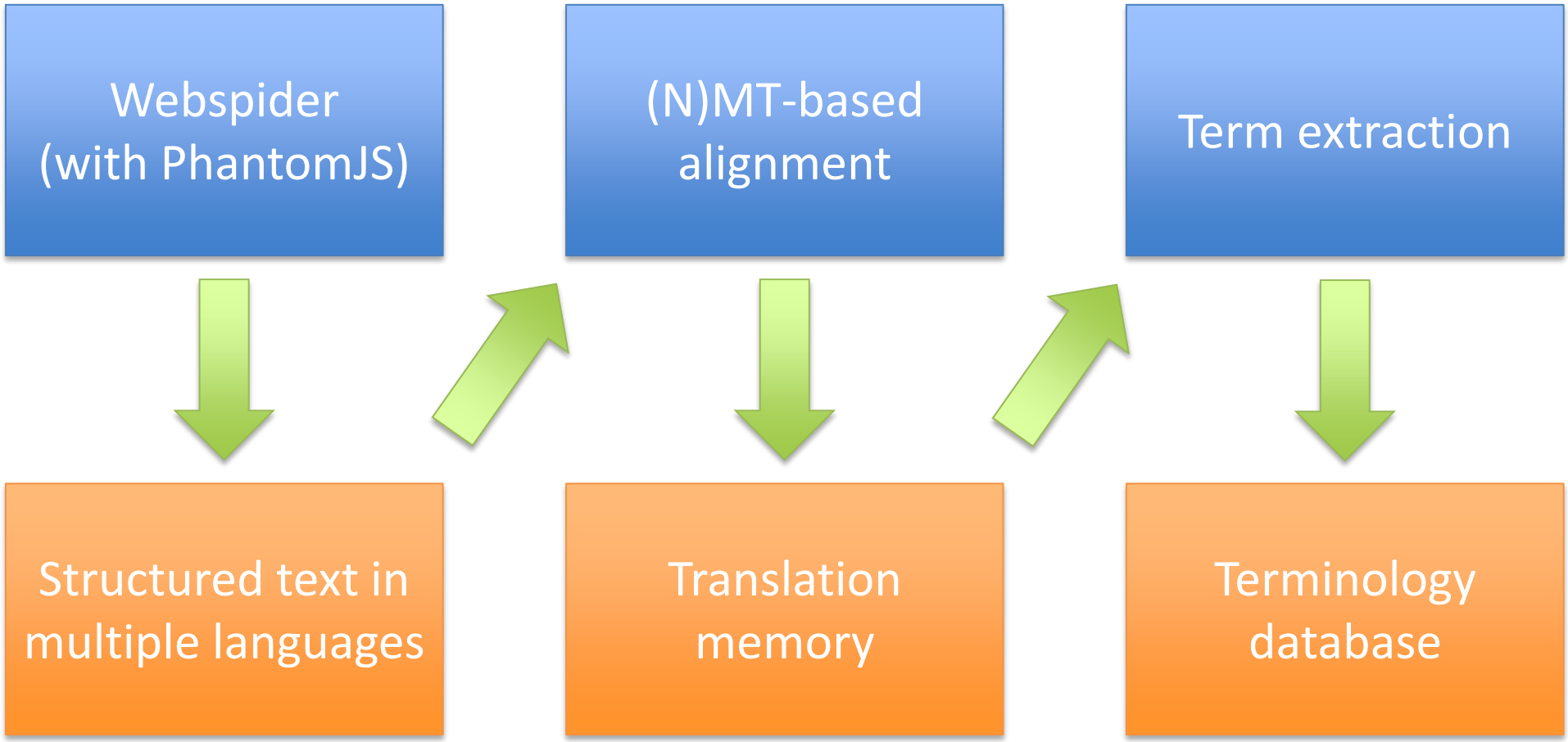
Improvements gained from changes in segment reporting.



Plugin additional algorithms

Works well for small fixes too

German	English
myCloud	myCloud
Wie kann ich	How can I
Swisscom Broadcast	Swisscom Broadcast
HomepageTool	HomepageTool
Abo	subscription
Swisscom (Schweiz) AG	Swisscom (Switzerland) Ltd
Swisscom Sharespace	Swisscom Sharespace
App	app
Healthi	Healthi
KMU	SME
SharePoint	SharePoint
Endresultat	Swisscom.”
Ihre Webseite	your website
KMU	SMEs
Dateien	Files
Swisscom myCloud	Swisscom myCloud
generelles Rauchverbot	cablex."
Dateien	Photos
inOne mobile	inOne mobile
Business Apps	Business Apps



The big picture: fast & easy client onboarding

Term extraction

German

Geänderte Segmentberichterstattung erhöht Aussagekraft.

English

Improvements gained from change in segment reporting.

☑	Segmentberichterstattung	17
☑	Swisscom (Schweiz) AG	28
☐	Endresultat	5
☑	Dateien	12

Segmentberichterstattung	17	90%
Swisscom (Switzerland) Ltd	28	87%
Swisscom.”	5	45%
Files	12	80%

Problems

- Speed (tests take multiple hours)
- Insufficient data (>50k TM units helps)
- Bad source data (HTML, Javascript, etc.)

Budget

$1,000 KTI Research Project

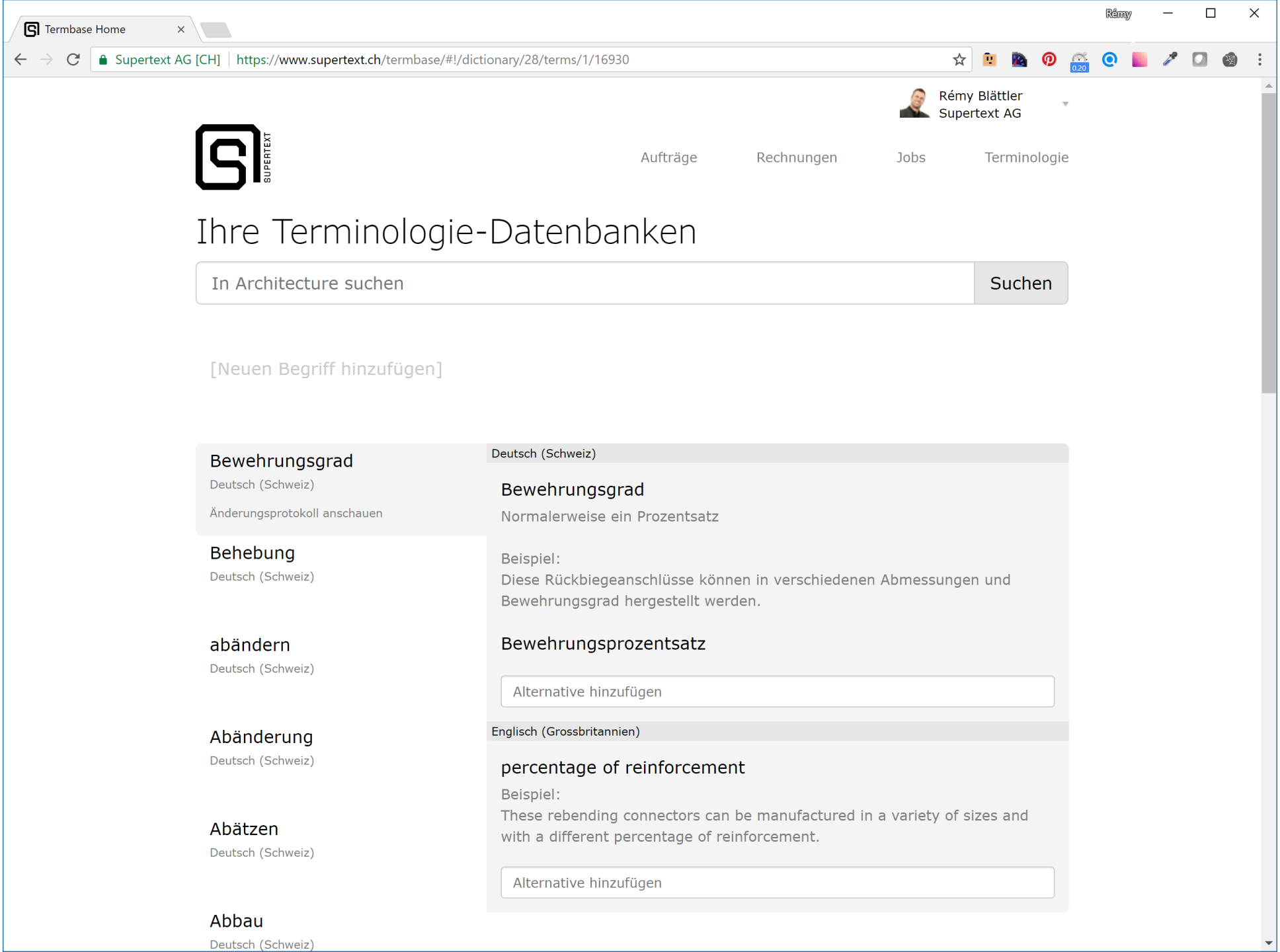
$20,000 SpinningBytes & University Cooperation

=> $20,000 more until online & 80h+ internal

Questions?

@remyblaettler (Twitter)

remy@supertext.ch


Termbase Home
Rémy
Supertext AG [CH] | https://www.supertext.ch/termbase/#!/dictionary/28/terms/1/16930
Rémy Blättler
Supertext AG
Aufträge
Rechnungen
Jobs
Terminologie
Ihre Terminologie-Datenbanken
In Architecture suchen
Suchen
[Neuen Begriff hinzufügen]
Bewehrungsgrad
Deutsch (Schweiz)
Änderungsprotokoll anschauen
Behebung
Deutsch (Schweiz)
abändern
Deutsch (Schweiz)
Abänderung
Deutsch (Schweiz)
Abätzen
Deutsch (Schweiz)
Abbau
Deutsch (Schweiz)
Deutsch (Schweiz)
Bewehrungsgrad
Normalerweise ein Prozentsatz
Beispiel:
Diese Rückbiegeanschlüsse können in verschiedenen Abmessungen und Bewehrungsgrad hergestellt werden.
Bewehrungsprozentsatz
Alternative hinzufügen
Englisch (Grossbritannien)
percentage of reinforcement
Beispiel:
These rebending connectors can be manufactured in a variety of sizes and with a different percentage of reinforcement.
Alternative hinzufügen

Files Edit Project Terminology Help

Documents | Extraction | TermBase | Tools & Config.

Source Extraction - 2222/20033

Expression	Occurren...	N-Gram
anywhere else	8	2
find things	8	2
personal and professional development	8	3
enough coins to enter the sweepstakes t...	8	5
gift certificates	8	2
manage and limit the effects of heart dis...	8	5
nutrition facts	8	2
available online	8	2
fully relax	8	2

Target Extraction

Expression	%	Occurren...
los certificados	46	12
los certificados de regalo	35	4
aplican términos y condiciones	29	6
nunca y pueden ser canjeados por un mil...	25	2
no caducan nunca y pueden ser canjead...	25	2

Source Selection, with 5 forms

gift certificates

Target Selection - 12 Occurrences

los certificados

Translation Units

☐ Terms and conditions are applied to ***gift certificates*** gift cards. Terms and conditions for jcpenney gift cards are listed on the eGift Card and also at jcp.com. jcpenney is not a sponsor or affiliated with this Program. The jcpenney name and logo are registered trademarks of 2012 J.C. [1]

☐ Se aplican Términos y condiciones a ***los certificados*** y tarjetas de regalo. Los Términos y condiciones de las tarjetas de regalo de JCPenney se enumeran en la tarjeta de regalo electrónica y también en jpc.com. JCPenney no es un patrocinador ni está afiliado con este programa. El nombre y el logo de JCPenney son marcas registradas de 2012 J.C. [1]

☐ Terms and conditions are applied to ***gift certificates*** gift cards. [1]

☐ Se aplican Términos y condiciones a ***los certificados*** y tarjetas de regalo. [1]

☐ Terms and conditions are applied to ***gift certificates*** gift cards. Terms and conditions for JCPenney gift cards are listed on the eGift Card and also at www.jcp.com. JCPenney is not a sponsor or affiliated with this Program. The JCPenney name and logo are registered trademarks of 2016 J.C. Penney Corporation Inc. [1]

☐ Se aplican Términos y condiciones a ***los certificados*** y tarjetas de regalo. Los Términos y condiciones de las tarjetas de regalo de JCPenney se enumeran en la eGift Card y también en www.jpc.com. JCPenney no es un patrocinador ni está afiliado a este programa. El nombre y el logo de JCPenney son marcas registradas de 2016 J.C. Penney Corporation Inc. [1]