— Lexical Chain Annotation Guidelines —

Steffen Remus

v1.0 – September 1, 2012

About the document

Use this document as a schema to identify and annotate lexical chains in documents of the current *Salsa*[1] *2.0* / *TiGer*[2] *2.1 Corpus*. This document guides you through common standards and potential pitfalls.

Introduction to lexical chains

When reading a text, whether short or long, the text is intelligible because of its structure. Each text is designed to transport a message, and a reasonable structure is the essential means for this. *Lexical chains* visualize this structure by connecting words or phrases in the text that are semantically related. These words or phrases are called *lexical items* and each item contributes to a specific meaning of a lexical chain. Consider the given example text below and some extracted lexical chains:

> COFFEE QUOTA TALKS CONTINUE BUT NO AGREEMENT YET
>
> LONDON, March 2 - Coffee quota talks at the International Coffee Organization council meeting here continued this afternoon, but producers and consumers still had not reached common ground on the key issue of how to estimate export quotas, delegates said.
>
> The 54 member contact group was examining a Colombian proposal to resume quotas April 1 under the ad hoc system used historically, with a pledge to meet again in September to discuss how quotas would be worked out in the future, they said.

1. SALSA – The Saarbrücken Lexical Semantics Acquisition Project, "see A. Burchardt et al. 2006. The SALSA Corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th international conference on language resources and evaluation (LREC-2006)*. Genoa, Italy".
2. TiGer – TiGer Treebank (Version 2.1), see "S. Brants et al. 2002. TIGER Treebank. In *Proceedings of the workshop on treebanks and linguistic theories (TLT02)*. Sozopol, Bulgaria".

1. { *COFFEE QUOTA TALKS, Coffee quota talks, Coffee Organization council meeting, quotas, quotas, quotas* }
2. { *CONTINUE, continued, resume* }
3. { *AGREEMENT, reached common ground, pledge, discuss* }
4. { *producers, consumers* }

Related terms are underlined with the same color, which build up the chains 1-4.

Although the example is incomplete yet (meaning there are more chains in the text), the central theme of the text (by just looking at the chains) clearly can be identified.

This technique can be utilized in many applications, and the main goal is to automatically identify these chains with the help of computer systems. In order to evaluate these systems, i.e. to say how good or bad they did the job, we need examples annotated by humans and compare them with those made by the system.

Short Description

- Lexical items are the basic elements of the chains occurring in the texts.
- Candidates for lexical items are words or phrases bearing some meaning.
- Lexical items describing the same concept or relating to the same topic form a so-called dense chain.
- Within dense chains each item must be of the same topic as every other item in the chain.
- Each lexical item must solely exist in exactly one chain.
- Chains consisting of only a single item may exist.
- Each dense chain can be assigned a description of its concept.
- Dense chains may be linked to describe broader concepts.
- Source and target of links are representative lexical items, which may link to several dense chains.

Schema

A Candidate items (items to annotate)

The question of which words or phrases in the text are adequate lexical items is crucial for forming lexical chains.

In general we will only consider **terms or phrases associated with meaning** as possible candidates, specifically:

- **Nouns** and **compounds** (examples for compounds are *cellar door*, *city council member*, *small talk*, ...).
- **Adjective noun combinations**, where the modifying adjective significantly influences the meaning of the item. E.g., the topic of *money* may be *economics*, but the phrase *dirty money* is more in the domain of *cheating* or more general *crime*.

  Always **prefer shorter expressions to longer ones**. Only consider longer phrases when the concept of the shorter phrase is very different and not cohesive at all. For example in the sentence `'Dorothy ate a hearty supper ...'` the phrase *hearty supper* is of minor interest, but the *supper* itself maybe indeed important for the context and an adequate lexical item.
- **Ignore words that bear no relevant meaning**. Strictly speaking, words that bring no relevant information about the context. These are typically words that frequently occur in texts. Examples for this include: *a*) numbers (*once, twice, hundreds, million, ...*) *b*) units (*percent, miles, meters, gallons, ...*) *c*) currencies (*dollar, euro, yen, ...*) *d*) times / periods (*monday, year, quarter, ...*) *e*) frequently occurring words (so called stop words) (*everybody*, *anybody*, *person*, *name*, *good*, *bad*, *start*, *end*, *go*, *walk*, ...) Specifically **ignore names, places and referring items** (*John, he, Toyota, it, ...*).

B Level one: dense lexical chains

- **Each item in a specific chain must be of the same topic as every other item in the chain.** This implies: *each chain can be assigned exactly one topical description.* More precisely: item X is of the same *narrow* topic as every other item Y in the chain. E.g. *car* is related to *driver* because they share the same topic (for instance *traffic* or *transport*) and every other item in the chain, e.g. the term *leg*, must also share this specific topic, as would obviously not be the case here.

Definition: *a*) **collocational** terms (i.e. semantic relationships of words / items that often co-occur) *b*) **repetition** of words or phrases (also including words with different flection (e.g. *tourist*, *tourists*)) *c*) **synonyms** or **near-synonyms** *d*) **superordinates** or **subordinates** (cf. hypernyms and hyponyms, e.g. *mammal* is a hypernym / superordinate term of *cow*) *e*) **opposites** or **antonyms** (also including converseness (doctor – patient), incompatibility, complementarity (male – female), ...) *f*) **part-whole-relations** (cf. meronymy, e.g. *arm*, *leg* are meronyms of *body*)

- Unfortunately the previous definition is accompanied with transitivity and uncertainty. As an example, consider the terms *car*, *bus*, and *airplane*. All of them share the same topic *vehicles*, thus one could create one dense chain {*car, bus, airplane*}. But *car* and *bus* are also *driving vehicles* whereas *airplane* is a *flying vehicle*, so one could also create two chains {*car*, *bus*} and {*airplane*}. On the other hand, unlike *car*, *bus* and *airplane* are both means of *public transport*, which would then result in {*bus, airplane*} and {*car*}.

  This example shows that the manual annotation process can only be done by human **intuition** (i.e. common sense and knowledge of the language) The context of candidate items is always of enormous importance and determinant for the choice of the chain. Always consider the following rule of thumb:

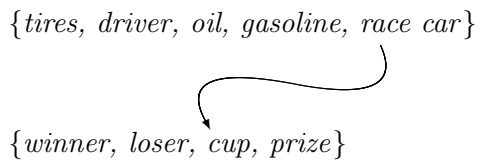
  If you "feel" that two candidate items are "semantically too far away" then better prefer a level two link to the broader item.

C Level two: links between dense chains

- **Dense chains can be linked via lexical items.** A link can be described as a semantic relation between two lexical items belonging to different dense chains. As in B, a semantic relationship exist if two lexical items share the same topic.

  This concept is best described with an example, consider therefore the term *race car* where the head of the compound signals the topic *car* (in general) and the modifier signals the topic *racing* or *contest*. Consider further the two fictional chains {*tires, driver, oil, gasoline*} describing the topic *car* and {*winner, loser, cup, prize*} describing the topic *contest*. The term *race car* belongs in principle into both chains, but its stronger affinity belongs to the former one (*car*), and since in our scenario a lexical item belongs to exactly one chain it will be put into the *car* chain. Stopping here would lead to inconsistencies, so we connect *race car* with the *contest* chain via that lexical item, that is

most coherent with *race car* (i.e. the most representative lexical item sharing the same topic), here *cup*, visually resulting in:



- Once two chains are connected there is no need to add further links. Each chain corresponds to its own description of a concept, and the links between chains express that these concepts are related and form a wider concept. E.g. the resulting description of the concept of the linked chains in the example above would be *auto racing*.
- Several links via the same lexical item are possible and allowed. E.g. think about *formula one race car* which may encode three topics.

D Potential pitfalls

- homonymy or polysemy – i.e. same form different meaning / topic.
- items share the same topic in general, but not in the current context. Consider for example the following two extracts of the same document:

  "... due to the falsification of documents, ..."

  and

  "... quoted the state department spokesman from the letter."

  Although the word *documents* is in fact a hypernym of the word *letter*, they should not be considered to be put into the same dense chain, because in this special case the context of the term *letter* refers to some topic like *writing*, whereas *documents* refers to something *official* e.g. ID cards, passports, or the like.
- Beware of the correct meaning of compounds, e.g. which part is the head and which the modifier, or is the meaning of the compound completely independent. Consider for example the words *waterbed* and *blackboard*, where *waterbed* is simply a special kind of *bed* but *blackboard* is not simply a *black board*.

Annotation recipe

READ – GRASP – SEARCH & MERGE – REVISIT & LINK

1. **Read** the document and **grasp** the main message behind the text.
2. **Pick** the next unchained lexical item as described in A.
3. **Search** for appropriate chains or other lexical items, and if they share the same topic as described in B **merge** them.

   Remember that **each item** in the dense chain should satisfy the condition **with every other** item in the chain.
4. Continue with 2 until all lexical items in the text are served.

   Remember that single item chains may exist.
5. **Revisit** the dense chains and check if you should do one of the following operations:

   - splitting of chains – often chains grow too large because a too broad topic was chosen.
   - merging of chains – often too many chains emerge because a too narrow topic was chosen.
6. **Add links** between dense chains via the lexical items as described in C. Try to **pick meaningful representatives**, because it often happens that many items serve as candidate links. It is thus sufficient to **link chains only once**. In other words, if a link exists between the chains A and B there's no need to seek for other links between A and B.

   Remember that several links via the same lexical item are allowed.

When you are done, you ideally have one or two "big chains" (when considering linked chains as a whole) describing the main message of the text, as well as a few small chains describing side topics.

Examples

The following examples show selected texts from the *Salsa* corpus. Terms in square brackets are preselected candidate lexical items (cf. schema section A). This preselection is done to ease the annotation process, schema A in particular. These items are automatically selected based on a part-of-speech pattern and are not necessarily exact. However, in most cases the candidate items are adequate enough and no effort must be put into finding additional candidate items. Consider only correct items for chaining and linking, though.

The examples comprise the original text with preselected candidate items, the identified dense chains and links including comments. The examples are supplemental to the annotation recipe above and may still be arguable, though they might give a good direction to reach common ground.

TEXT:

USA

Weiteres [Geständnis] im [Touristenmord-Prozeß]

RIVERSIDE , 17. November (ap / dpa) .

Der dritte [Beteiligte] an der [Ermordung] der [deutschen [Touristin]] Gisela Pfleger hat sich des [Raubes] und [Totschlags] schuldig bekannt .

Die [Staatsanwaltschaft] in Riverside in Kalifornien machte das [Bekenntnis] des [20jährigen] Xou Yang jetzt publik .

Des [Mordes] an der [64jährigen] aus Emmerich sind bereits der 19jährige Thongxay Nilakout und der 20jährige Khamchan Bret Ketsouvannasane schuldig gesprochen worden .

Die [Strafmaße] werden später bekanntgegeben .

Zu [lebenslangen [Freiheitsstrafen]] wurden zwei [junge [US-Amerikaner]] in Miami wegen des [Mordes] an dem [deutschen [Touristen]] Uwe-Wilhelm Rakebrand verurteilt .

DENSE CHAINS:

1. {[Geständnis], [Bekenntnis]}
   (the usage of "Geständnis" and "Bekenntnis" is synonymous)
2. {[Touristenmord-Prozeß], [Staatsanwaltschaft], [Strafmaße], [Freiheitsstrafen]}
   (All items share the same topic (e.g. "Prozeß"). "Touristenmord-Prozeß" is an endocentric compound consisting of "Tourist", "Mord", and "Prozeß", but the predominant part is "Prozeß". The fact that in this case "Freiheitsstrafen" is favored over "lebenslangen Freiheitsstrafen" is that the adjective "lebenslang" is not relevant to get the message of the text.)
3. {[Ermordung], [Raubes], [Totschlags], [Mordes], [Mordes]}
   ("Ermordung" and "Mordes" can be used synonymous and all words share the same topic (e.g. "Verbrechen"))
4. {[Touristin], [Touristen]}
   (Same words but different flection. "Touristin" and "Tourist" are favored over "deutschen Touristin" and "deutschen Touristen" because the modifying adjective is irrelevant.)

LINKS:

1. {Touristenmord-Prozeß → Geständnis }
   (A trial ("Prozeß") often ends with a confession ("Geständnis"), thus the terms are related.)
2. {Touristenmord-Prozeß → Ermordung}
   ("Ermordung" is synonymous with "Mord" which is a part of the compound "Touristenmord-Prozeß")
3. {Touristenmord-Prozeß → Touristen}
   ("Touristen" is a part of the compound "Touristenmord-Prozeß'.')
4. {Ermordung → Beteiligte}
   ("Ermordung" is a criminal act with at least two participants ("Beteiligte") the murderer and the victim, in this case two murderers.)

TEXT:

[Franzosen] gegen [Atomtests]

PARIS (ap) .

Zwei von drei [Franzosen] sind nach einer neuen [Meinungsumfrage] gegen die [Atomversuche] ihres [Landes] .

Nach den am Montag in Paris [veröffentlichten [Ergebnissen]] des [Instituts] TMO Consultants sprachen sich 23 Prozent für die [Versuche] im Südpazifik aus - zehn Prozent äußerten keine [Meinung] .

DENSE CHAINS:

1. {[Franzosen],[Franzosen]}
   (Repetion of the same word.)
2. {[Atomtests],[Atomversuche],[Versuche]}
   (Each word is about a test (nuclear tests in particular).)
3. {[Meinungsumfrage],[Ergebnissen],[Meinung]}
   (Each word is about the same topic which is some opinion poll.)

LINKS:

1. {[Franzosen] → [Landes]}
   (The term "Landes" builds up a single element chain and is linked to "Franzosen" because it is France which is meant by the term "Landes".)
2. {[Meinungsumfrage] → [Instituts]}
   ("Instituts" builds up a single chain element and is linked to "Meinungsumfrage" because a poll is mostly carried out by institutions.)

TEXT:

[Lärm] gegen (Ver) [schweigen]

BELGRAD , 3. Januar (rtr) .

Tausende [Bewohner] der [jugoslawischen [Hauptstadt]] Belgrad sind dem [Aufruf] der [Opposition] gefolgt und haben am Donnerstag während der [TV-Abendnachrichten] mit [Töpfen] und [[Pfannen] [Lärm]] gemacht .

[Studenten] hatten zu dem [Protest] aufgerufen , um gegen die einseitige [Berichterstattung] des [staatlichen [Fernsehens]] zu protestieren .

Der [Krach] war in ganz Belgrad zu hören .

Die [Opposition] warf der [Stadtverwaltung] vor , das [Streuen] [vereister [Straßen]] absichtlich zu verzögern .

Dadurch kam es zu vielen [Knochenbrüchen] .

Erstmals hat auch die [Serbisch-Orthodoxe [Kirche]] der [[Regierung] [Wahlbetrug]] vorgeworfen .

DENSE CHAINS:

1. {[Lärm],[schweigen],[Lärm],[Krach]}
   (The main shared topic of these terms is noise. Here, the term "schweigen" is only partly correct extracted (cf. "(Ver) schweigen"), but the message is the same.)
2. {[Bewohner],[Hauptstadt]}
   (Each town has citizens.)
3. {[Aufruf],[Protest]}
   (The meaning of the term "Aufruf" is in fact "Protestaufruf", which is in the same topic as "Protest".)
4. {[Opposition],[Opposition],[Stadtverwaltung],[Regierung]}
   (An example were each term is connected with the topic "politics".)
5. {[TV-Abendnachrichten],[Berichterstattung],[Fernsehens]}
   (Each term is about public media.)
6. {[Streuen],[vereister Straßen]}
   (Icy streets should be salted.)

LINKS:

1. {Protest → Lärm}
   (Most protests are noisy.)
2. {Studenten → Protest}
   (In most protests students are involved.)
3. {vereister Straßen → Knochenbrüchen}
   (Not salting icy streets increases the probability of injuries of any kind.)
4. {Regierung → Wahlbetrug}
   (This link tells that the government is playing wrong indicated by the last part of the compound "Wahlbetrug", where the first part of the compound ("Wahl") is directly connect to the chain describing the topic politics.)