LIRICS

Deliverable 3.3B

Interim Report: WD of syntactic annotation standard CD ballot

Pro ect reference number	e-Content-22236-LIRICS
Pro ect acronym	LIRICS
Pro ect fu t t e	L ngu st c Infrastructure for Interoperab e Resource and Systems
Pro ect contact po nt	Laurent Romary, INRIA-Lor a 615, rue du ard n botan que BP101. 54602 V ers ès Nancy (France) romary@ or a.fr
Pro ect web s te	ttp:// r cs. or a.fr
EC pro ect off cer	Erw n Va ent n
Document t t e	Inter m Report: CD document, 1st vers on
De verab e ID	3.3B
Document type	Report
D ssem nat on eve	Conf dent a
Contractua date of de very	M27 for D3.3B
Actua date of de very	09.02.2007
Status & vers on	Draft
Wor pac age, tas & de verab e respons b e	DFKI
Aut or(s) & aff at on(s)	T erry Dec erc , M r am Kess er, U r c Kr eger, Tan a Avgust onva and Va a Kordon (DFKI)
Add t ona contr butor(s)	Many experts from nat ona standard sat on bod es and ISO.
Keywords	Syntax, Annotat on, Standards, Tree-Ban s

Document evolution

Vers on	date	vers on	date
0.9	22nd Dec. 2006		
1.0	31st Jan 2007		
2.0	30t August 2007		

ISO 24615:2007

Introduction

We present n t s an updated vers on of t e de verab e D3.3A t e actua state of wor of SynAF (Syntact c Annotat on Framewor), w c w be soon subm tted as a CD (end of September). SynAF as been rev ewed n t e ne of o nt d scuss ons w t t e ed tors of LAF, LMF and MAF, d scuss ons w c started at t e p enary ISO TC37 n Be ng (21-26 August 2006), and ere we nc ude some mod f cat ons suggested by part c pants to d scuss on on SynAF at a ISO meet ng n Par s (Ma 2007) and at t e p enary ISO TC37 meet ng n Provo (August 2007), w c we dec ded to nc uded n t e vers on to be subm tted for a CD ba ot, postpon ng t us t e subm ss on to September 2007.

.

Reference number of wor ng document: ISO/TC 37/SC 4 N??? Rev.4

Date: 2007-08-22

ISO CD 24615:2007

Comm ttee dent f cat on: ISO/TC 37/SC 4

Secretar at: KATS

Language resource management—Syntact c Annotat on Framewor (SynAF)

Gest on des ressources ngu st ques — Cadre d' Annotat on Syntact que —

Document type: Internat ona standard
Document subtype: f app cab e
Document stage: 30.00
Document anguage: en

TABLE OF CONTENTS

WARNING 5

COPYRIGHT NOTICE II

FOREWORD IV

1 SCOPE 6

2 NORMATIVE REFERENCES 7

3 TERMS AND DEFINITIONS 7

4 KEY STANDARDS USED BY SYNAF 9

4.1 UNICODE 9
4.2 ISO 12620 DATA CATEGORY REGISTRY (DCR) 9
4.3 UNIFIED MODELING LANGUAGE (UML) 9

5 EMBEDDING SYNAF IN THE LAF MODEL 9

6 THE SYNAF METAMODEL 10

6.1 INTRODUCTION 10
6.2 THE SYNAF DIAGRAM (TO BE REPRESENTED IN UML) 11
6.2.1 T Nodes class 11
6.2.2 NT Nodes class 11
6.2.3 Edges class 11
6.2.4 Syntactic Annotation class 11

ANNEX A : (INFORMATIVE) DATA CATEGORIES FOR SYNAF 13

A.1 CONSTITUENCY 13
A.1.1 The TIGER Tagset for Node Labels ***Erreur ! Signet non défini.***
A.1.2 The ISST Tagset for Node Labels ***Erreur ! Signet non défini.***
A.2 DEPENDENCY 14
A.2.1 The Sparkle Tagset for Edge Labels (Grammatical Relations in Sparkle) ***Erreur ! Signet non défini.***
The Summary of the Sparkle tagset for Dependencies in a Table: ***Erreur ! Signet non défini.***
A.2.2 The Tiger Tagset for Edge Labels ***Erreur ! Signet non défini.***

ANNEX B (INFORMATIVE) ANNOTATION EXAMPLES 17

Foreword

ISO (t e Internat ona Organ zat on for Standard zat on) s a wor dw de federat on of nat ona standards bod es (ISO member bod es). T e wor of prepar ng Internat ona Standards s norma y carr ed out t roug ISO tec n ca comm ttees. Eac member body nterested n a sub ect for w c a tec n ca comm ttee as been estab s ed as t e r g t to be represented on t at comm ttee. Internat ona organ zat ons, governmenta and non-governmenta , n a son w t ISO, a so ta e part n t e wor . ISO co aborates c ose y w t t e Internat ona E ectrotec n ca Comm ss on (IEC) on a matters of e ectrotec n ca standard zat on.

Internat ona Standards are drafted n accordance w t t e ru es g ven n t e ISO/IEC D rect ves, Part 3.

Draft Internat ona Standards adopted by t e tec n ca comm ttees are c rcu ated to t e member bod es for vot ng. Pub cat on as an Internat ona Standard requ res approva by at east 75 % of t e member bod es cast ng a vote.

Internat ona Standard 24615 was prepared by Tec n ca Comm ttee ISO/TC 37, *Terminology and other language resources*, Subcomm ttee SC 4, *Language resource management*, n co aborat on w t t e European eContent Pro ect "LIRICS" (**L** ngu st c **I**nf**r**astructure for **I**nteroperab e Resour**c**es and **S**ystems), under t e contract e-Content-22236-LIRICS.

ISO 24615 s des gned to coord nate c ose y w t ISO AWI 24612, *Linguistic Annotation framework (LAF)*, and ISO CD 24613, *Lexical Markup Framework (LMF)*, and ISO CD 24611, *Morphosyntactic Annotation Framework (MAF)*, and ISO NP 2461x-1, *Semantic Annotation Framework - Part 1: Time and events (SemAF-Time)*.

Annexes A forms an ntegra part of t s Internat ona Standard.

Introduction

T ere ave been n t e past no t oroug standard sat on act v t es n t e doma n of syntact c annotat on, desp te t e numerous pro ects (see Abe é, 2003) t at ave des gned ways to mp ement ngu st c TreeBan s, .e. syntact ca y annotated corpora. For severa years t e Penn Treeban n t at ves ave served as a de facto standard, but more recent wor (e.g. t e Negra/T ger n t at ve[1] n Germany or t e ISST n t at ve n Ita y[2]) as s own t at a more co erent framewor cou d be des gned to account for bot (erarc ca) const tuency and dependency p enomena n syntact c annotat on.
W t n t e European eContent LIRICS pro ect, a group of nternat ona experts as started t e ISO process, ca ed SynAF (Syntact c Annotat on Framewor). T e actua document s a rev s on of ISO WD 24615, which is t e resu t of a more extended d scuss on, nc ud ng feedbac and comments from ISO experts, and w be subm tted for ts acceptance as a CD.

T e document proposes a metamode for syntact c annotat on and sts n t e annex cand date data-categor es for syntact c annotat on, to be descr bed n more deta s n ISO/TC 37/SC 4 Ad oc T emat c Doma n Group 4: Syntax (on syntact c data-categor es). T e estab s ment of t s group as been reso ved at t e ISO TC37/SC4 annua meet ng n Be ng (2006-08-21/25).

[1] See: ttp://www. ms.un -stuttgart.de/pro e te/TIGER/TIGERCorpus/

[2] See Montemagn (2003).

1 Scope

T s Internat ona Standard descr bes t e Syntact c Annotat on Framewor (SynAF), a g eve mode for represent ng t e syntact c annotat on of textua documents.
SynAF s bu d ng on t e ISO MAF proposa (CD 24611). MAF (Morp o-Syntact c Framewor) s dea ng w t t e morp o-syntact c annotat on of spec f c segments of textua documents. T e morp o-syntact c annotat on framewor s about *part of speech* (noun, ad ect ve, verb, etc.), *morphological* and *grammatical* features (suc as number, gender, person, mood, verba tense).
SynAF s about t e annotat on of t e syntact c const tuency of suc (groups of) morp o-syntact ca y annotated fragments and t e syntact c dependency re at ons ex st ng between t ose (groups of) morp o-syntact ca y annotated fragments. We cons der t at t e sentence w def ne t e boundar es of t e fragments of textua documents to w c SynAF w app y.
As suggested ust above, syntact c annotat on as at east two funct ons n anguage process ng:

1) To represent ngu st c const tuenc es, e Noun P rases (NP), descr b ng a structured sequence of morp o-syntact ca y annotated tems[3], w ere we cons der a so const tuents bu t from non-cont guous e ements, and

2) To represent dependency re at ons, e ead-mod f er re at on[4]. T e dependency nformat on can ex st between morp o-syntact ca y annotated tems w t n a p rase (an ad ect ve s t e mod f er of t e ead noun w t n an NP) or descr be a spec f c re at on between syntact c const tuents at t e c ausa and sentent a eve (.e. an NP be ng t e "sub ect" of t e ma n verb of a c ause or sentence). T e dependency re at on can a so be stated nc ud ng empty e ements (e t e pro-drop property n romance anguages[5])

SynAF s dea ng w t t e descr pt on of a metamode for syntact c annotat on, w c means t at SynAF w descr be e ementary ngu st c (n fact syntact c) abstract ons t at support t e construct on and t e nteroperab ty of (syntact c) annotat ons and resources. T e T emat c Doma n Group 4 (TDG 4) “Syntax” assoc ated to SynAF w propose t e def n t on of t e re ated data categor es, w c w represent a po nt of reference for part cu ar tagsets used for t e syntact c annotat on of var ous anguages, a so n t e context of var ous app cat on scenar os.
To summar ze: SynAF s concerned w t a metamode t at covers bot d mens ons of syntact c *constituency* and *dependency*, and SynAF w propose a mu t - ayered annotat on framewor t at a ows t e comb ned and nterre ated annotat on of anguage data a ong bot nes of cons derat on. A so t e data-categor es to be proposed w t n TDG4 w be about t e bas c annotat on concern ng bot d mens ons.
T s standard s des gned to be used n c ose con unct on w t t e metamode presented n ISO AWI 24612, L ngu st c resource framewor (LAF) and w t ISO 12620, Term no ogy and ot er anguage resources — Data categor es.

---

[3] But SynAF s a so des gned for dea ng w t e empty e ements or traces generated by movements at t e const tuency eve .

[4] Inc ud ng a so re at ons between same categor es, e t e ead- ead re at on between nouns n appost ons or nom na coord nat ons.

[5] T s po nt as been part cu ar y stressed by t e aut ors of t e ISST framewor , s ow ng ere an advantage of t e two- eve approac , w ere t e dependency nformat on do not ave to map ent re y to t e const tuencey approac . In t s sense, bot eve s of annotat on ave a certa n ndependency n re at on to eac ot er (see Montemagn , 2003).

2 Normative references

T e fo ow ng normat ve documents conta n prov s ons t at, t roug reference n t s text, const tute prov s ons of ISO 24615. For dated references, subsequent amendments to, or rev s ons of, any of t ese pub cat ons do not app y. However, part es to agreements based on ISO 24615 are encouraged to nvest gate t e poss b ty of app y ng t e most recent ed t ons of t e normat ve documents nd cated be ow. For undated references, t e atest ed t on of t e normat ve document referred to app es. Members of ISO and IEC ma nta n reg sters of current y va d Internat ona Standards.
ISO 639-1:2002, Codes for t e representat on of names of anguages – Part 1: A p a-2 Code.
ISO 639-2:1998, Code for t e representat on of anguages – Part 2: A p a-3 Code.
ISO DIS 639-3:2005, Codes for t e representat on of anguages – Part 3: A p a-3 Code for compre ens ve coverage of anguages.
ISO 1087-1:2000, Term no ogy – Vocabu ary – Part 1: T eory and app cat on.
ISO 1087-2:1999, Term no ogy – Vocabu ary – Part 2: Computer app cat on.
ISO/IEC 10646-1:2003, Informat on tec no ogy – Un versa Mu t p e-Octet Coded C aracter Set (UCS).
ISO/IEC 11179-3:2003, Informat on Tec no ogy – Data management and nterc ange – Metadata Reg str es (MDR) – Part 3: Reg stry Metamode (MDR3)
ISO 12620:200?, Term no ogy and ot er anguage resources – Data Categor es – Spec f cat on of data categor es and management of a data category reg stry for anguage resources.

3 Terms and definitions

For t e purposes of t s Internat ona Standard, t e terms and def n t ons g ven n ISO 1087-1, ISO 1087-2, ISO 12620:200? and t e fo ow ng app y:

3.1

Annotat on: Some code assoc ated w t parts of text and prov d ng for add t ona nformat on about t s part of text. In t s document we use "annotat on" as a s ort form for " ngu st c" annotat on, mean ng t e nd of textua enr c ment t at can be prov ded by ngu st c nformat on, w c s ere m ted to morp o-syntax and syntax.

Category: a feature va ue prov d ng t e content of a node.

C ause: a group of *phrases*, usua y conta n ng a verb, w c va ency a so determ nes t e number of ob gatory c ause e ements (*phrases*). A c ause can be e t er a *main clause* or a *subordinated clause*. C auses can be e t er f n te or non-f n te, n dependency of t e mode of ts verb. Usua y, a f n te c ause conta ns at east a *subject* n add t on to t e verb. A ma n c ause a one can bu d a comp ete sentence. In our mode , a c ause s a spec a case of a const tuent.
Const tuent: a types of nodes we f nd n t e syntact c annotat on are bu d ng a const tuent (to be rev sed)
Const tuency re at on: a syntact c group ng of words *(i*nto *phrases)*, *phrases* (nto *clauses*) or *clauses (* nto a *sentence)* on t e base of structura (or erarc ca) propert es
Dependency re at on: a re at on between const tuents on t e base of grammat ca funct ons const tuents p ays n re at on to eac ot er w t n t e arger const tuent t ey are embedded n.
Edge: a tr p et w t a source node, a target node, and a abe . Non-Term na nodes ave an outgo ng const tuency edges (to be def ned)

Grammat ca funct on: Const tuents can ave a grammat ca funct on w t n t e r embedd ng syntact c env ronment. So a NP can act as a sub ect w t n a *sentence.* We spea ere a so of a grammat ca re at on between t e sub ect-NP and t e ma n verb n a sentence. We subsume a t ose grammat ca re at ons (Sub ect-Pred cate, Head-Mod f er, etc.) under t e concept of *dependency* re at ons.

Grap : a we understood mode for represent ng ob ects t at can be v ewed as a connected set of more e ementary sub-ob ects
Head: t e most mportant word n a const tuent; t e word t at carr es t e ma n mean ng of t e p rase. T e ead of a const tuent cannot be eft out.

H erarc y: t e re at ve pos t on of const tuents n a syntact c tree.
Human anguage tec no ogy: tec no ogy as app ed to natura anguages

Labe : a feature va ue prov d ng t e content of an edge.
Ma n c ause: a f n te c ause, w c can act on ts own as a comp ete sentence. Examp e: *The train has some delay.*
Mod f er: a mod f er s a part of t e const tuent w c ascr bes a property to t e ead of t e const tuent. A mod f er may be p aced before or after t e ead of t e p rase (pre-mod f er or post-mod f er). Mod f ers are opt ona n a const tuent

Natura anguage process ng NLP: f e d cover ng now edge and tec n ques nvo ved n t e process ng of ngu st c data by a computer

Node: pa r cons st ng of a (poss b y mu t p e) span and a category,

Non-Term na nodes ave an outgo ng const tuency edges (to be def ned)

P rases: a word or group of words w c can fu f a *grammatical function* n a c ause. But we a ow empty p rases (as t e examp e of t e empty NP n Ita an and Span s , be ng no-rea sed pronouns and av ng t e ro e of sub ects n *clauses*). A p rase s typ ca y named after t e most mportant word n t (w c we a so ca t e *head*), so we ave for examp e noun p rases, verb p rases, ad ect ve p rases, adverb p rases and prepos t ona p rases.). P rases ave been nforma y descr bed as "b oated words", n t at t e parts of t e p rase t at are added to t e ead e aborate and spec fy t e reference of t e ead word. In our mode , a p rase s a spec a case of a const tuent.

Sentences: a sequence of words, start ng very often w t a cap ta etter up to a f na punctuat on mar . But s def n t on s too restr cted to ayout property of certa n anguage sty es. A usage ru e says t at a comp ete sentence must conta n a sub ect and a verb (n f n te mode). A sentence cons sts of one or more *clauses.* In descr b ng speec , t s common to ta about 'utterances' rat er t an sentences.

Span: a pa r of po nts dent fy ng a segment of t e document subm tted to syntact c annotat on. T e f rst po nt s ess or equa to t e second po nt. A Mu t p e span s sequence of spans w ere t e end ng po nt of eac span s ess or equa to t e start ng po nt of t e subsequent span.

Spec f er: a spec f er n a const tuent spec f es t e ead (or t e comb nat on of mod f er and ead) w t nformat on about number, def n teness, prox m ty and owners p

Subord nated c ause: a c ause w c fu f s a *grammatical function* n a *phrase* (for examp e a re at ve c ause mod fy ng t e ead noun of a nom na p rase) or n anot er c ause. A subord nated c ause can not act on ts own as a *sentence*.

Sucategor zat on frame: set of restr ct ons nd cat ng t e propert es of t e syntact c arguments t at can or must occur w t t.
Examp e: A fred (syntact c argument) read a boo (syntact c argument) today (ad unct)
NOTE T e sub ect, nd rect ob ect and d rect ob ect are poss b e grammat ca re at on for a sentence.
Syntax: T e way words are grouped toget er n ngu st ca y mean ngfu un ts, and t e re at ons t at ex st between t ose un ts.
Syntact c argument: one of t e essent a and funct ona e ements n a c ause t at dent f es t e part c pants n t e process referred to by a verb

Syntax Tree: a syntact c grap n w c eac node as a s ng e parent.
Term na node: refers to a s ng e wordFrom/ ex ca un t or a span w t entg =0, and t e node and t e wordForm/ ex ca un t ave dent ca span.

4 Key standards used by SynAF

4.1 Unicode

SynAF s Un code comp ant and presumes t at a data are represented us ng Un code c aracter encod ngs.

4.2 ISO 12620 Data Category Registry (DCR)

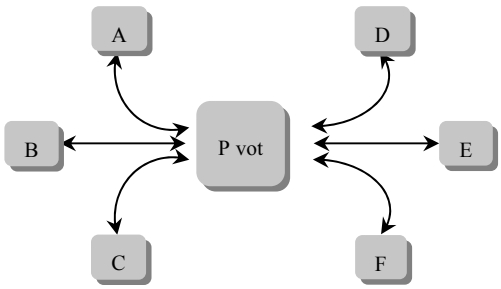
T e des gners of an SynAF conformant annotat on s a use data categor es from t e ISO 12620 Data Category Reg stry (DCR), or a tagset t at can be mapped onto t e data categor es.

4.3 Unified Modeling Language (UML)

SynAF comp es w t t e spec f cat ons and mode ng pr nc p es of UML as def ned by t e Ob ect Management Group (OMG) [4]. SynAF uses a subset of UML t at s re evant for ngu st c descr pt on. (not done yet).

5 Embedding SynAF in the LAF model[6]

We want to embed t e meta-mode of SynAF n t e more gener c L ngu st c Annotat onFramewor (LAF) and reuse ts annotat on strategy. LAF prov des a genera framewor for represent ng annotat ons t at as been descr bed e sew ere n deta (Ide and Romary, 2004, 2006). Its deve opment as bu t on common pract ce and convergence of approac n ngu st c annotat on over t e past 15-20 years. T e core of t e framewor s spec f cat on of an abstract mode for annotat ons nstant ated by a *pivot format*, nto and out of w c annotat ons are mapped for t e purposes of exc ange.



[6] T e w o e sect on 5 s ta en from (Ide, 2007).

Figure 1: Use of the LAF pivot format

F gure 1 s ows t e overa dea for s x d fferent user annotat on formats (abe ed A – F), w c requ res two mapp ngs for eac sc eme—one nto and one out of t e p vot format, prov ded by t e sc eme des gner. T e max mum number of mapp ngs among sc emes s t erefore 2*n*, vs. n^2-*n* mutua mapp ngs w t out t e p vot.

To map to t e p vot, an annotat on sc eme must be (or be rendered v a t e mapp ng) somorp c to t e abstract mode , w c cons sts of (1) a *referential structure* for assoc at ng stand-off annotat ons w t pr mary data, nstant ated as a d rected grap ; and (2) a *feature structure representation* for annotat on content. An annotat on t us forms a d rected grap referenc ng *n*-d mens ona reg ons of pr mary data as we as ot er annotat ons, n w c nodes are abe ed w t feature structures prov d ng t e annotat on content. Forma y, LAF cons sts of:

- A data mode for annotat ons based on d rected grap s def ned as fo ows: A grap of annotat ons *G* s a set of vert ces *V*(*G*)[7] and a set of edges *E*(*G*). Vert ces and edges may be abe ed w t one or more features. A feature cons sts of a quadrup e (*G', VE, K, V*) w ere, *G'* s a grap , *VE* s a vertex or edge n *G'*, *K* s t e name of t e feature and *V* s t e feature va ue.
- A *base segmentation* of pr mary data t at def nes edges between v rtua nodes ocated between eac "c aracter" n t e pr mary data.[8] T e resu t ng grap *G* s treated as an *edge graph G'* w ose nodes are t e edges of *G*, and w c serve as t e eaf ("s n ") nodes. T ese nodes prov de t e base for an annotat on or severa ayers of annotat on. Mu t p e segmentat ons can be def ned over t e pr mary data, and mu t p e annotat ons may refer to t e same segmentat on.
- Ser a zat ons of t e data mode , one of w c s des gnated as t e p vot.
- Met ods for man pu at ng t e data mode .

Note t at LAF does not prov de spec f cat ons for annotat on *content categories* (.e., t e abe s descr b ng t e assoc ated ngu st c p enomena), for w c standard zat on s a muc tr c er matter. T e LAF arc tecture nc udes a *Data Category Registry* (DCR) conta n ng pre-def ned data e ements and sc emas t at may be used d rect y n annotat ons, toget er w t means to spec fy new categor es and mod fy ex st ng ones (see Ide and Romary, 2004).

6 The SynAF Metamodel

6.1 Introduction

W e prepar ng SynAF, we dent f ed some ex st ng n t at ves s ar ng a some ow common data mode t at seems to offer a good bas s for t e SynAF meta-mode (T ger and ISST for examp e, but a so a onger st of corpora as been stud ed, see De verab e D.3.1 of LIRICS). Base on t s study we strong y suggest t e adopt on of a mu t - ayered annotat on strategy nterre at ng syntact c annotat on for bot const tuency and dependency n a sound representat on sc eme. T e stud ed n t at ves are a so offer ng a qu te comp ete st of descr ptors, w c we started to "merge" nto a f rst st of cand date data-categor es, to be extended by data categor es cover ng syntact c p enomena (const tuency and dependency) for ot er anguages t en German and Ita an. Our st of cand date data categor es s presented n Annex A. TIGER and ISST are summar zed n Annexes
T e SynAF mode w be represented by UML c asses and by a set of ISO 12620 data categor es t at funct on as UML attr bute-va ue pa rs. T e data categor es are used to decorate

[7] T e word "vert ce" s er esynonym to "node".

[8] A c aracter s def ned to be a cont guous byte sequence of a spec f ed engt .For text, t e defau t s UTF-16.

t e UML c asses t at prov de a g eve v ew of t e mode . SynAF spec f cat ons n t e form of textua descr pt ons t at descr be t e semant cs of t e mode ng e ements prov de more comp ete nformat on about t e SynAF c asses, re at ons ps, and extens ons t an can be nc uded n t e UML d agram. Deve opers s a def ne a data category se ect on (DCS) as spec f ed for SynAF data category se ect on procedures (see be ow).

6.2 The SynAF diagram (to be represented in UML)

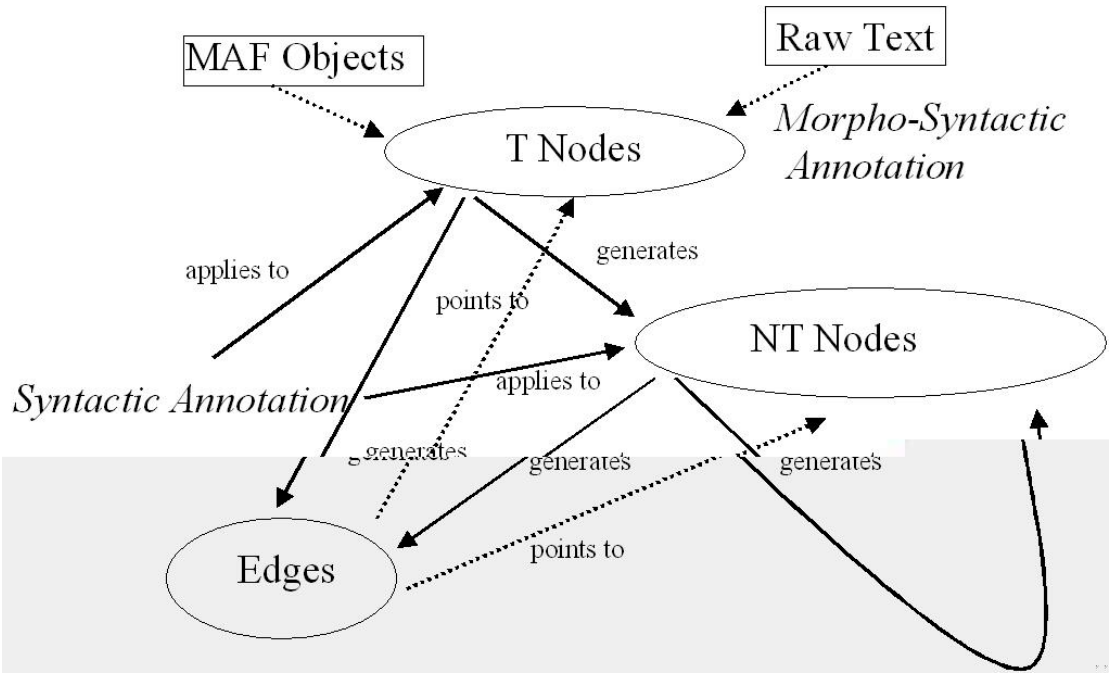


Figure 1: The SynAF metamodel

6.2.1 T Nodes class

T e *t_nodes* c ass represents t e term na nodes of a syntax tree, most y cons st ng of morp o-syntact ca y annotated words, but empty e ements are a owed. T e t_nodes are def ned over a *span.* T s can be a mu t p e span (for account ng for d scont nuous const tuents). T e t_nodes are abe ed w t syntact c categor es va d for t e word eve .

6.2.2 NT Nodes class

T e n*t_nodes* c ass represents t e non-term na nodes of a syntax tree, most y cons st ng of t_nodes and nt_nodes, but empty e ements are a owed. T e nt_nodes are a so def ned over a (poss b y mu t p e) *span.* T e nt_nodes are abe ed w t syntact c categor es va d at t e p rasa eve and g er (c ausa , sentent a).

6.2.3 Edges class

T e *Edges* c ass represents t e dependency re at on between nodes (bot term na and non-term na nodes). T e dependency re at on s a b nary one and cons sts of a abe name and a pa r of source and target nodes.

6.2.4 Syntactic Annotation class

T e *Syntactic Annotation* c ass represents t e app cat on of syntact c nformat on to MAF annotated nput. It can be e t er a manua or an automat c app cat on. W en syntact c

annotat on s app ed to nodes (non-term na or term na), t en t generates e t er a new (non-term na) node or a dependency edge.

Annex A: (informative) Data Categories for SynAF

Our strategy cons sted n co ect ng some of t e most consensua syntact c annotat on def n t ons for ga n ng a st of data categor es for const tuency (node abe s) and dependency (edge abe s) annotat on, w c w be estab s ed n t e document resu t ng from t e wor n ISO TC37/SC4 TDG 4 "Syntax". In t s document we present t e actua st of cand dates, as t ey ave been detected n annotat on n t at ves e TIGER, ISST, Spar e and EAGLES, and mod f ed/ armon zed for t e purpose of t s document. We do not quote t e spec f c or g n of eac cand date data category. We nd cate, w ere appropr ate, anguage spec f c data catagor es.

A.1Constituency

Constituency_labels	**Meaning**
AA	super at ve p rase w t am (for German)
AP	ad ect ve p rase
AVP	adverb a p rase
CAC	coord nated adpos t on
CAP	coord nated ad ect ve p rase
CAVP	Coord nated adverb a p rase
CCP	Coord nated comp ement ser
CH	C un (non-recurs ve const tuent)
CNP	Coord nated noun p rase
CO	coord nat on
CPP	Coord nated adpos t ona p rase
CVP	Coord nated verb p rase (non-f n te)
CVZ	Coord nated nf n t ve w t zu (for German)
NP	noun p rase
PN	proper noun
PP	adpos t ona p rase (prepos t ona and postpos t ona p rases
S	Sentence
VP	verb p rase (non-f n te)
VZ	nf n t ve w t zu (for German)

SPD	prepos t ona p rase *di* 'of' (for Ita an)
SPDA	prepos t ona p rase *da* 'by, from' (for Ita an)
IBAR	verba nuc eus w t f n te tense and a ad o ned

	e ements e c t cs, adverbs and negat on
SV2	nf n t va c ause
SV3	part c p a c ause
SV5	gerund ve c ause
FAC	sentent a comp ement
FS	subord nate sentence
FINT	+*wh* nterrogat ve sentence
F2	re at ve c ause
CP	d s ocated or fronted sentent a ad uncts
COMPC	copu at ve/pred cat ve comp ement

A.2Dependency

In t e fo ow ng we present t e cand date data categor es for dependency structures (t e abe s of edges n t e annotat on grap). Source of nsp rat on ere were t e Spar e and t e T ger tagsets for dependency. We use a so some examp es ta en from Spar e (t e s ort be ow some data categor es.)

mod: nd cates t e word ntroduc ng t e dependent n a ead-mod f er re at on

mod(of,g ft,boo) t e g ft of a boo

mod(by,g ft,Peter) t e g ft of a boo by Peter

mod(of,exam nat on,pat ent) t e exam nat on of t e pat ent

mod('s,doctor,exam nat on) t e doctor's exam nat on of t e pat ent

cmod, xmod, ncmod: C ausa and non-c ausa mod f ers may (opt ona y) be d st ngu s ed by t e use of cmod / xmod, and ncmod respect ve y, eac w t t e same s ots as **mod**. T e GR cmod s for w en t e ad unct s contro ed from w t n, and xmod for contro from w t out t e const tuent under cons derat on.

cmod(because,eat,be) e ate t e ca e because e was ungry

xmod(w t out,eat,as) e ate t e ca e w t out as ng

subj: nd cates t e sub ect n t e grammat ca re at on Sub ect-Pred cate. T e re at on between a pred cate and ts sub ect; w ere appropr ate, t e **initial_gr** nd cates t e syntact c n between t e pred cate and sub ect before any GR-c ang ng process.

sub (arr ve,Jo n,_) Jo n arr ved n Par s

sub (emp oy,M crosoft,_) M crosoft emp oyed 10 C programmers

sub (emp oy,Pau ,ob) Pau was emp oyed by M crosoft

W t pro-drop anguages suc as Ita an, w en t e sub ect s not overt y rea sed t e annotat on s, for examp e, as fo ows:

sub (arr vare,Pro,_) arr va n r tardo '(I) arr ved ate'

W ere t e dependent s ot s f ed by t e abstract f er **pro**, w c nd cates t at person and number of t e sub ect can be recovered from t e nf ect on of t e ead verb form.

csubj, xsubj, ncsubj: T e Grammat ca Rea t ons (RL) s **csubj** and **xsubj** may be used for c ausa sub ects, contro ed from w t n, or w t out, respect ve y. **ncsubj** s a non-c ausa sub ect.

csub (eave,mean,_) t at Ne e eft w t out say ng good-bye meant s e was st angry

xsub (w n,requ re,_) to w n t e Amer ca's Cup requ res eaps of cas

dobj: Ind cates t e ob ect n t e grammat ca re at on between a pred cate and ts d rect ob ect.
dob (read,boo ,_) read boo s
dob (ma ,Mary, ob) ma Mary t e contract

iobj T e re at on between a pred cate and a non-c ausa comp ement ntroduced by a prepos t on; **type** nd cates t e prepos t on ntroduc ng t e dependent.
ob (n,arr ve,Spa n) arr ve n Spa n
ob (nto,put,box) put t e too s nto t e box
ob (to,g ve,poor) g ve to t e poor

obj2: T e re at on between a pred cate and t e second non-c ausa comp ement n d trans t ve construct ons.
ob 2(ead,dependent)
ob 2(g ve,present) g ve Mary a present
ob 2(ma ,contract) ma Pau t e contract

dependent: T e most gener c re at on between a ead and a dependent dependent(ntroducer, ead,dependent)
dependent(n, ve,Rome) Mar sa ves n Rome

Dependency Rel	ID	Definition	Parent
Adpos t ona Case Mar er	AC	Prepos t on/postpos t on n a PP, annotated as a s ster const tuent of t e determ ner, ad ect ves, noun etc	PP
Ad ect ve Component	ADC	Component of a mu t -to en ad ect ve (MTA)	MTA
Appos t on	APP	" nserted" p rase, furt er spec fy ng/mod fy ng t e ent ty descr bed by t e matr x NP.	NP PP
Adverb a p rase Component	AVC	Component of a ead- ess AVP	ADV
con unct	CJ	Const tuent part c pat ng n coord nat on	any
comparat ve con unct on	CM	L ngu st c part c es ntroduc ng a compar son n comparat ve construct ons (for examp e “grosser a s” n German)	
dat ve	DA	Dat ve ob ect/`free dat ve' (for anguages av ng t s case n t e morp o ogy/syntax))	S VP AP AVP
ead	HD	T e ma n e ements n a nd of cons tutents	S VP AP AVP
postnom na mod f er	MNR	Postnom na NP/PP mod f er	NP PP
negat on	NG	t e negat on part c e `n c t' (a so mod f ed)	any
gen t ve ob ect	OG	Gen t ve ob ects of verbs, part c p es and certa n ad ect ves (for anguage av ng t e gene t ve case n	

		t e morp o ogy/syntax)	
pred cate	PD	Pred cat ve AP/NP/PP, typ ca y n a copu ar construct on	S VP
morp o og ca part c e	PM	two cases: t e nf n t va `zu' (zu ge en) t e ad ect va (super at ve) `am' (am besten)	VZ AA
re at ve c ause	RC		NP PP S VP AP

Annex B (informative) Annotation example

T e fo ow ng examp e s ows ow a mu t - ayered approac to syntact c annotat on can be encoded n XML. T e tagset n use s not po nt ng yet to t e data categor es, but suc a n ng w be nc uded n t e next vers on of t e document. T e Grammat c Funct ons (dependenc es) are encoded ere w t n t e "c ause" XML e ements. T e dependenc es w t n const tuents are not annotated exp c t y n t s examp e.

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<document id="ww92-short.xml" lang="de">
<paragraph id="p1" corresp="">
 <sentence id="s4" corresp="">
  <clauses>
    <clause id="cl1" from="c1" to="c22" pred_struct="c18+1"
pred_subcat_stem="sei" GF_Subj="c19+1" NP_List="c19+1"
VG_List="c18+1"/>
    <clause id="cl2" from="c23" to="c26" pred_struct="c22+1"
pred_subcat_stem="find" GF_Subj="c24+1" GF_Acc_Obj="c23+1 "
NP_List="c23+1 c24+1" VG_List="c22+1"/>
  </clauses>
  <chunks>
    <chunk id="c1" from="1" to="1" type="VG"/>
    <chunk id="c2" from="2" to="3" type="NP"/>
    <chunk id="c3" from="4" to="4" type="W"/>
    <chunk id="c4" from="5" to="5" type="W"/>
    <chunk id="c5" from="6" to="6" type="VG"/>
    <chunk id="c6" from="7" to="7" type="NP"/>
    <chunk id="c7" from="8" to="10" type="NP"/>
    <chunk id="c8" from="11" to="11" type="W"/>
  </chunks>
  <text>
    <token id="1" infl="[92 93 94 95]" pos="3" lemma="sei"
tc="22">Sind</token>
    <token id="2" infl="[2 5 20 6 13 23 9 16]" pos="8"
lemma="kein" tc="21">keine</token>
    <token id="3" infl="[6 7 8 9]" pos="1" lemma="angabe"
tc="22">Angaben</token>
    <token id="4" infl="[25]" pos="5" lemma="erhaeltlich"
tc="21">erhaeltlich</token>
    <token id="5" infl="[439]" pos="21" lemma=","
tc="1">,</token>
    <token id="6" infl="[204 205 209 206]" pos="2"
lemma="find" tc="21">findet</token>
    <token id="7" infl="[423 431 428 432 424 433 430 434]"
pos="11" lemma="sich" tc="21">sich</token>
    <token id="8" infl="[10 12]" pos="7" lemma="d-det"
tc="21">das</token>
    <token id="9" infl="[10 11 12 13 14 16]" pos="1"
lemma="kuerzel" tc="22">Kuerzel</token>
    <token id="10" tc="19">KA</token>
    <token id="11" infl="[440]" pos="21" lemma="."
tc="1">.</token>
```

```
  </text>
 </sentence>
 <sentence id="s5" corresp="">
  <clauses>
    <clause id="cl1" from="c1" to="c30" pred_struct="c27+1"
pred_subcat_stem="bedeut" GF_Subj="c26+1" NP_List="c26+1"
VG_List="c27+1"/>
  </clauses>
  <chunks>
    <chunk id="c1" from="1" to="2" type="NP"/>
    <chunk id="c2" from="3" to="3" type="VG"/>
    <chunk id="c3" from="4" to="14" type="SUBORD_CLAUSE"/>
    <chunk id="c4" from="15" to="15" type="W"/>
  </chunks>
  <text>
    <token id="1" infl="[17 10 12]" pos="8" lemma="ein"
tc="22">Ein</token>
    <token id="2" infl="[17 18 19]" pos="1" lemma="strich"
tc="22">Strich</token>
    <token id="3" infl="[204 205 209 107 206]" pos="2"
lemma="bedeut" tc="21">bedeutet</token>
    <token id="4" tc="1" lemma="," infl="[439]"
pos="21">,</token>
    <token id="6" tc="21" lemma="dass" pos="20">dass</token>
    <token id="8" tc="21" lemma="zutreff" infl="[204]"
pos="2">zutrifft</token>
    <token id="9" tc="21" lemma="nicht" pos="22">nicht</token>
    <token id="11" tc="22" lemma="wert" infl="[18 20 21 23]"
pos="1">Werte</token>
    <token id="12" tc="21" lemma="d-det" infl="[2 5 20 6 13 23
9 16]" pos="7">die</token>
    <token id="14" tc="21" lemma="nicht"
pos="22">nicht</token>
    <token id="16" tc="21" lemma="vergleichbar" infl="[25]"
pos="5">vergleichbar</token>
    <token id="18" tc="21" lemma="sei" infl="[92 93 94 95]"
pos="3">sind</token>
    <token id="20" tc="21" lemma="oder" pos="19">oder</token>
    <token id="22" tc="22" lemma="kriterium" infl="[10 11 12]"
pos="1">Kriterium</token>
    <token id="23" tc="21" lemma="d-det" infl="[10 12]"
pos="7">das</token>
    <token id="25" tc="22" lemma="spalt" infl="[18 20 21 23]"
pos="1">Spalte</token>
    <token id="26" tc="21" lemma="dies" infl="[17 3 4 21 7
14]" pos="7">dieser</token>
    <token id="28" tc="21" lemma="d-det" infl="[10 12]"
pos="7">das</token>
    <token id="29" tc="21" lemma="fuer" infl="[102]"
pos="23">fuer</token>
    <token id="30" tc="22" lemma="unternehmen" infl="[10 11 12
13 14 15 16]" pos="1">Unternehmen</token>
    <token id="31" infl="[440]" pos="21" lemma="."
tc="1">.</token>
  </text>
 </sentence>
```

```
 <sentence id="s6" corresp="">
  <clauses>
    <clause id="cl1" from="c1" to="c37" pred_struct="c31+1"
pred_subcat_stem="bezieh" GF_Subj="c30+1" GF_Acc_Obj="c32+1 "
PP_Adjunkt="c35+1" NP_List="c30+1 c32+1" PP_List="c33+1 c34+1
c35+1" VG_List="c31+1"/>
  </clauses>
  <chunks>
    <chunk id="c1" from="1" to="3" type="NP"/>
    <chunk id="c2" from="4" to="4" type="VG"/>
    <chunk id="c3" from="5" to="5" type="NP"/>
    <chunk id="c4" from="6" to="8" type="PP"/>
    <chunk id="c5" from="9" to="10" type="PP"/>
    <chunk id="c6" from="11" to="11" type="PP"/>
    <chunk id="c7" from="12" to="12" type="W"/>
  </chunks>
  <text>
    <token id="1" infl="[2 5 20 6 13 23 9 16]" pos="7"
lemma="d-det" tc="22">Die</token>
    <token id="2" infl="[41 42 43 44 45 46 47 48 49 50 51 52
53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73
74 75 76 77 78 79 80 81 82 83 84]" pos="6" lemma="angeb"
tc="21">angegebenen</token>
    <token id="3" infl="[18 20 21 23]" pos="1" lemma="wert"
tc="22">Werte</token>
    <token id="4" infl="[92 93 94 95 96 97 98 99 100 101]"
pos="2" lemma="bezieh" tc="21">beziehen</token>
    <token id="5" infl="[423 431 428 432 424 433 430 434]"
pos="11" lemma="sich" tc="21">sich</token>
    <token id="6" infl="[24 102]" pos="23" lemma="in"
tc="21">in</token>
    <token id="7" infl="[17 3 4 21 7 14]" pos="7" lemma="d-
det" tc="21">der</token>
    <token id="8" infl="[2 3 4 5]" pos="1" lemma="regel"
tc="22">Regel</token>
    <token id="9" infl="[24 102]" pos="23" lemma="auf"
tc="21">auf</token>
    <token id="10" infl="[20 21 23]" pos="1" lemma="abschluss"
tc="22">Abschluesse</token>
    <token id="11" tc="11">31.</token>
    <token id="12" tc="11">1991</token>
    <token id="13" tc="21" lemma="zum" infl="[24]"
pos="23">zum</token>
    <token id="14" tc="22" lemma="dezember" infl="[17 18 19 20
21 22 23]" pos="1">Dezember</token>
    <token id="15" infl="[440]" pos="21" lemma="."
tc="1">.</token>
  </text>
 </sentence>
 <sentence id="s7" corresp="">
  <clauses>
   </clauses>
  <chunks>
    <chunk id="c1" from="1" to="8" type="PP"/>
    <chunk id="c2" from="9" to="9" type="VG"/>
    <chunk id="c3" from="10" to="15" type="PP"/>
```

```
    <chunk id="c4" from="16" to="16" type="VG"/>
    <chunk id="c5" from="17" to="17" type="W"/>
  </chunks>
  <text>
    <token id="1" infl="[24 102]" pos="23" lemma="in"
tc="22">In</token>
    <token id="2" infl="[41 42 43 44 45 46 47 48 49 50 51 52
53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73
74 75 76 77 78 79 80 81 82 83 84]" pos="5" lemma="einig"
tc="21">einigen</token>
    <token id="3" infl="[22]" pos="1" lemma="fall"
tc="22">Faellen</token>
    <token id="4" pos="20" lemma="wie" tc="21">wie</token>
    <token id="5" pos="22" lemma="beispielsweise"
tc="21">beispielsweise</token>
    <token id="6" tc="22" lemma="siemens" infl="[10 0 11 12 13
14 16]" pos="1">Siemens</token>
    <token id="7" pos="19" lemma="oder" tc="21">oder</token>
    <token id="8" tc="19">MAN</token>
    <token id="9" infl="[204]" pos="2" lemma="werd"
tc="21">wird</token>
    <token id="10" infl="[102]" pos="23" lemma="per"
tc="21">per</token>
    <token id="11" infl="[10 11 12]" pos="1" lemma="ende"
tc="22">Ende</token>
    <token id="12" infl="[17 18 19 20 21 23]" pos="1"
lemma="september" tc="22">September</token>
    <token id="13" pos="17" lemma="beziehungsweise"
tc="21">beziehungsweise</token>
    <token id="14" infl="[10 11 12]" pos="1" lemma="ende"
tc="22">Ende</token>
    <token id="15" infl="[17 18 19]" pos="1" lemma="juni"
tc="22">Juni</token>
    <token id="16" infl="[204 205 107 206]" pos="2"
lemma="bilanzier" tc="21">bilanziert</token>
    <token id="17" infl="[440]" pos="21" lemma="."
tc="1">.</token>
  </text>
 </sentence>
…
</Paragraph>
</document>
```

Bibliography

Abe é A., S. Hansen-Sc rra & H. Usz ore t (eds.), 2003. Proceed ngs of t e 4t Internat ona Wor s op on L ngu st ca y Interpreted Corpora (LINC-03).

Ca zo ar N., Mc Naug t J., Zampo A. 1996 Eag es, ed tors ntroduct on. ttp://www. c.cnr. t/EAGLES96/ed ntro/ed ntro. tm

Ca zo ar N., Bertagna F., Lenc A., Monac n M. ed tors, 2003. Standards and best Pract ce for Mu t ngua Computat ona Lex cons. MILE (T e Mu t ngua ISLE Lex ca Entry). ISLE CLWG De verab e D2.2 & 3.2 P sa.

Rumbaug J., Jacobson I., Booc G. T e un f ed mode ng anguage reference manua , second ed t on Add son Wes ey 2004

S. Montemagn , F. Barsott , M. Batt sta, N. Ca zo ar , A. Lenc O. Corazzar , A. Zampo , F. Fanc u , M. Massetan , R. Bas R. Raffae , M.T. Paz enza, D. Sarac no, F. Zanzotto, F. P anes N. Mana, and R. De monte. Bu d ng t e Ita an Syntact c-Semant c Treeban . In Anne Abe é, ed tor, Bu d ng and Us ng syntact ca y annotated corpora, pages 189--210. K uwer, Dordrec t, 2003.

Nancy Ide and Laurent Romary, 2004. A Reg stry of Standard Data Categor es for L ngu st c Annotat on. Proceed ngs of t e Fourt Language Resources and Eva uat on Conference (LREC), L sbon, 135-39.

Nancy Ide and Laurent Romary, 2004. Internat ona Standard for a L ngu st c Annotat on Framewor . Journa of Natura Language Eng neer ng, 10:3-4, 211-225.

Nancy Ide and Laurent Romary, 2006. Represent ng L ngu st c Corpora and T e r Annotat ons. Proceed ngs of t e F ft Language Resources and Eva uat on Conference (LREC), Genoa, Ita y.

Nancy Ide, GrAF: A Grap -based Format for L ngu st c Annotat ons. Proceed ngs of t e LAW Wor s op at ACL 2007, Prague.

T e EAGLES n t at ve: ttp://www. c.cnr. t/EAGLES96/ ome. tm

T e LIRICS Pro ect: ttp:// r cs. or a.fr

T e SPARKLE Pro ect: ttp://www. c.cnr. t/spar e/spar e. tm

T e TIGER pro ect: ttp://www. ms.un -stuttgart.de/pro e te/TIGER/TIGERCorpus/