RUHR UNIVERSITÄT BOCHUM

RUB

F-AG 7: Angewandte Sprachwissenschaft, Computerlinguistik

Kurationsprojekt 2

Linguistische Annotation von Nichtstandardvarietäten — Guidelines und „Best Practices"

Guidelines Koreferenz

Version 1.1

Stand: 21.11.13

Marc Reznicek

Marc.Reznicek@hu-berlin.de

Die folgenden Guidelines bauen hauptsächlich auf die Vorarbeiten der folgenden Arbeiten auf

- TüBa-D/Z (Naumann 2007)
- PoCos erweitertes Schema (Kaupat et al. 2013)

Alle grundlegenden Kategorien und Annotationsanweisungen stammen aus diesen beiden Quellen und wurden danach für die Strukturen in den NoSta-D Nicht-Standard-Subkopora erweitert. Sie kann daher eher als Anlage, denn als komplett eigenständiges Annotationshandbuch verstanden werden. Die wichtigen Begriffe, Regeln und Prinzipien werden hier zwar erwähnt, für die Annotation wird die Kenntnis der beiden genannten Dokumente allerdings vorausgesetzt.

Inhalt

1 Koreferenz

Koreferenz bezeichnet die Relation zwischen zwei Textelementen, die auf dieselbe Entität (denselben Diskursreferenten) verweisen.

Dabei unterscheiden wir zwischen:

- anaphorischer Koreferenz
- nicht-anaphorische Koreferenz

1.1. Nicht-anaphorische Koreferenz:

Zwei referierende Ausdrücke, deren Diskursreferenten **unabhängig voneinander interpretierbar** sind, stehen in einer **nicht-anaphorischen Koreferenz-Relation**.

1) [$_1$ ***Die deutsche Kanzlerin***] *hat sich gegen große Reformen entschieden.* [$_2$ **Angela Merkel**] *hat eine wichtige Chance verpasst.*

1.2. Anaphorische Koreferenz:

Wenn ein **referierender Ausdruck** allein nicht interpretierbar ist, sondern sein **Diskursreferent** sich erst unter Rückgriff auf einen **zuvor genannten koreferenten Ausdruck** ermitteln lässt, handelt es sich um eine **Anapher**. Der vorangehende koreferente Ausdruck ist dann das zugehörige **Antezedens**.

Anapher:

2) *Susanne mag* [$_1$ ***das Turnen***] *nicht, denn* [$_2$ ***es***] *ist sehr schwer.*

Wenn ein **referierender Ausdruck** allein nicht interpretierbar ist, sondern sein **Diskursreferent** sich erst unter Rückgriff auf einen **in der Folge genannten koreferenten Ausdruck** ermitteln lässt, handelt es sich um eine **Katapher**. Der nachfolgende koreferente Ausdruck ist dann das zugehörige **Antezedens**.

Katapher:

3) [$_1$ [$_2$ ***Sein***] allererstes *Gedicht*] *wollte* [$_3$ ***Michael Stich***] *dann doch nicht vorlesen.*

1.3. Referentielle Kette

Wir nennen die Serie der Erwähnungen desselben Referenten eine referentielle Kette. Ob ein referierender **Ausdruck *e*** zu einer **Kette *k*** gehört, kann mit einem Ersetzungstest geprüft werden: Wenn für jedes **Substantiv *s*** (Nomen, Eigenname) in ***k*** gilt, dass die Ersetzung von **e** durch ***s*** die Interpretation des Textes nicht verändert, so gehört **e** zur **Kette *k*** und es ist eine **Koreferenz-Relation** zum letzten Element der Kette zu annotieren.

1.4. Nominale Diskursreferenten

Dieses Annotationsschema beschränkt sich auf **nominale Diskursreferenten**. Das bedeutet, es werden nur referentielle Ketten behandelt, deren Komponenten jeweils durch eine (im weiten Sinne) Nominalphrase ausgedrückt sind. Relationen, die durch sog. Sachverhaltsanaphern gestiftet werden, bei denen das Antezedens eines Pronomens eine Verbalphrase, ein Satz oder auch ein ganzer Teiltext sein können, werden hier nicht behandelt.

2 Annotation

Die Annotation von Koreferenzrelationen in NoSta-D verläuft in vier Schritten.

1) Zuerst werden alle definiten Nominalphrasen (inklusive Personal-, Demonstrativ-, Relativpronomen) als sogenannte „primäre Referenzausdrücke“ mit dem Label „CM“ (chain member) annotiert. Sie dienen als Ausgangspunkte für die Annotation der im Text enthaltenen Refefenzketten.

2) Für alle primären Refer*enzausdrücke* werden die passenden Referenzrelationen zu ihren Antezendenten annotiert.
   a) Jedes Mal, wenn ein primärer Refer*enzausdruck* auf einen Ausdruck verweist, der selbst kein primärer Referenzausdruck ist, wird dieser als sog. „sekundärer Referenzausdruck“ ebenfalls mit „CM“ annotiert.
3) Referiert ein Ausdruck auf eine Gruppe, für die es geteilte Antezendens gibt, so werden die einzelnen Mitglieder der Gruppe als „GROUP“ annotiert.
4) Zuletzt werden alle primären Referenzausdrücke gelöscht, die keinen Antezendens haben.

3 Referenzausdrücke (RA = Markables)

Referenzausdrücke sind maximale NPen inklusive

1) Komplementen
   [$_1$ *Die Frage nach der Ehre*]
2) Attributen
   [$_1$ die ihn störende Ungenauigkeit seiner Kollegen heute]
3) Appositionen
   [$_1$ Ihr Geschäft, eine Privatbank im Zentrum von New York]

Referenzausdrücke können ineinander verschachtelt sein.

   a) Mit Eigennamen
      [$_1$ [$_2$ Angela Merkels] wenig gefeiertes Kabinett]

Verweisen Teile eines Referenzausdrucks auf den gleichen Referenten, wie der gesamte Ausdruck, wird der eingebettete nicht auch noch annotiert. Das bedeutet, dass Appositionen keine eigene Referenz tragen. Im folgenden Beispiel wird „Hans Taake“ nicht als eigener Referenzausdruck annotiert.

Im Januar hat die AWO [$_1$ **ihren Geschäftsführer Hans Taake**] *fristlos entlassen*. [$_2$ **Er**] *musste gehen.*

4 Koreferenzrelationen

Die in NoSta-D annotierten Referenzrelationen, folgen dem Namen nach zwar der ursprünglichen Kategorisierung nach MUCC, der „Koreferenz“-Analyse liegt aber ein Diskurs-Modell zugrunde[1].

[1] *MUCCS was designed to encode information deemed useful for a subtask of information extraction, and the instructions provided to annotators were meant to ensure that all information provided by a text about a certain entity would be marked using a single device, the **IDENT relation**. As van Deemter and Kibble (2000) point out,*

Es werden nur Relationen annotiert, die sich auf die gleichen Referenten beziehen, nicht auf darüber hinausgehenden rein semantische Referenzen wie beispielsweise „Teil-Ganzes“- Beziehungen oder Metonymien (vgl. *Naumann 2007:1*).

Eine indefinite NP kann ein indefinites Antezedens besitzen. Im Gegensatz zum Vorgehen in Nauman (2007) wird hier eine Relation etabliert, wenn die Zuordnung aus dem Kontext möglich ist.

> [1 *Ein Bundeswehrsoldat*] *ist am Dienstag in Griechenland von* [2 *einem Nato-Gegner*] *verletzt worden. Medienberichten zufolge schleuderte* [3 *ein 29Jahre alter Grieche*] *in Salonica seinen Helm auf die Windschutzscheibe eines Militärfahrzeugs, das* [4 *der Soldat*] *fuhr*. (*Naumann 2007:4*).

Anaphorische Relationen werden immer über kataphorischen bevorzugt. Koreferenzen verweisen immer nach links.

4.1. Relationstypen

Die folgenden Relationstypen werden in NoSta-D vergeben.

4.1.1 coreferential

Eine Relation wird dann als „coreferential“ zwischen einem Ausdruck und seinem Antezedens etabliert, wenn die Referenz des ersten auch ohne die des letzteren erschließbar ist.

coreferential

4) *Der Vorhang geht wieder auf im* [1 **Metropol**]. *Kultursenator will* [2 ***das Theater am Nollendorfplatz***] *an Privatinvestor verkaufen.*

4.1.2 anaphoric

Eine Relation wird dann als „anaphoric“ zwischen einem Ausdruck und seinem Antezedens etabliert, wenn die Referenz des ersten nur unter Rückgriff auf die Referenz des letzteren erschließbar ist. Dabei wird davon ausgegangen, dass ein Antezedens, das selbst eine Anapher ist, seine Referenz bereits etabliert hat. Prototypischer Weise wird eine Anapher durch ein Pronomen ausgedrückt

anaphoric

[1 ***Ein klarer Ton***] *breitet sich*[2] *aus, warm und satt, bis* [3 ***er***] *den ganzen Saal erfüllt.*

Auch definite NPen sind sehr oft Anpahern.

anaphoric

Der Vorhang geht wieder auf im [1 **Metropol**]. *Kultursenator will* [2 ***das Theater***] *an Privatinvestor verkaufen.*

---

however, the result is rather ad hoc; the IDENT relation as defined by the instructions doesn't capture any coherent definition of 'coreference' […] The MATE proposals, by contrast, while still labeled as proposals for 'coreference annotation,' because the name has become a de facto standard as a result of the MUC initiative, are explicitly based on the DISCOURSE MODEL assumption adopted almost universally by linguists (computational and not) working on anaphora resolution and generation.(Poesio 2004:154f.)

[2] „sich“ wird hier nicht annotiert, da es lexikalisch vom Verb „ausbreiten“ gefordert wird und daher keine Referenz im eigentlichen Sinne trägt. Siehe Punkt 5.2.x

4.1.3 cataphoric

Eine Relation wird dann als „cataphoric“ zwischen einem Ausdruck und seinem Antezedens etabliert, wenn die Referenz des ersten nur unter Rückgriff auf die Referenz des letzteren erschließbar ist. Hierbei muss außerdem gelten, dass der Referenzausdruck links von seinem Antezedens realisiert ist. In jedem Fall ist eine mögliche anaphorische Interpretation über eine kataphorische zu bevorzugen.

[1 ***Er***] *hatte noch keine Ahnung wohin es gehen sollte.* [2 ***Lucky Luke***] *ritt der Abendsonnte entgegen.*

4.1.4 bound (=TüBa-DZ; MATE: bound anaphors)

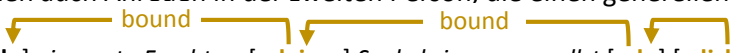
Ist der Referent eines Ausdrucks **RA** koreferent mit seinem Antezedens, ist dessen Referent allerdings nicht erschließbar, wird die Relation als „bound“ annotiert[3].

[1 **Niemand**] verliert gerne [2 **seinen**] *Arbeitsplatz.*

[1 **Wer**] *einen Sitzplatz haben will,* [2 **der**] *muß um 19 Uhr, wer im Treppenhaus noch etwas hören will, sollte spätestens um 20 Uhr da sein.*

Berücksichtigt [1 **man**] *den weiblichen Alkoholismus, kann* [2 **man**] *den Anpassungsdruck erahnen.*

Hierunter fallen auch Anreden in der zweiten Person, die einen generellen Charakter haben.

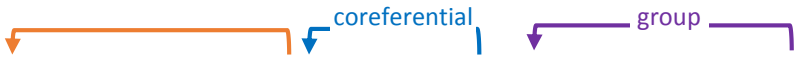


Und willst [1 **du**] *eine gute Frucht zu* [2 **deiner**] *Seele bringen, so sollst* [3 **du**] [4 **dich**] *üben in guten werken.*
(NoSTa-D_AnselmBerlin)

4.1.5 group (≠TueBa; Kaupat: groups)

Referenzrelationen, die zwischen mehreren Referenzaudrücken und einem gemeinsamen Antezedens bestehen, werden als Gruppe annotiert. Dafür werden zuerst alle gemeinsam referierenden Ausdrücke auf der Ebene der Markables als „GROUP“ annotiert und dann miteinander mit er „GROUP“-Relation verbunden. Folgt das Antezedens dem letzten Element der Gruppe, so ist dieses der *Gruppenanker*. In allem anderen Fällen ist das erste Element der Gruppe der *Gruppenanker*. Die Relation zwischen der Gruppe und dem Antezedens wird dann regulär zwischen dem Gruppenanker und dem Antezedens hergestellt. Im folgenden Beispiel ist **RA3** der Gruppenanker und verweist per „**coreferential**“-Relation auf **RA2** „das“.

Kontrolliert werden die Geschäftsführer von …



… [1 **den Revisoren des AWO-Landesverbandes**], [2 **das**] *sind* [3 **Detlev Griesche**] *und* [4 **Karin Freudenthal**]

[3] Im Gegensatz zu einer Koreferenzbeziehung im eigentlichen Sinne, bei der die einzelnen RAs auf den gleichen Referenten außerhalb des Textes verweisen (coreferential) , bzw. ein etablierter Referent übernommen wird (anaphoric/cataphoric), werden hier nur die Ausdrücke miteinander gekopppelt (vgl. Naumann 2007:6)

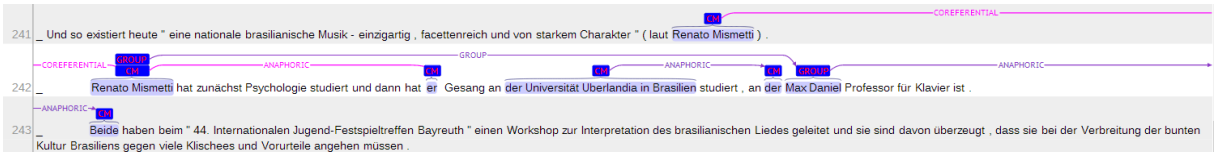


Abbildung 1: Gruppenreferenzen in NoSta-D (NoSta-D_tuebadz-r8.0:241-143)

Der Umgang mit Gruppen in NoSta-D ersetzt die Relation *split_antecedent* in Naumann (2007).

4.1.6 group_antecedent (=TüBa-DZ; MATE: element & evoked entities)

Verweist ein Referenzausdruck auf einen Referenten aus einer größeren Gruppe, so wird die Relation mit „group_antecedent“ markiert. Dies kann entweder ein RA einer annotierten Gruppe sein (siehe 0) oder ein implizites Element eines Pluralausdrucks.

group group_antecedent

[1 **das**] *sind* [2 **Detlev Griesche**] *und* [3 **Karin Freudenthal**]. [4 **Freudenthal**] *wollte gestern nichts dazu sagen*.

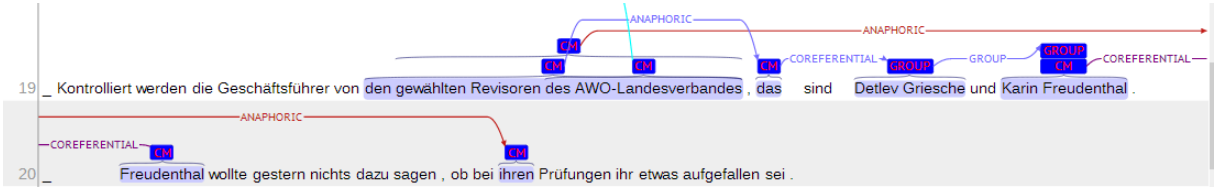


Abbildung 2: Mehrfache RA-Spannen in zukünftiger WebAnno-Version

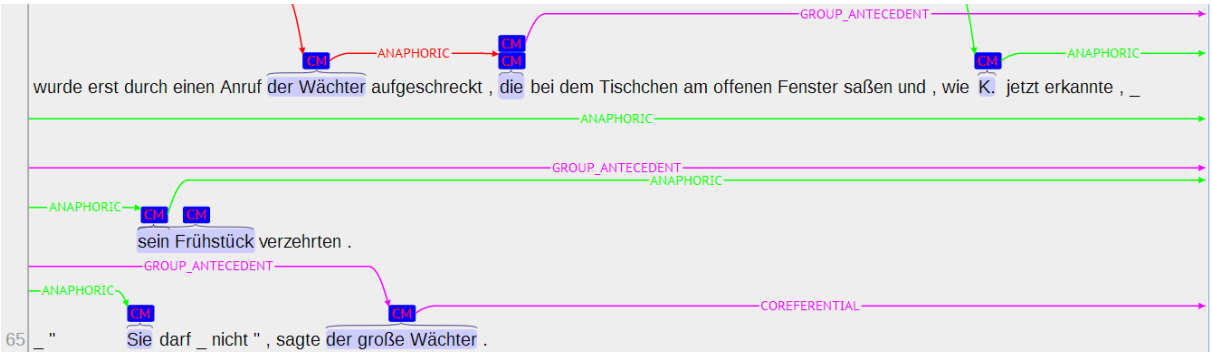


Abbildung 3: group_antecedent in NoSta-D_kafka

5 Zweifelsfälle

5.1. Koreferenzrelationen

5.1.1 Apposition

NPn innerhalb einer Apposition wird nicht innerhalb der gleichen maximalen NP referenziert.

[1 **Hulk Hogan, der letzte Ringer alter Schule**].

5.1.2 Kopulakonstruktionen

Predikationen sind keine Referenten und werden nicht annotiert.

[1 **Hulk Hogan***] ist* [2 **der letzte Ringer alter Schule**].

[1 **Hulk Hogan**] *ist* [2 **ein Ringer alter Schule**].

5.1.3 Vergleichskonstruktionen mit “als”

NPn in Vergleichskonstruktionen mit “als” werden nicht koreferenziert (vgl. Naumann 2007:13)

[1 **Hulk Hogan**] gilt als [2 **der letzte Ringer alter Schule**].

5.2. Anaphorische Relationen (=TueBa; = Kaupat)

5.1.4 Lexikalische Reflexive

Reflexivpronomen in lexikalischen Reflexivverbkonstruktionen tragen keine Referenz und werden daher nicht annotiert (vgl. Naumann 2007:37; Kaupat 2013:18).

Im Yellowstone Nationalpark ereignete [1 **sich**] *vor 100 Millionen Jahren ein Massensterben.*

5.1.5 Personalpronomen 1. und 3. Person Plural (=TüBa)

Personalpronomen, die in direkter Rede auftauchen, werden mit Referenten im Matrixsatz verbunden unabhängig von unterschieden in der Zeit.

[1 **Der Fraktionsvorsitzende der SPD**] *begrüßte dies: „* [2 **Ich**] finde es gut.“

Dies gilt nur, wenn beide Referenzausdrücke auf den gleichen Referenten verweisen. Nicht dann, wenn auch andere Referenten miteingeschlossen werden (vgl. Naumann 2007:17). Im folgenden Beispiel bleibt die Beziehung offen.

„[1 **Wir**] *bestreiten heute ein wichtiges Spiel.“, sagte* [2 **Manager Willi Lemke**].

[1 **Izet**] *deutet auf die Plastiksandalen, die* [2 **ihm**] *geblieben sind.*

„[3 **Wir**] *haben nur noch das, was* [4 **wir**] *auf dem Leibe tragen.“*

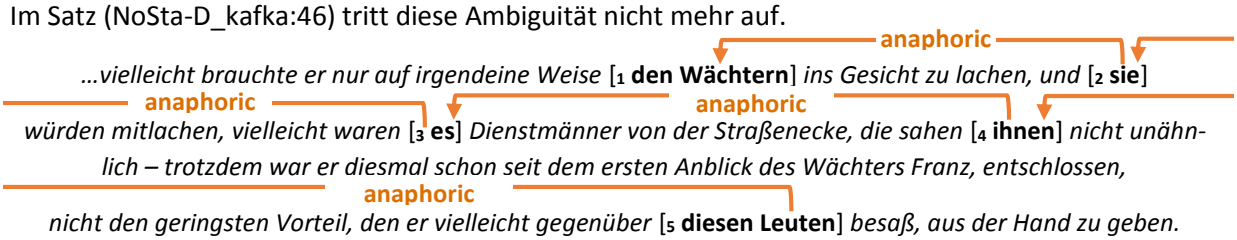
5.3.2 Gruppen

Gruppenbezeichnungen werden nur dann annotiert, wenn sie im Text unzweideutig sind.

Im folgenden Beispiel (NoSta-D_kafka:46) wird der RA „diese Leute“ nicht annotiert, da er entweder nur auf K. beide Wärter oder auf die Wärter und den Aufseher gerichtet sind.

In Gegenwart [1 **dieser Leute**] *konnte er aber nicht einmal nachdenken.*

Im Satz (NoSta-D_kafka:46) tritt diese Ambiguität nicht mehr auf.



…vielleicht brauchte er nur auf irgendeine Weise [1 **den Wächtern**] *ins Gesicht zu lachen, und* [2 **sie**] *würden mitlachen, vielleicht waren* [3 **es**] *Dienstmänner von der Straßenecke, die sahen* [4 **ihnen**] *nicht unähnlich – trotzdem war er diesmal schon seit dem ersten Anblick des Wächters Franz, entschlossen, nicht den geringsten Vorteil, den er vielleicht gegenüber* [5 **diesen Leuten**] *besaß, aus der Hand zu geben.*

5.3.3 Relativpronomen (=TüBa; ≠ Kaupat 2013:26)

Er zeigt auf [1 **die Stelle**], *an* [2 **der**] *er gestürzt war.*

„Wo“ als Relativpronomen wird regulär annotiert

5.3.4 Expletive (≠ TüBa-D/Z)

Expletive tragen keine Referenz und werden nicht annotiert.

[1 Es] *ist heute wieder wirklich kalt*

5.3.5 Vokative

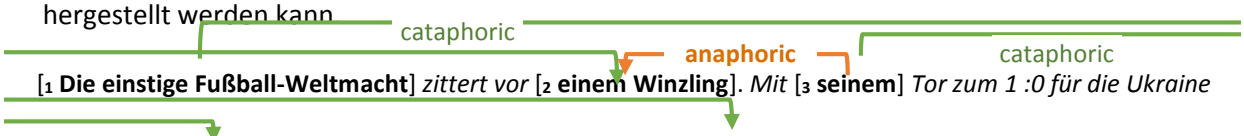
Vokative sind keine Referenzträger im Sinne des Diskursmodells und werden daher nicht mit annotiert.

lass es leiber **[1 zor]**

[2 tapfere kleine zoraaaay]

5.3.6 Diskursanaphern und -kataphern

Nominale NPen, deren Referenz erst durch den Diskurs etabliert wird, werden dann als „anaphoric" bzw. „cataphoric" mit dem nächsten Antezedens verbunden, wenn die Referenzgleichheit **eindeutig** hergestellt werden kann.



(Kaupat et al.2013:32)

5.3. Generische NPen ohne Referenz

NPen, die in generischen Kontruktionen verwendet werden, tragen keine Referenz und werden daher nicht annotiert. Als Test dient die Attribuierungsprobe. Kann man den Referenten nicht mit einem weiteren Attribut erweitern, ohne eine die Lesart zu ändern, wird keine Referenz annotiert. Im folgenden Satz ist „Kino" beispielsweise nicht durch „große" modifiziert werden, ohne dass sich die Bedeutung von „Kino" ändert.

Ich gehe gerne ins Kino	→ irgendein Kino
Ich gehe gerne ins **große** *Kino*	→ ein bestimmtes Kino
Anna soll mir das Frühstück machen.	→ irgendein Frühstück
Anna soll mir das **große** *Frühstück machen*	→ ein bestimmter Typ Frühstück

6 Literatur

Hirschman, Lynette; Chinchor, Nancy (1998): MUC-7 Coreference Task Definition: Version 3.0. In: Chinchor, Nancy; Marsh, Elaine; Perzanowski, Dennis (Hgg.), *Proceedings of the 7th Message Understanding Conference (MUC 7)*, http://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html.

Kaupat, David; Warzecha, Saskio; Stede, Manfred (2013): Koreferenz: Chapter 5: Erweiterung des PoCoS-Kernschemas.

Naumann, Karin (2007): Manual for the Annotation of In-Document Referential Relations. Seminar für Sprachwissenschaft, Abt. Computerlinguistik Universität Tübingen, http://www.sfb441.uni-tuebingen.de/a1/Publikationen/tuebadz_relations_man.pdf.

Poesio, Massimo (2004): The MATE/GNOME Proposals for Anaphoric Annotation, Revisited. In: Strube, Michael; Sidner, Candy (Hgg.), *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*. Cambridge and Massachusetts and USA: Association for Computational Linguistics, 154–162, http://sigdial.org/workshops/workshop5/proceedings/pdf/poesio.pdf.

van Deemter, Kees; Kibble, Rodger (2000): On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics* 26(4), 629–637. , http://dl.acm.org/citation.cfm?id=971882.971888.

Alle Quellen wurden am 6.11.2013 geprüft.