INEL

Eine Infrastruktur zur Dokumentation indigener nordeurasischer Sprachen

Beata Wagner-Nagy, Hanna Hedeland, Timm Lehmberg (Universität Hamburg)
Michael Rießler (Universität Freiburg)

hzsk hamburger zentrum für sprachkorpora

Institut für Finnougristik/Uralistik IFUU

INEL

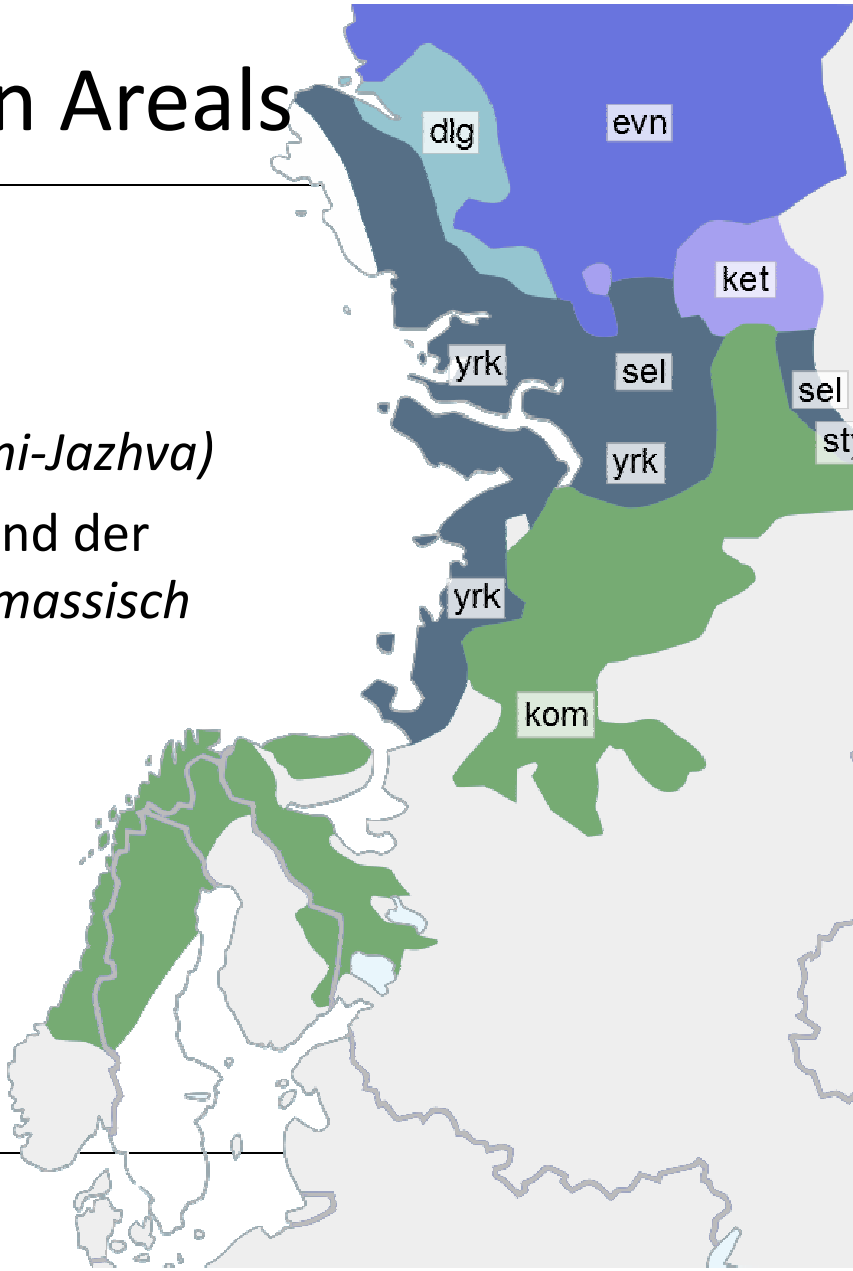
Grammatical Descriptions, Corpora, and Language Technology for Indigenous Northern Eurasian Languages

Die Sprachen des nordeurasischen Areals

- zwei Sprachfamilien + isolierte Sprache:
  - *Uralisch*
    - *Komi (2 Varietäten: Komi-Permjakisch* und *Komi-Jazhva)*
    - *Nenzisch (2 Varietäten:*Dialekten der Tajmyr- und der Kanin-Halbinsel*), Selkupisch (3 Varietäten), Kamassisch*
  - *Altaisch*
    - *Dolganisch, Evenkisch (*Norddialekt), *Sibirisch-Tatarisch*
  - *Ket*

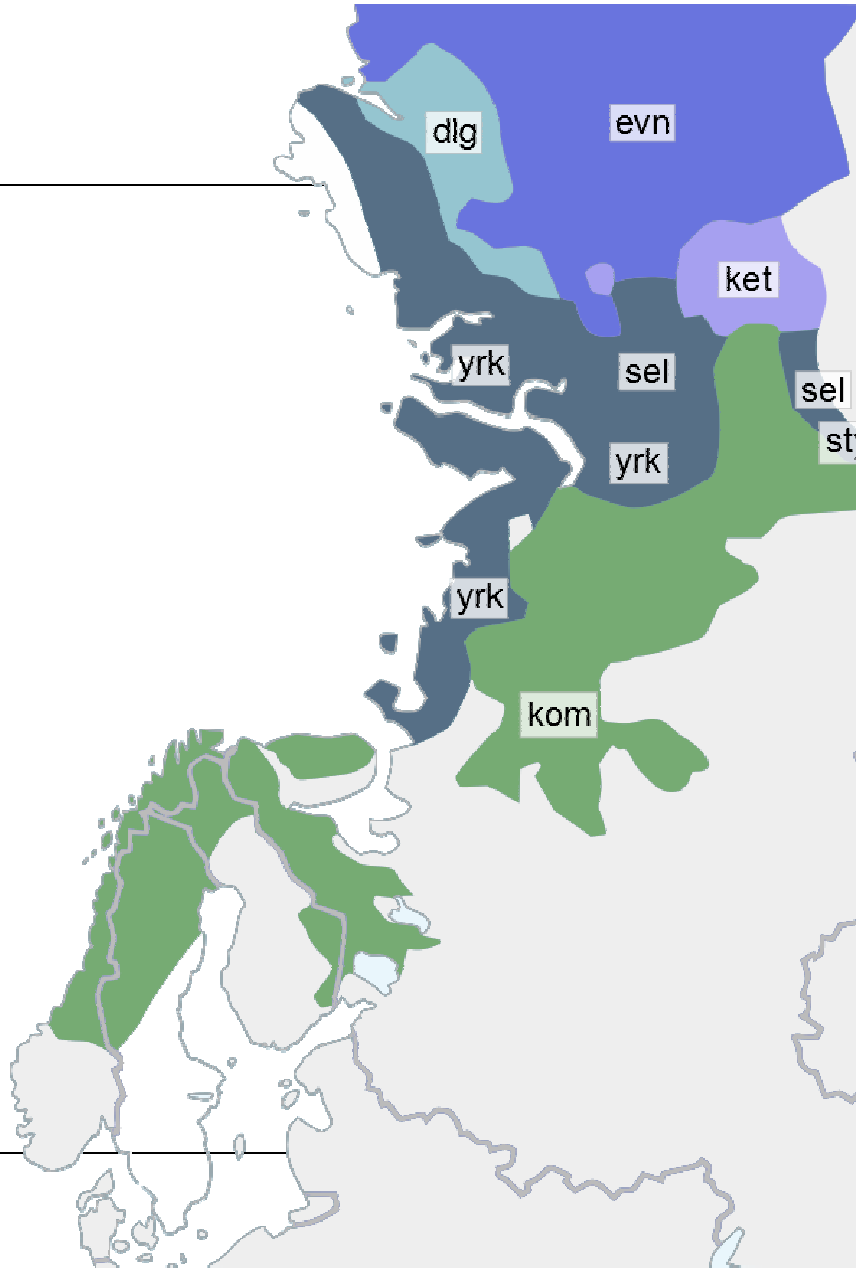
Die Sprachen des nordeurasischen Areals

- zwei Sprachfamilien + isolierte Sprache:
  - *Uralisch*
    - *Komi (2 Varietäten: Komi-Permjakisch* und *Komi-Jazhva)*
    - *Nenzisch (2 Varietäten:*Dialekten der Tajmyr- und der Kanin-Halbinsel*), Selkupisch (3 Varietäten), Kamassisch*
  - *Altaisch*
    - *Dolganisch, Evenkisch (*Norddialekt), *Sibirisch-Tatarisch*
  - *Ket*



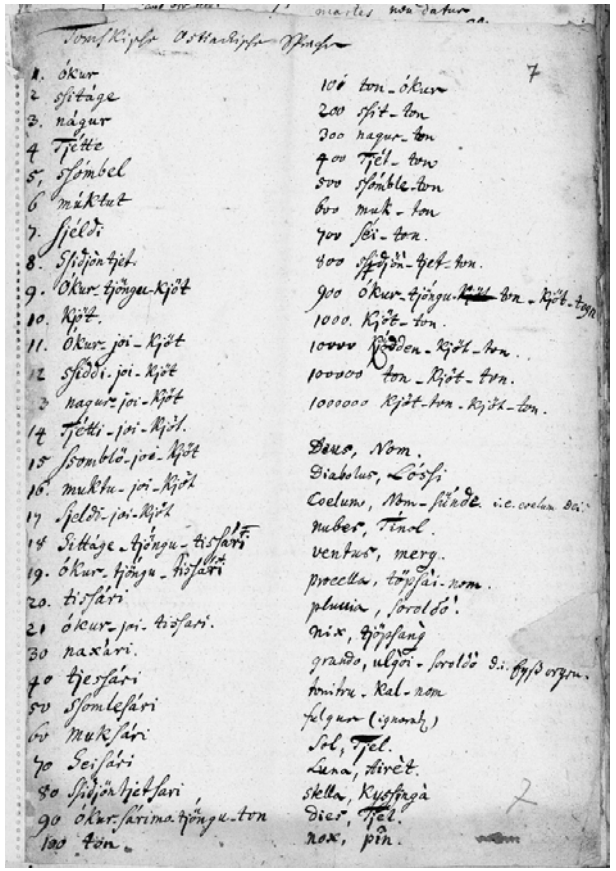
Die Sprachen im Projektfokus

xas:	Kamassisch †
ket:	Ketisch
sel:	Selkupisch
kom:	Komi
yrk:	Tundra-Nenzisch
dlg:	Dolganisch
evn:	Nord-Ewenkisch
sty:	Sibirisch-Tatarisch



Stand der linguistischen Erfassung

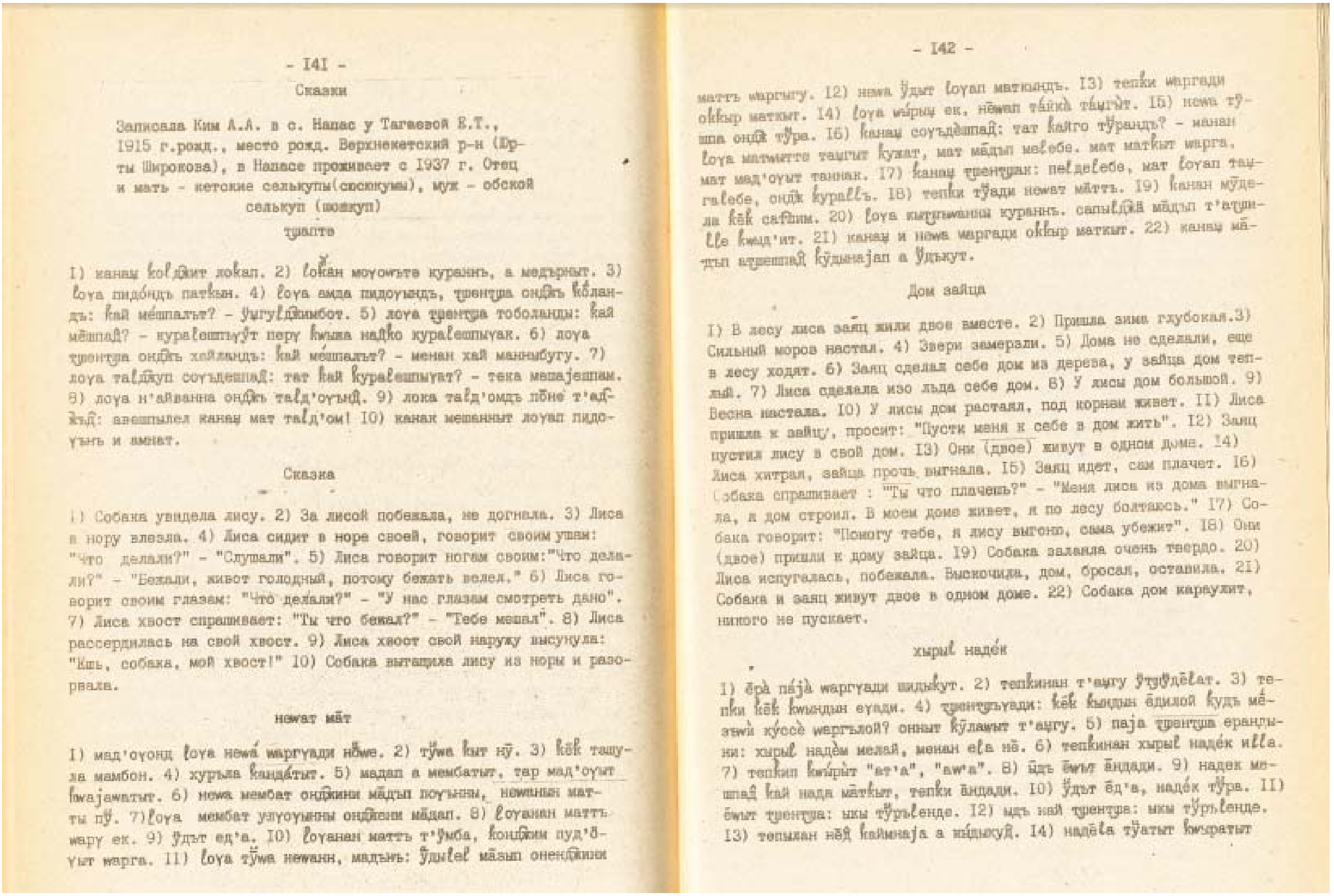
- Wortlisten vom 18. Jh.
- Textsammlungen und sprachwissenschaftliche Beschreibungen einiger Sprachen vom 19. Jh
- Textsammlungen, Grammatiken von 20. Jh



Materialsammlungen

- große Mengen an Sprachdaten in den Archiven und in Veröffentlichungen
  - Daten mit oder ohne Aufnahmen
- Archive in Tomsk, St.Petersburg, Moskau, Hamburg, Tartu

Rohdaten



- 141 -

Сказки

Записала Ким А.А. в с. Напас у Тагаевой Е.Т., 1915 г.рожд., место рожд. Верхнекетский р-н (Юрты Широкова), в Напасе проживает с 1937 г. Отец и мать - кетские селькупы(сюсюкумы), муж - обской селькуп (шошкуп)

[illegible]

Сказка

1) Собака увидела лису. 2) За лисой побежала, не догнала. 3) Лиса в нору влезла. 4) Лиса сидит в норе своей, говорит своим ушам: "Что делали?" - "Слушали". 5) Лиса говорит ногам своим:"Что делали?" - "Бежали, живот голодный, потому бежать велел." 6) Лиса говорит своим глазам: "Что делали?" - "У нас глазам смотреть дано". 7) Лиса хвост спрашивает: "Ты что бежал?" - "Тебе мешал". 8) Лиса рассердилась на свой хвост. 9) Лиса хвост свой наружу высунула: "Кшь, собака, мой хвост!" 10) Собака вытащила лису из норы и разорвала.

[illegible]

- 142 -

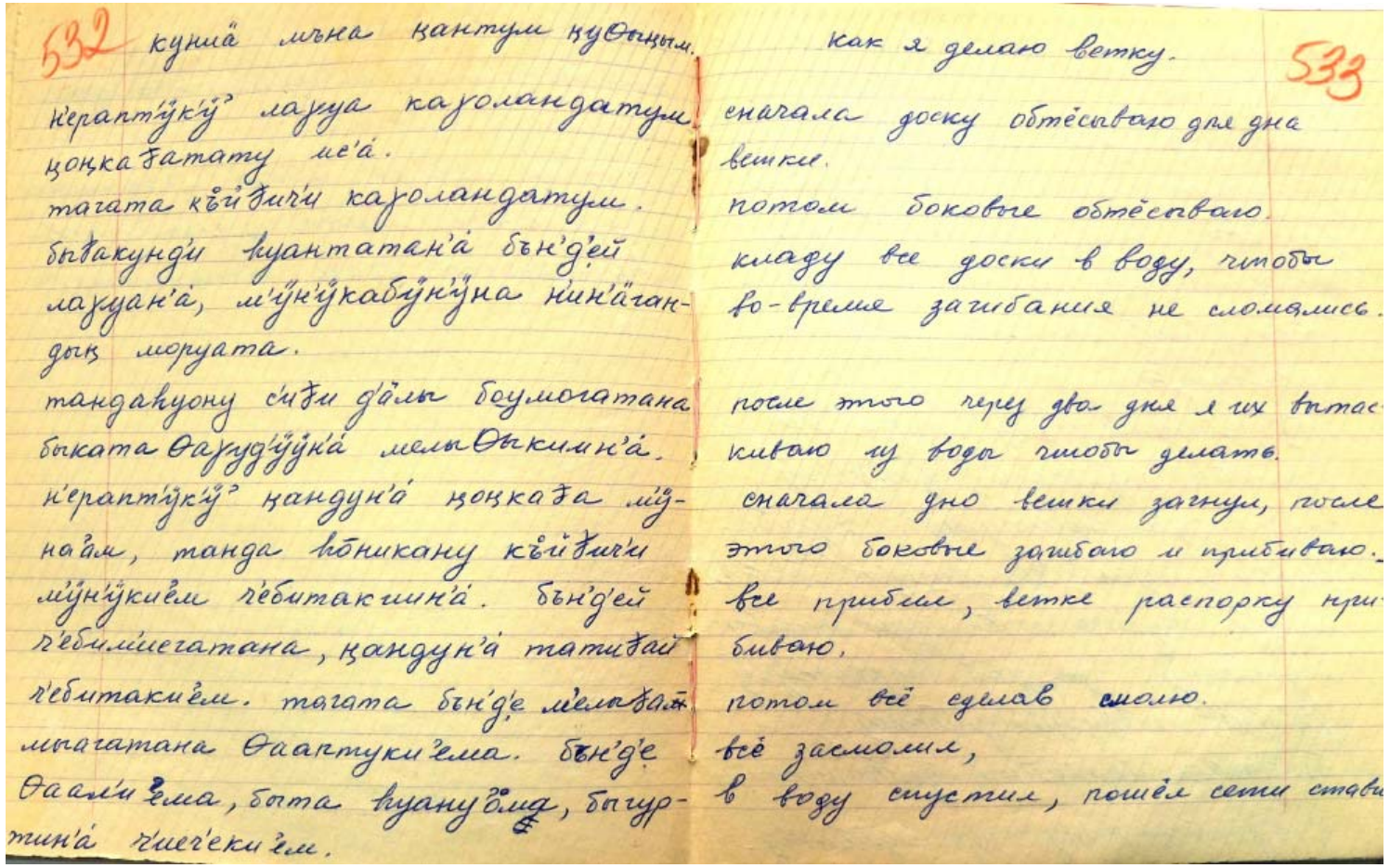
[illegible]

Дом зайца

1) В лесу лиса заяц жили двое вместе. 2) Пришла зима глубокая.3) Сильный мороз настал. 4) Звери замерзали. 5) Дома не сделали, еще в лесу ходят. 6) Заяц сделал себе дом из дерева, у зайца дом теплый. 7) Лиса сделала изо льда себе дом. 8) У лисы дом большой. 9) Весна настала. 10) У лисы дом растаял, под корнем живет. 11) Лиса пришла к зайцу, просит: "Пусти меня к себе в дом жить". 12) Заяц пустил лису в свой дом. 13) Они (двое) живут в одном доме. 14) Лиса хитрая, зайца прочь выгнала. 15) Заяц идет, сам плачет. 16) Собака спрашивает : "Ты что плачешь?" - "Меня лиса из дома выгнала, я дом строил. В моем доме живет, я по лесу болтаюсь." 17) Собака говорит: "Помогу тебе, я лису выгоню, сама убежит". 18) Они (двое) пришли к дому зайца. 19) Собака залаяла очень твердо. 20) Лиса испугалась, побежала. Выскочила, дом, бросал, оставила. 21) Собака и заяц живут двое в одном доме. 22) Собака дом караулит, никого не пускает.

[illegible]

Rohdaten



532 кунйа мына кантуи кудоцын.

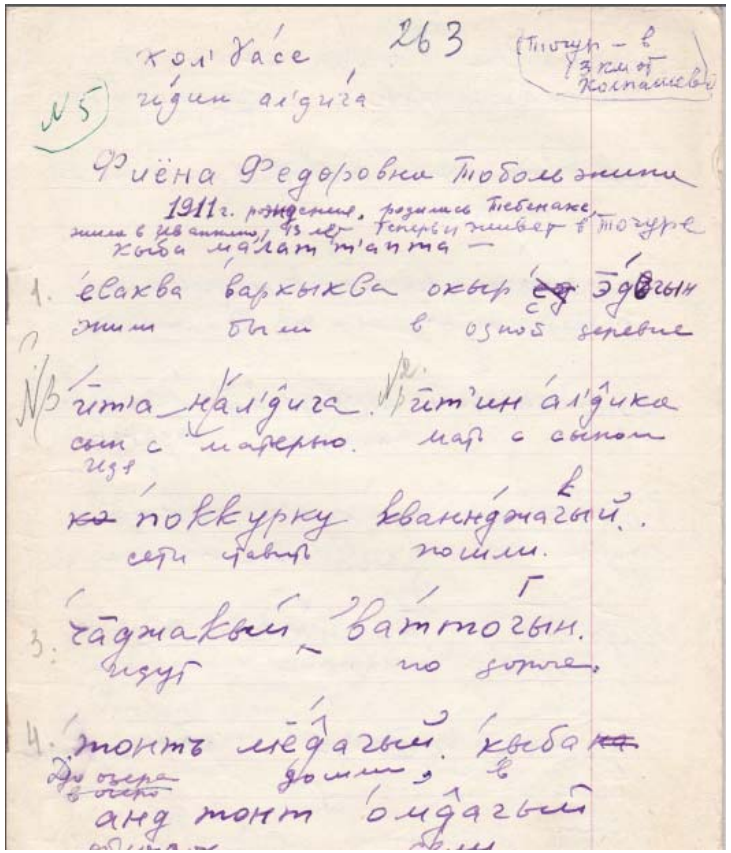
керапт'ук'ў лаууа каголандатуи цонкаʒатату ме'а.
тагата къйʒич'и каголандатуи.
бывакунди куантатан'а бын'дей лаууан'а, м'ўн'ўкабўн'ўна нин'ағандоқ моруата.
тандакуону сиʒи дʒалы боумогатана быката ʒаууддўўн'а мелыʒкимин'а.
н'ерапт'ук'ў? кандун'а цонкаʒа м'ўнаʒи, танда вбникану къйʒич'и м'ўн'ўкиʒем ч'ебитакин'а. бын'дей ч'ебим'егатана, кандун'а татиʒай ч'ебитакиʒем. тагата бын'дʒе м'елаʒай мыгатана ʒаактуки'ема. бын'дʒе ʒаал'и'ема, бата куануʒомд, багуртин'а ч'иеч'еки'ем.

533

как я делаю ветку.

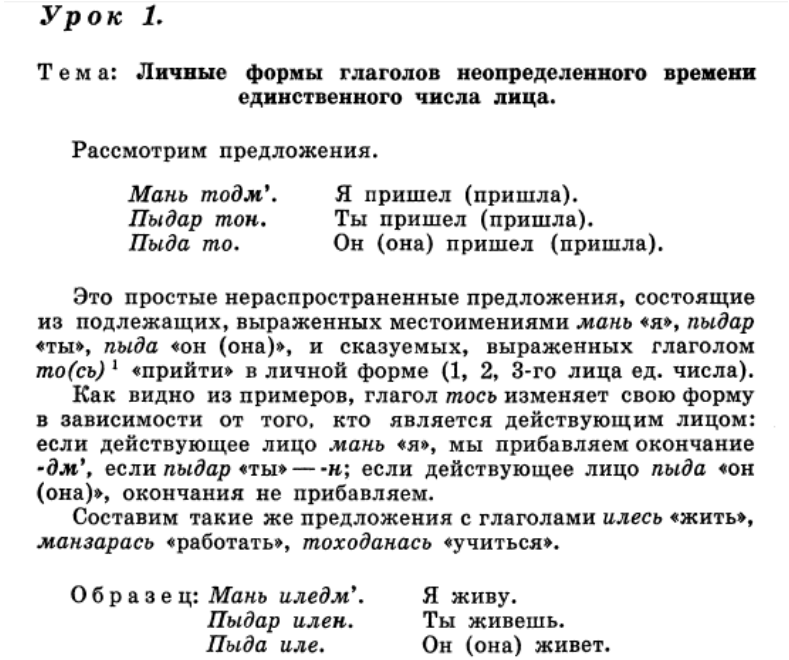
сначала доску обтёсываю для дна ветки.
потом боковые обтёсываю.
кладу все доски в воду, чтобы во-время загибания не сломались.
после того через два дня я их вытас
киваю из воды чтобы делать.
сначала дно ветки загну, после этого боковые загибаю и прибиваю.
все прибив, ветке распорку при
бываю.
потом всё сделав смолю.
всё засмолив,
в воду спустив, пошёл сети стави

Rohdaten



vorhandene Grammatiken

- von unterschiedlichem Niveau
- meistens russischsprachig



Урок 1.

Тема: Личные формы глаголов неопределенного времени единственного числа лица.

Рассмотрим предложения.

Мань тодм'.	Я пришел (пришла).
Пыдар тон.	Ты пришел (пришла).
Пыда то.	Он (она) пришел (пришла).

Это простые нераспространенные предложения, состоящие из подлежащих, выраженных местоимениями *мань* «я», *пыдар* «ты», *пыда* «он (она)», и сказуемых, выраженных глаголом *то(сь)* [1] «прийти» в личной форме (1, 2, 3-го лица ед. числа).

Как видно из примеров, глагол *тось* изменяет свою форму в зависимости от того, кто является действующим лицом: если действующее лицо *мань* «я», мы прибавляем окончание *-дм'*, если *пыдар* «ты» — *-н*; если действующее лицо *пыда* «он (она)», окончания не прибавляем.

Составим такие же предложения с глаголами *илесь* «жить», *манзарась* «работать», *тоходанась* «учиться».

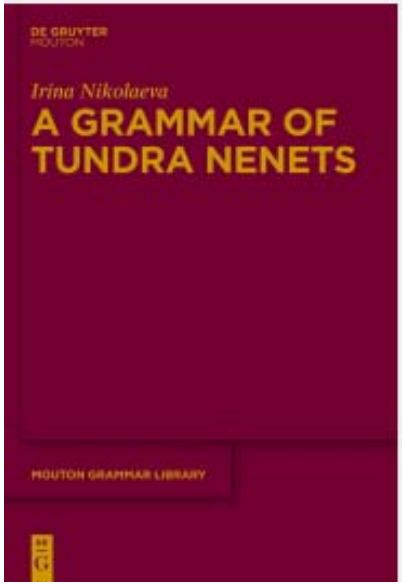
Образец:	*Мань иледм'.*	Я живу.
	Пыдар илен.	Ты живешь.
	Пыда иле.	Он (она) живет.

vorhandene Grammatiken

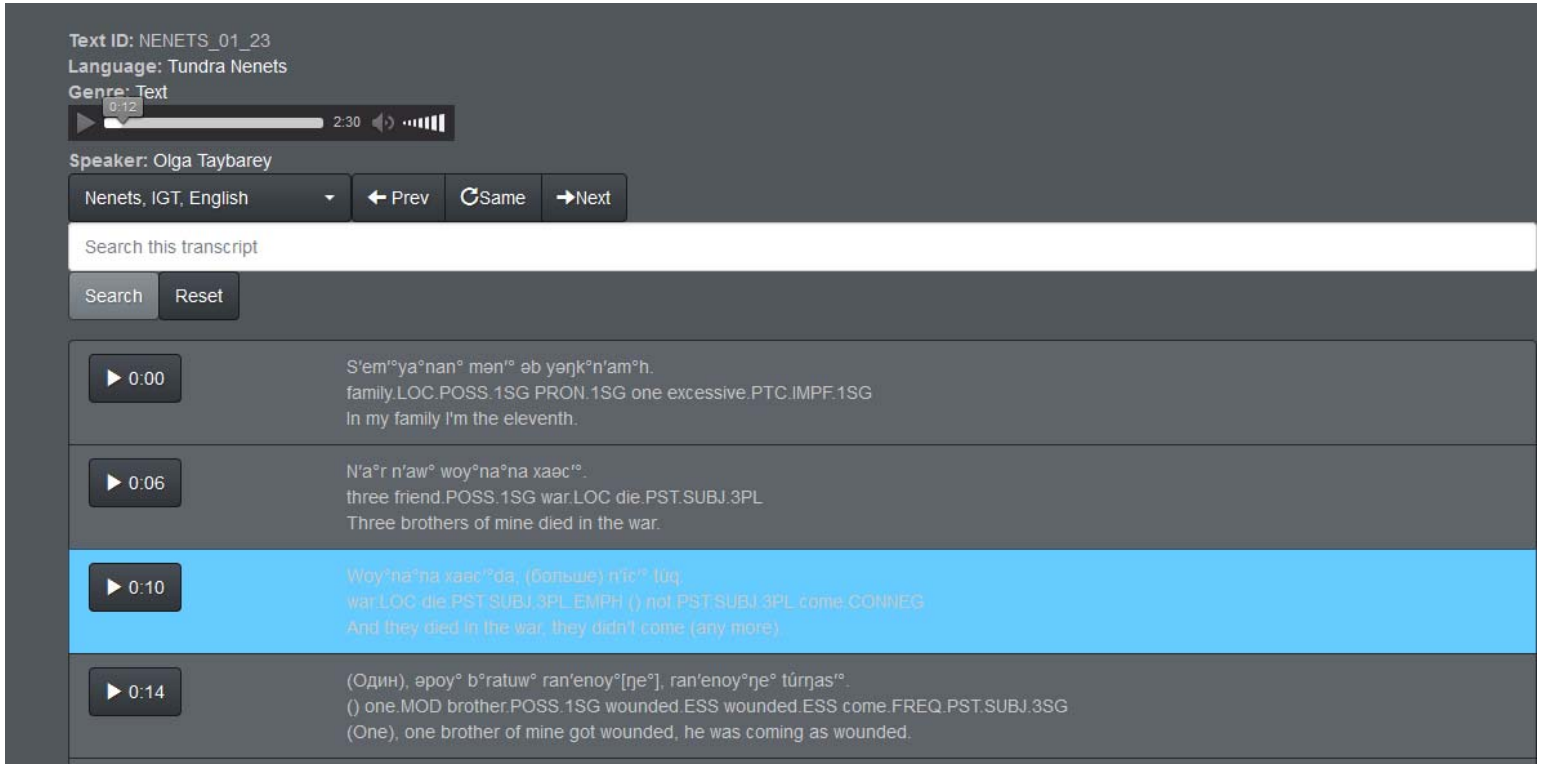
- angeblich korpusbasiert, aber die Beispiele sind im Korpus nicht identifizierbar

Sentence (30) was spontaneously produced when the speaker and the addressee looked at several moving cars. The speaker wanted to refer to one of these cars and used the word 'car' with the 2nd person possessive affix, although the car did not belong to the addressee. The reason for using the possessive suffix is that the speaker intended to call the attention of the addressee to the car. In other words, the car is 'yours' because 'I am talking to you about it'.

(30)	*siər-°q,*	*taki°*	*mašina-r°*	*xəya*
	look-IMP.2SG	that	car-2SG	go
	'Look, that car of yours left.'			



vorhandene Online-Ressourcen



Problemfelder: Bedarfslage

- die existierende Grammatiken sind meistens nicht korpusbasiert
- die Transkriptionen sind sehr unterschiedlich
- Sprachaufnahmen und Texte sind nicht verknüpft
- Texte sind meistens nicht glossiert, nicht annotiert

Anforderungen

- **Auffindbarkeit** von Ressourcen und Dokumenten

  → ?
- **Referenzierbarkeit**

  → ?
- **Zugänglichkeit** existierender sprachlichen Ressourcen

  → ?
- **Durchsuchbarkeit und Benutzbarkeit** der Ressourcen

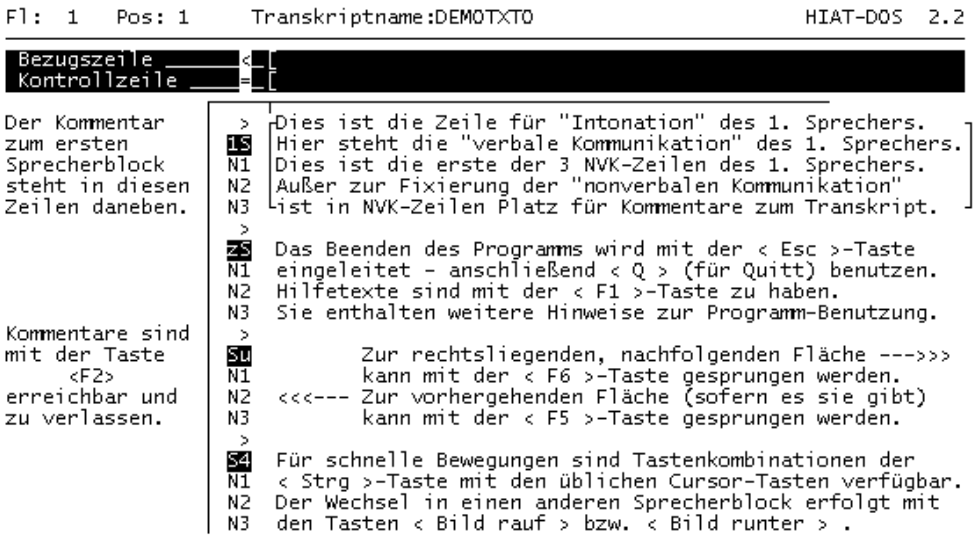
  → ?

Anforderungen

- **Auffindbarkeit** von Ressourcen und Dokumenten
  → Einsatz generischer Metadatenformate
- **Referenzierbarkeit**
  → Vergabe von Persistant Identifiern
- **Zugänglichkeit** existierender sprachlichen Ressourcen
  → Authentifizierungsstruktur und offene Lizenzmodelle
- **Durchsuchbarkeit und Benutzbarkeit** der Ressourcen
  → generische Datenformate und intuitive GUIs

Vor 18 Jahren …

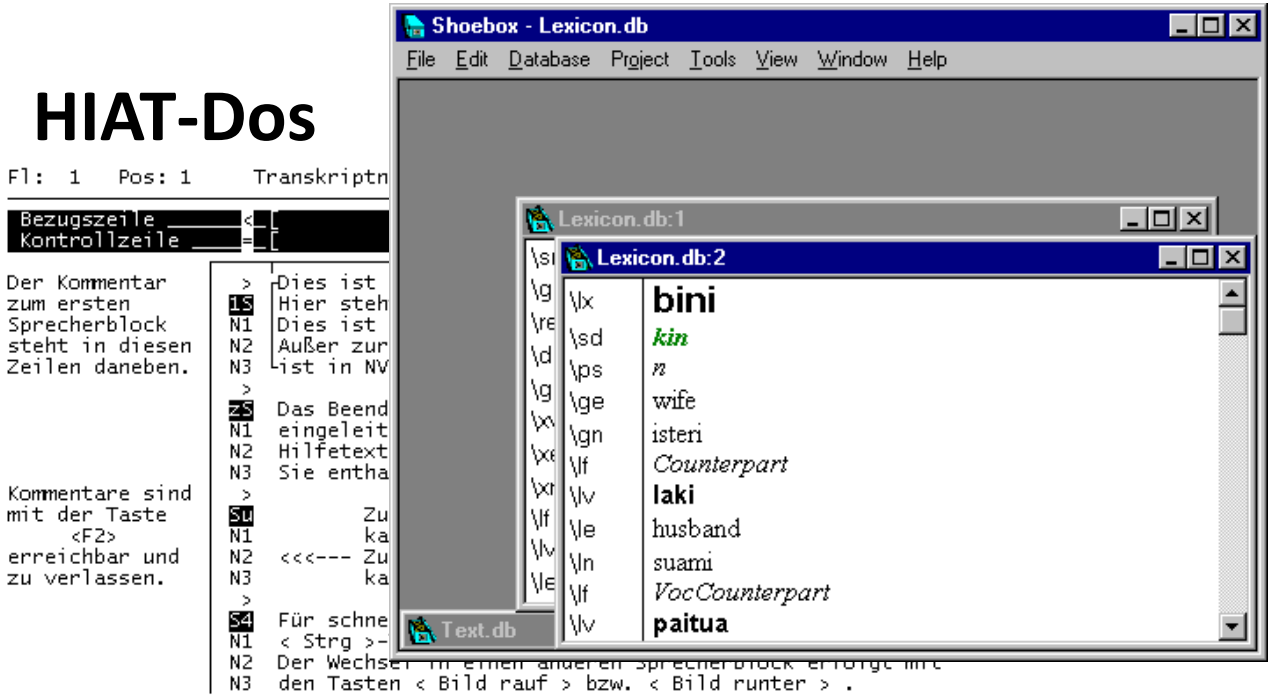
HIAT-Dos



Vor 18 Jahren …

Shoebox

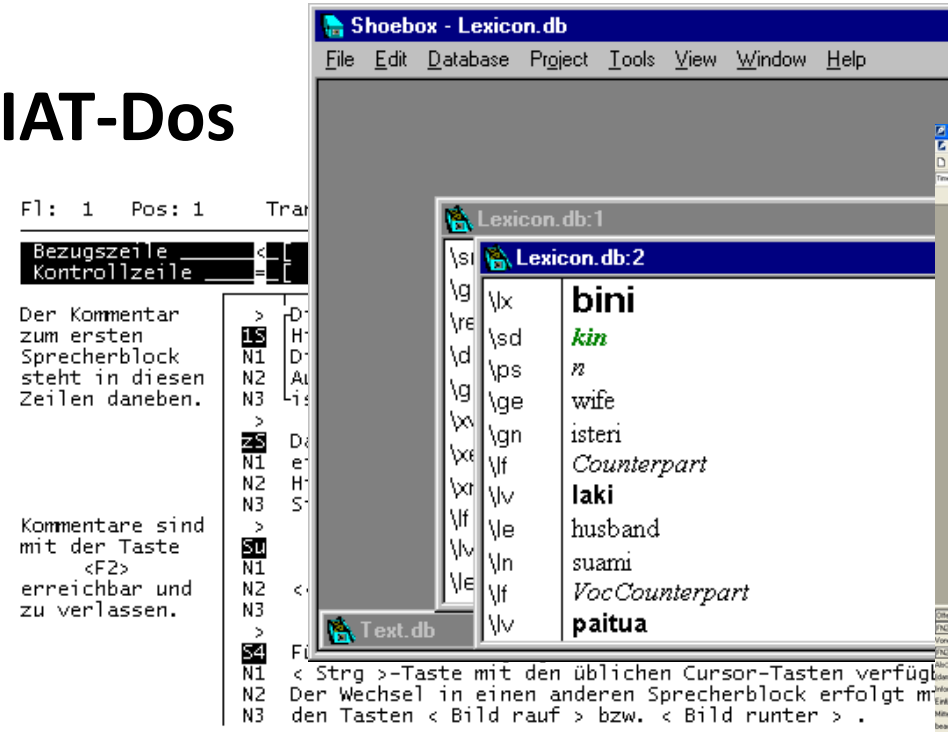
HIAT-Dos



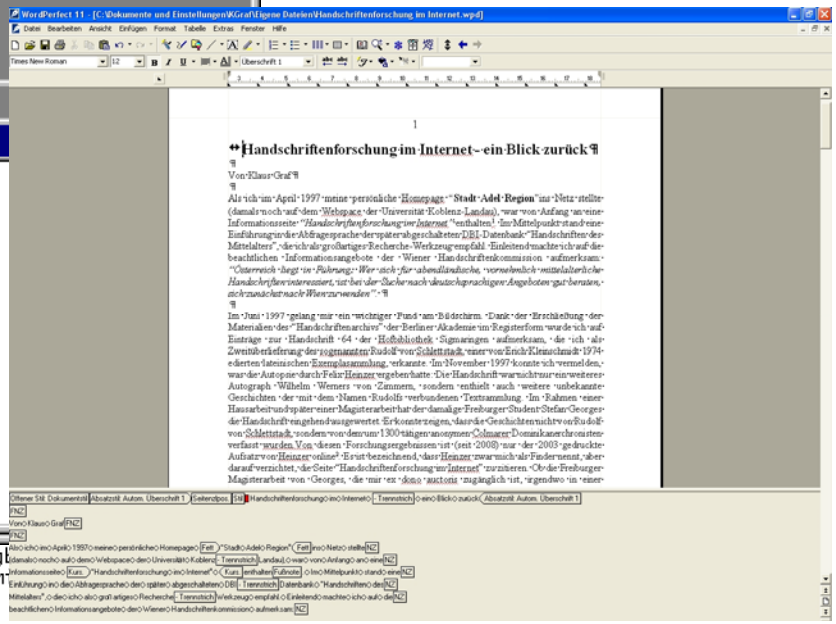
Vor 18 Jahren …

Shoebox

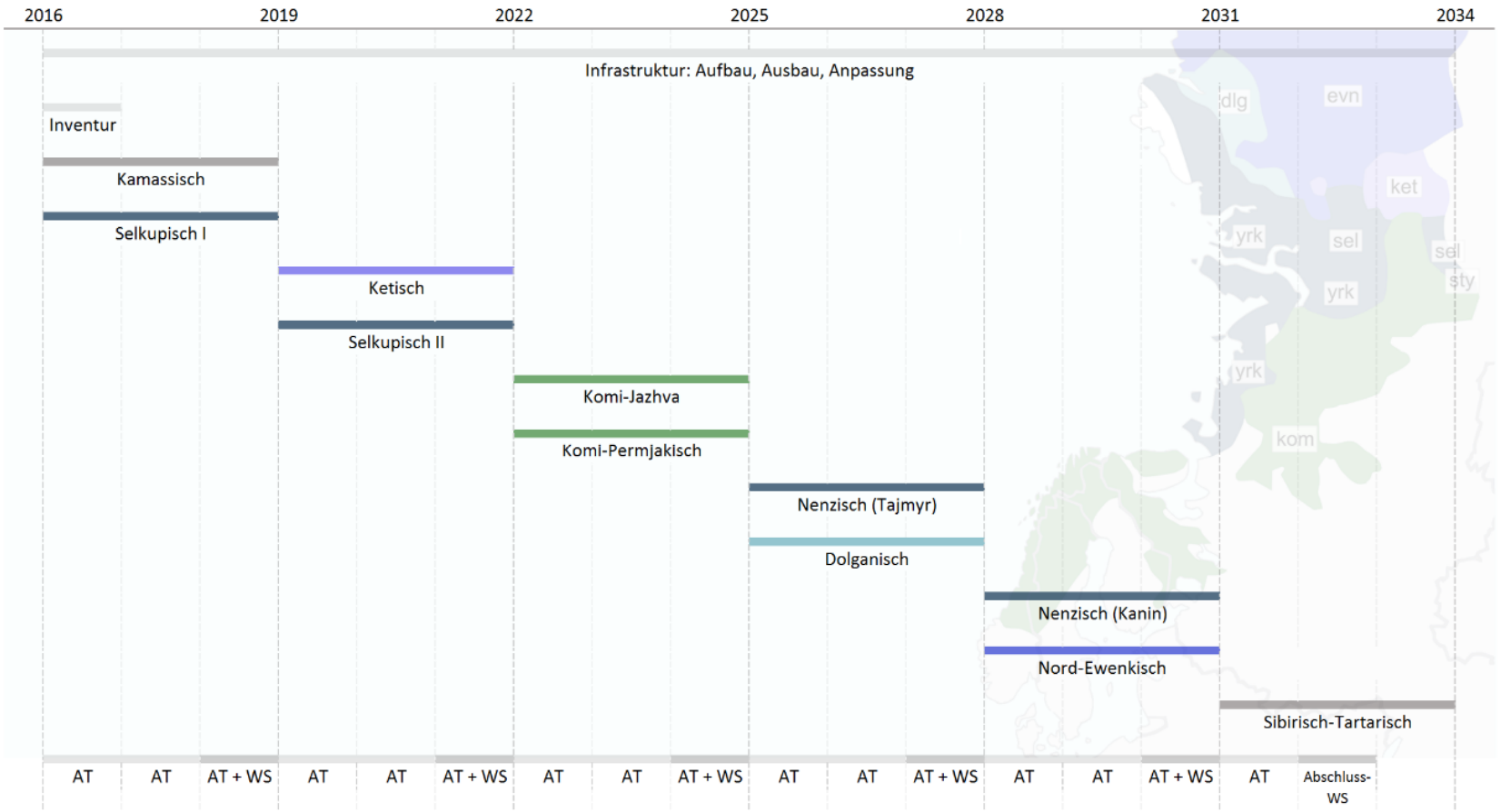
HIAT-Dos



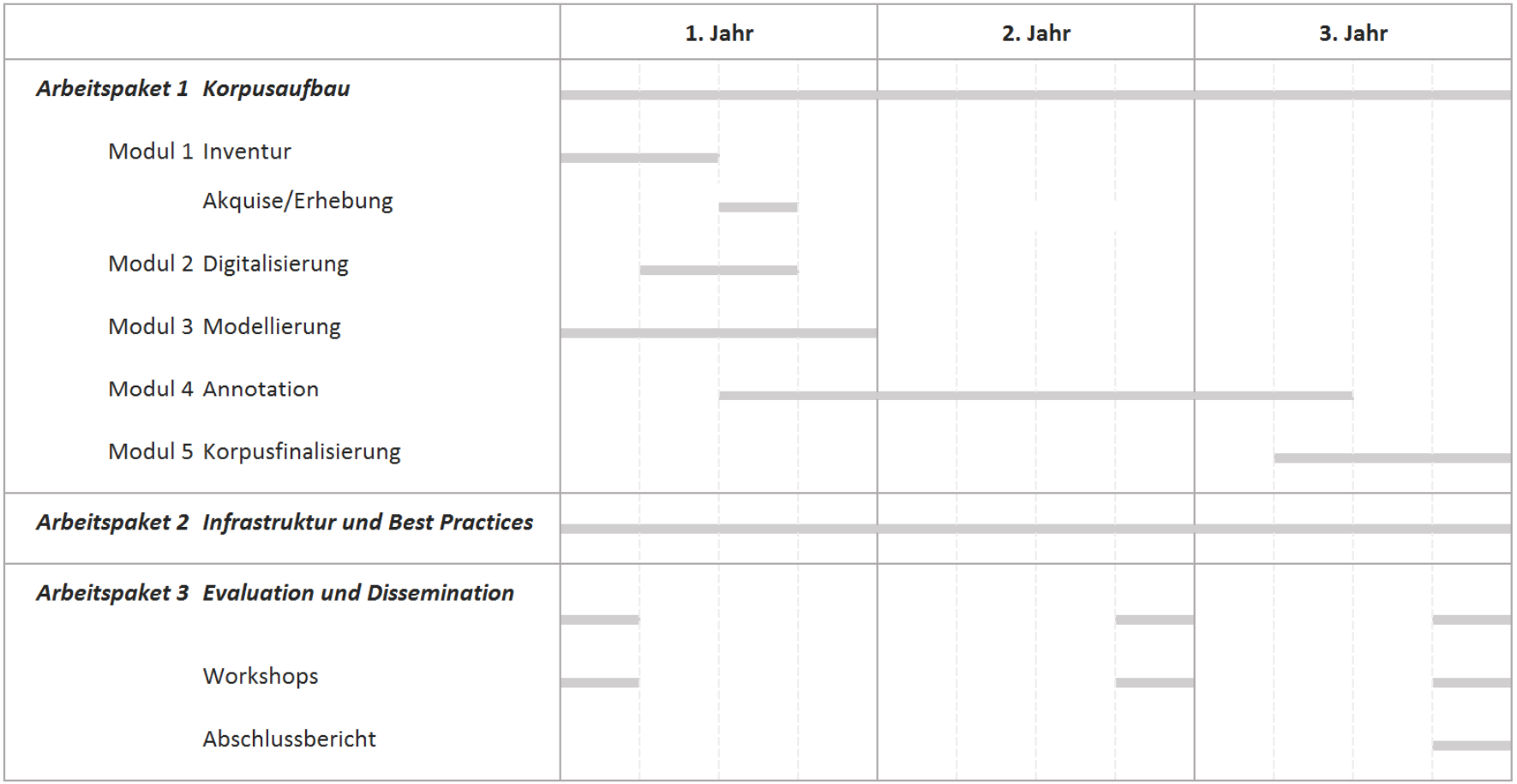
Word Perfect



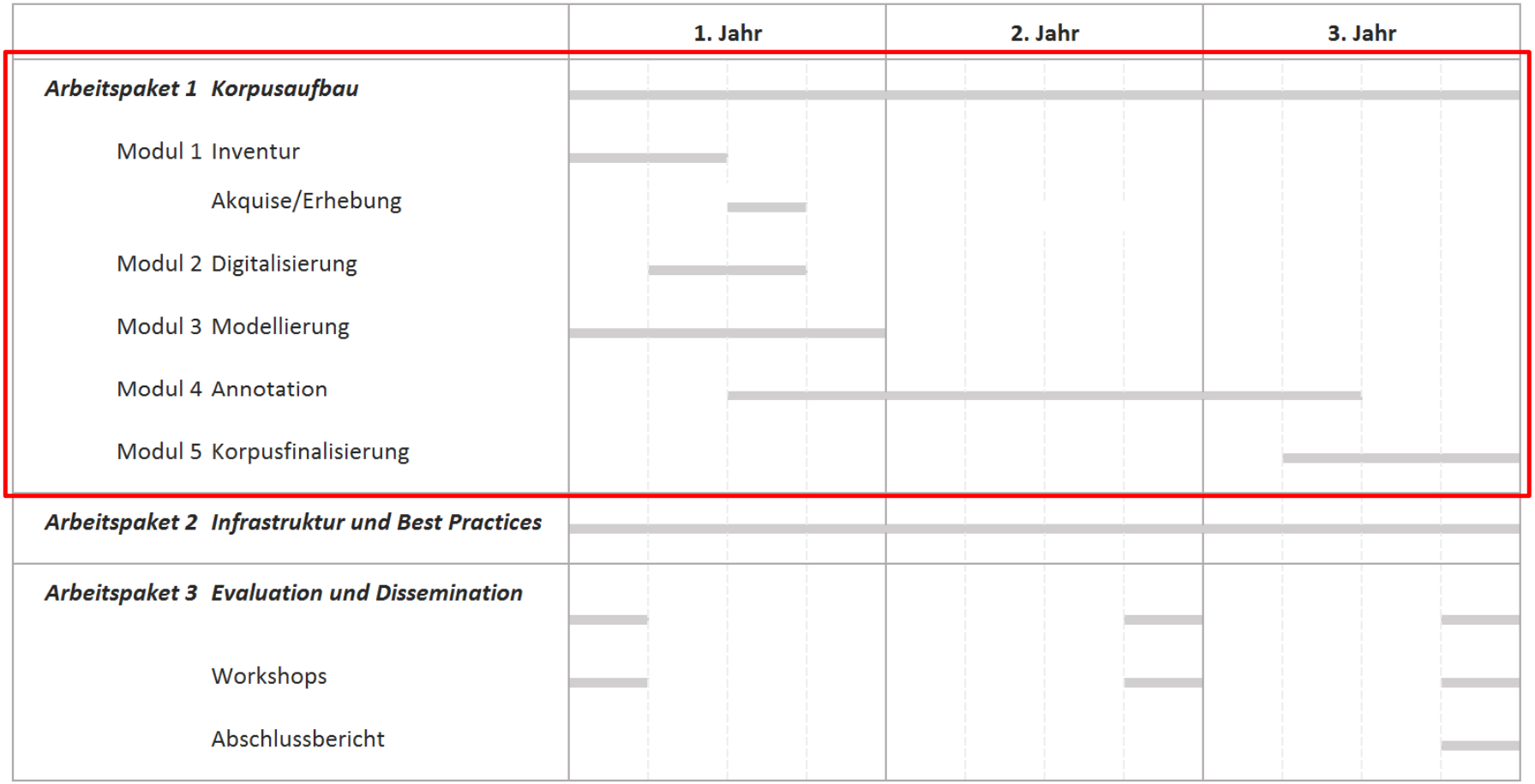
Arbeitsprogramm



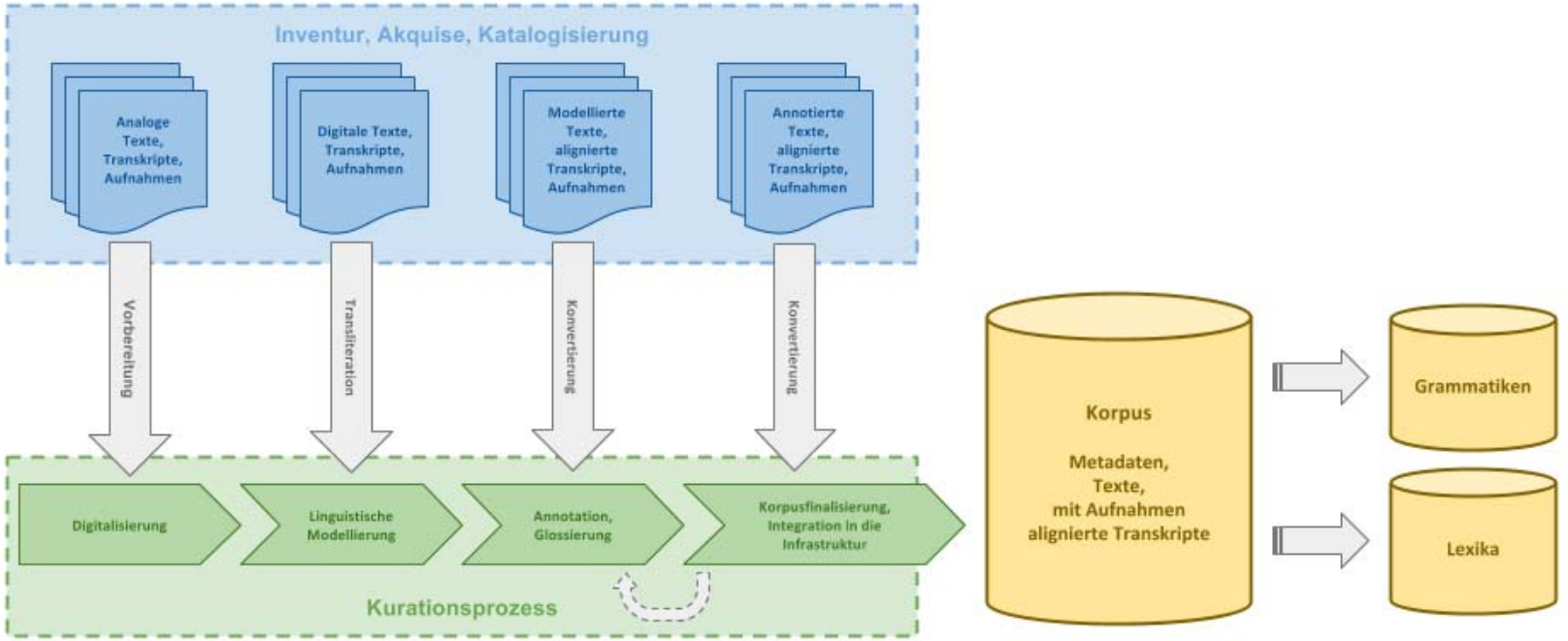
Arbeitsprogramm



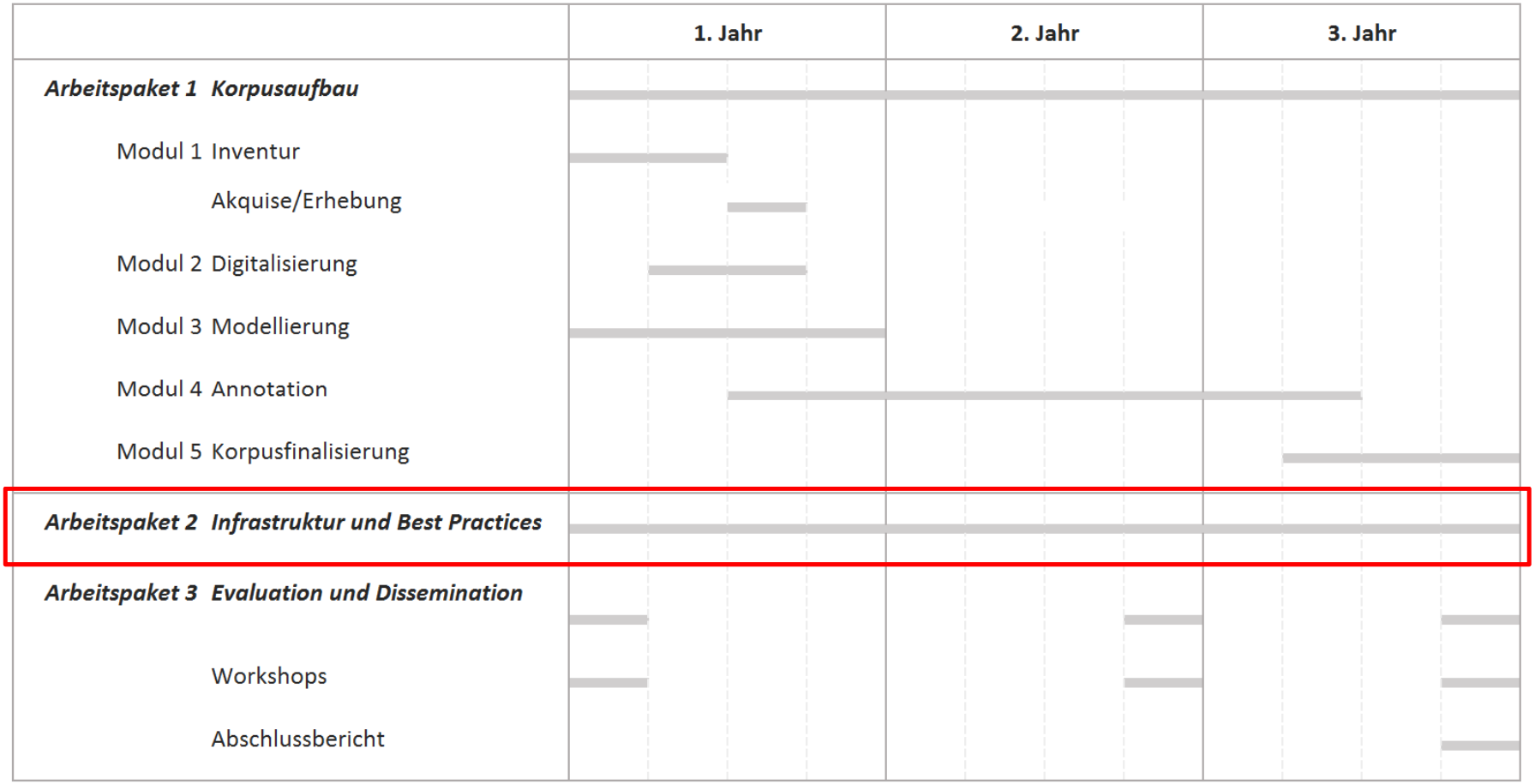
Arbeitsprogramm – AP1 Korpusaufbau



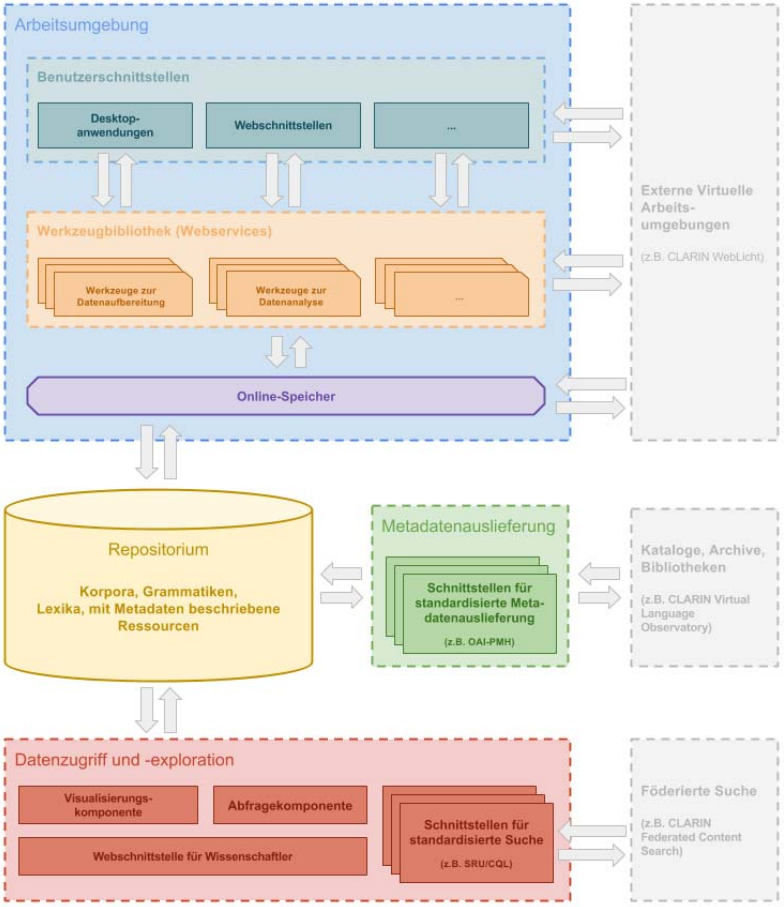
Arbeitsprogramm – AP1 Korpusaufbau



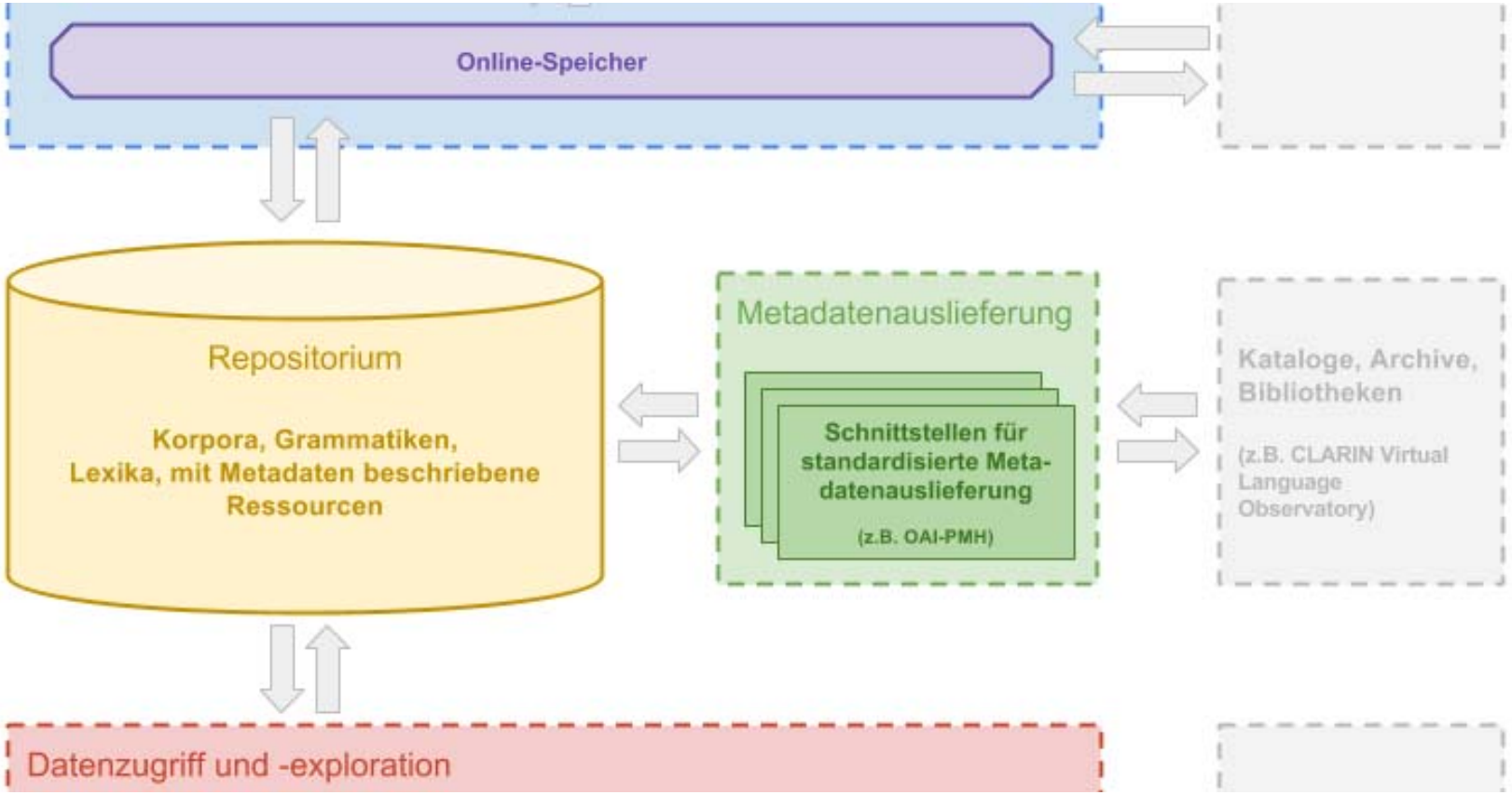
Arbeitsprogramm – AP1 Korpusaufbau



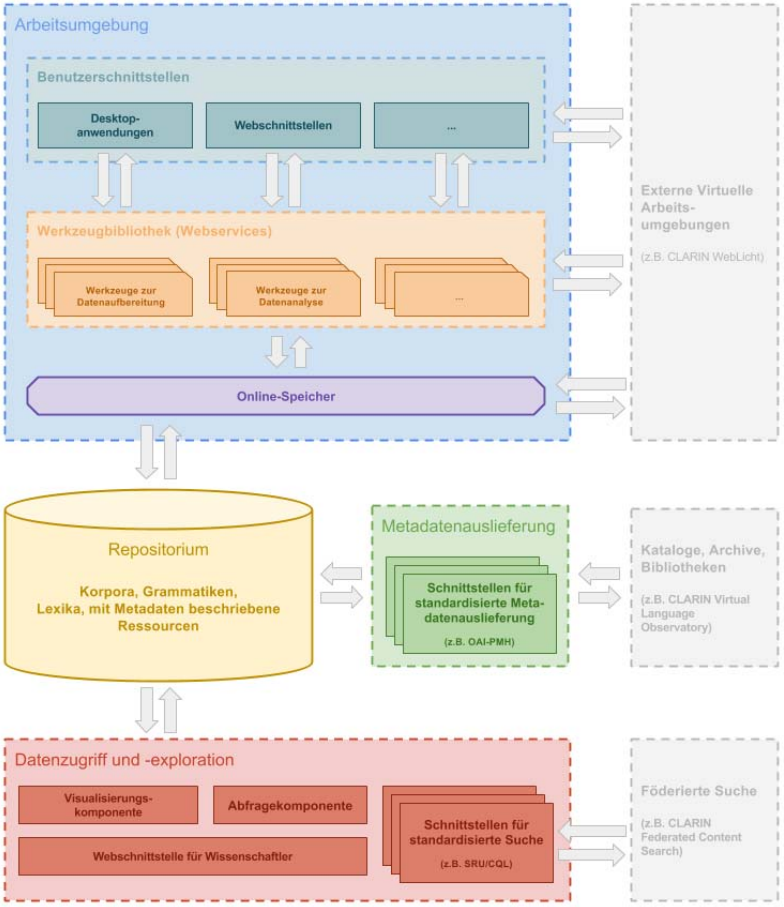
Forschungsumgebung



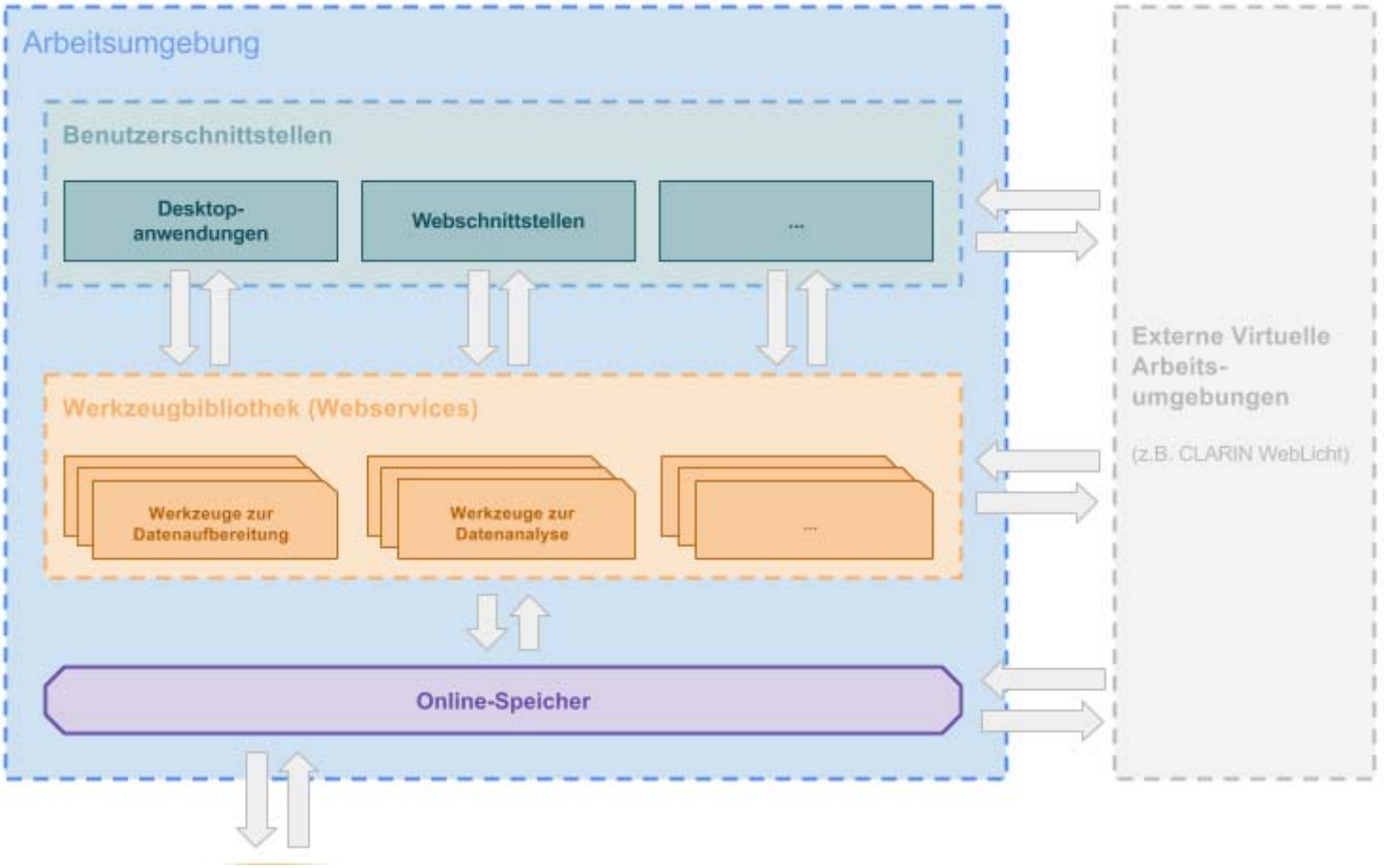
Forschungsumgebung



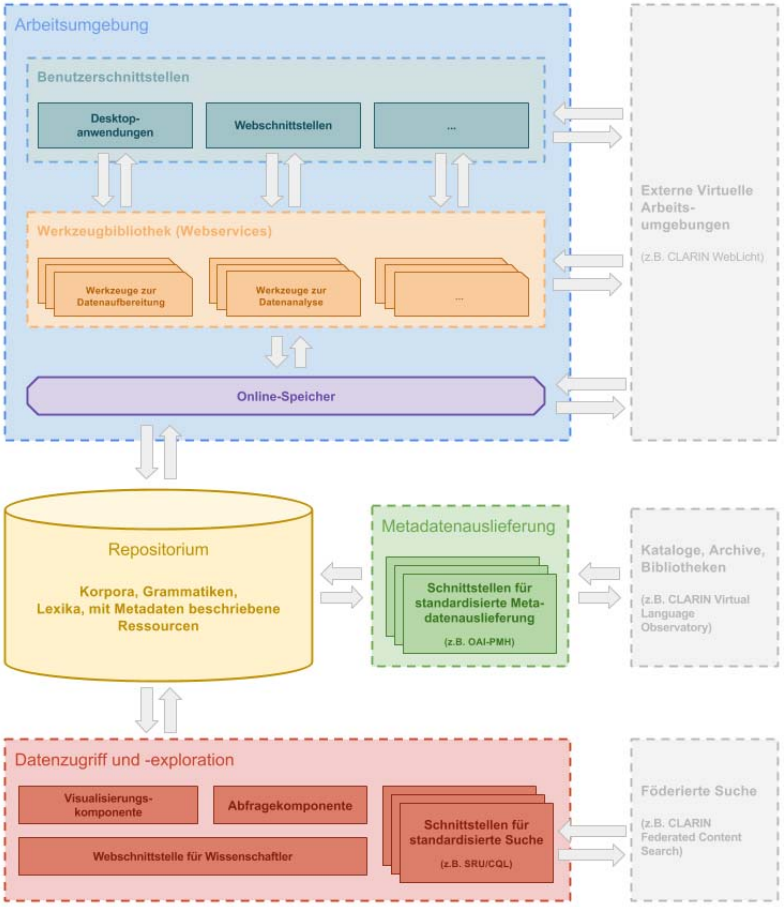
Forschungsumgebung



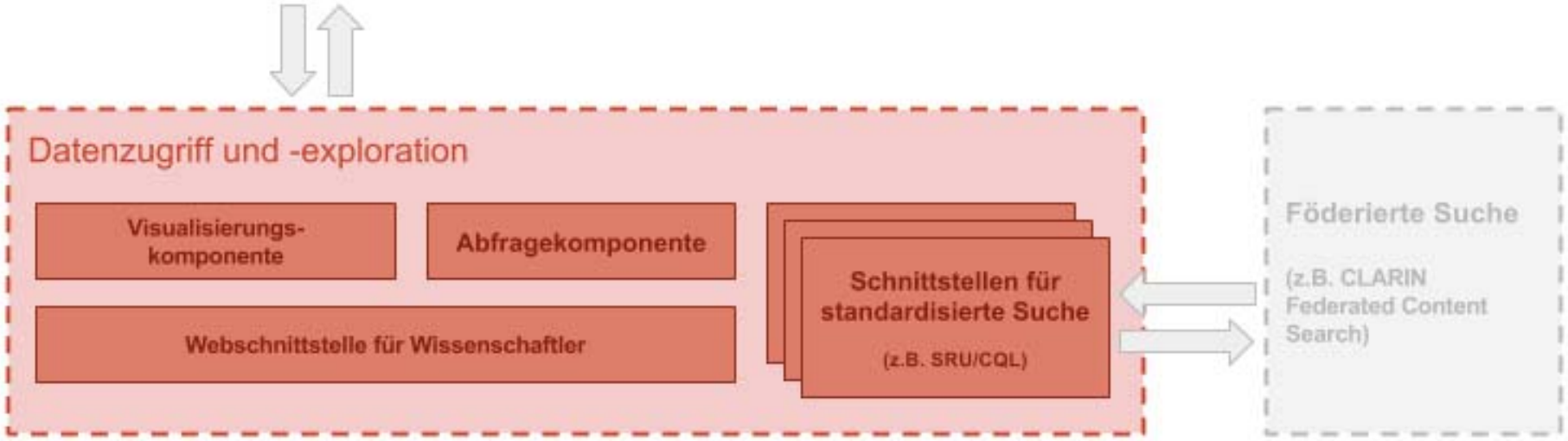
Forschungsumgebung



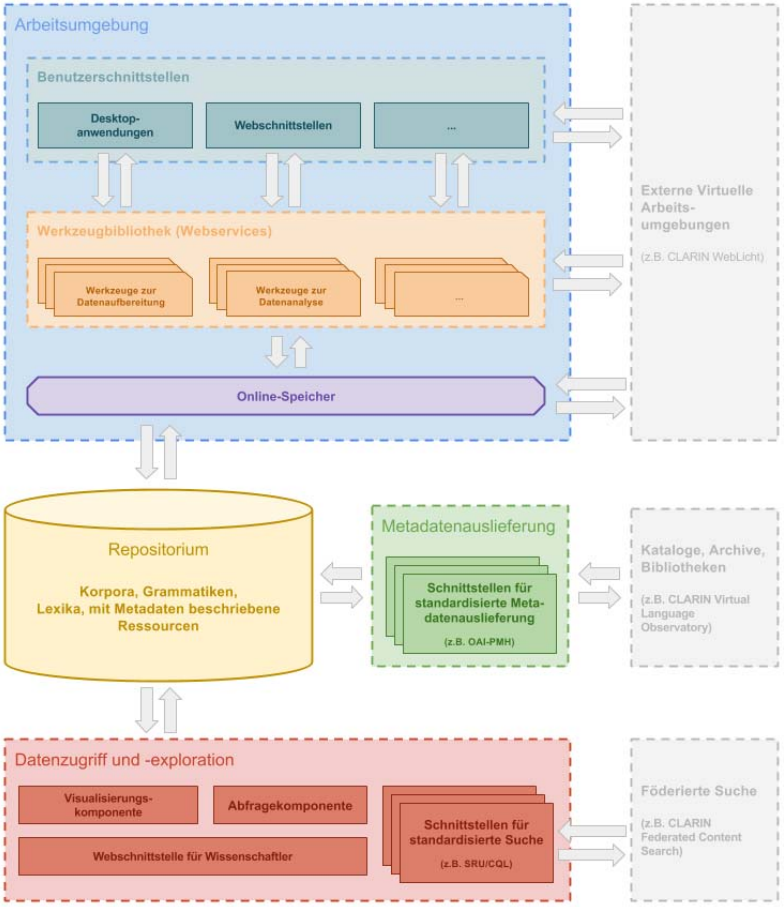
Forschungsumgebung



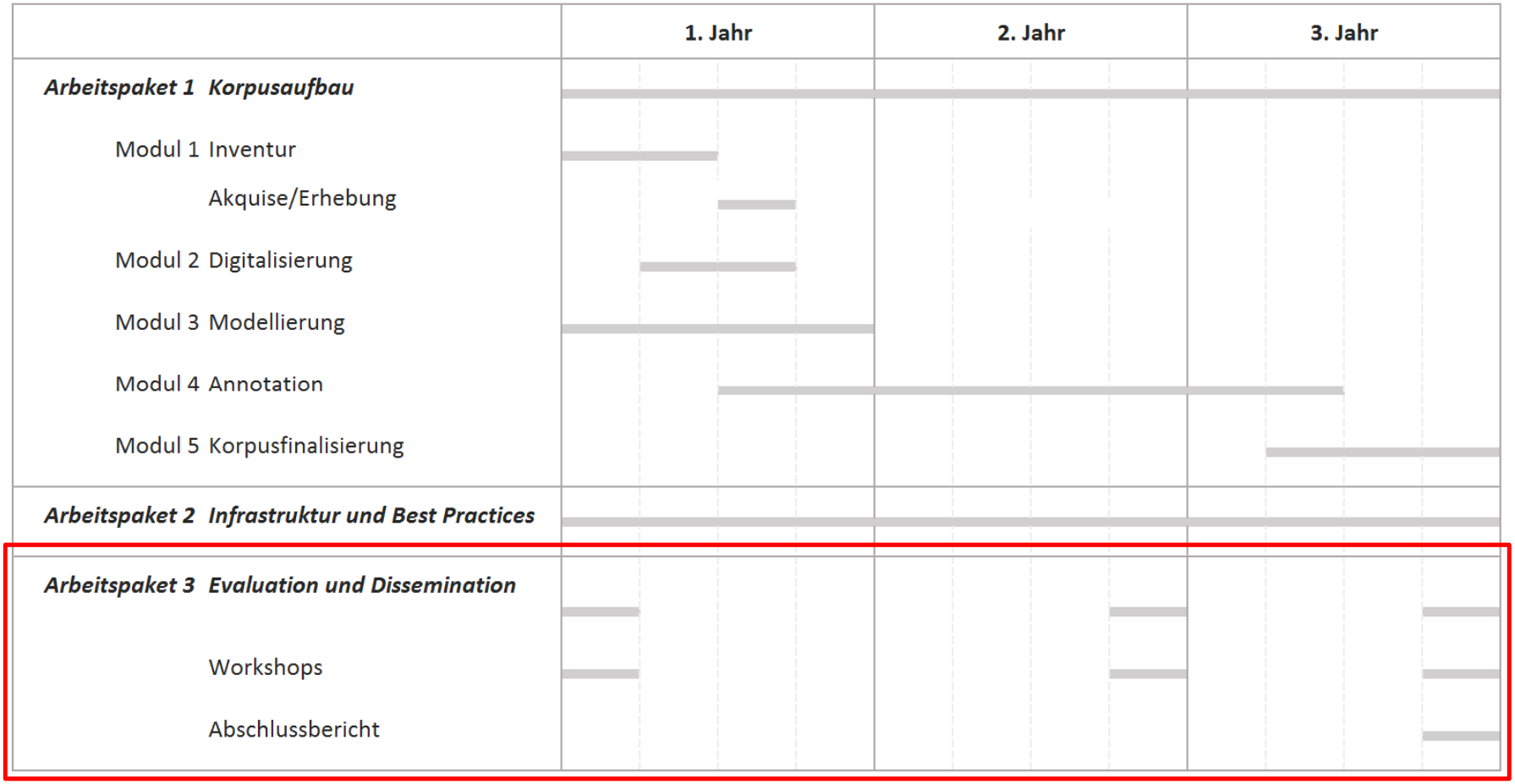
Forschungsumgebung



Forschungsumgebung



Arbeitsprogramm – AP1 Korpusaufbau



Alles wird gut!

- Usecase für digitale Forschungsinfrastrukturen
- Sprachbeschreibung(en) auf der Grundlage adäquater empirischer Ressourcen
- Integration der uralistischen Community in bestehende Forschungsinfrastrukturen